BigPicture 2026

The Big Picture v2: Crafting a Research Narrative

Proceedings of the Workshop

July 4, 2026

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-416-3

Introduction

Welcome to the Proceedings of the first iteration of the Big Picture Workshop (The Big Picture: Crafting a Research Narrative). The workshop is hosted at ACL 2026, in San Diego, USA, on July 4th, 2026.

The Big Picture Workshop provides a dedicated venue for exploring and distilling broader NLP research narratives. All research exists within a larger context, and progress is made by standing on the shoulders of giants: building on the foundations laid by earlier researchers. In light of rapid publication rates and concise paper formats, it has become increasingly difficult, however, to recognize the larger story to which a paper is connected. The Big Picture Workshop invites researchers to reflect on how their individual contributions fit within the overall research landscape and what stories they are telling with their bodies of research. The goals of the workshop are to enhance communication and understanding between different lines of work, highlight how works connect and build on each other, generate insights that are difficult to glean without combining and reconciling different research narratives, encourage broader collaboration and awareness of prior work in the NLP community, and facilitate understanding of trajectories and insights within the field of NLP.

We received 15 submissions, of which we accepted 12 for presentation at the workshop. Those 12 accepted papers are contained in this volume.

The workshop schedule features one standard invited talk, and three special invited presentations designed to foster live engagement between different lines of related work. In these special presentations, two to three invited presenters speak on their individual lines of work and the connections between them, followed by a moderated discussion further exploring the overall narrative that emerges from these works in aggregate. In addition to invited presentations, the workshop features an in-person poster session, and spotlight talks.

We extend heartfelt thanks to our program committee, our participants, and all authors who submitted papers for consideration—your engagement has been critical to the success of the workshop. Finally, we thank the ACL 2026 organizers and workshop chairs for their hard work and support.

The Big Picture Workshop Organizers,

Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder

Organizing Committee

Organizers

Yanai Elazar, Bar-Ilan University
Allyson Ettinger, Allen Institute for AI
Nora Kassner, Google DeepMind
Sebastian Ruder, Meta Superintelligence Labs

Program Committee

Reviewers

David Ifeoluwa Adelani, Maria Antoniak

Alexandra Chronopoulou, Marta R. Costa-jussà

Matthias Gallé

Tiago Pimentel, Barbara Plank

Philip Resnik, Kyle Richardson

Vered Shwartz

Yogarshi Vyas

Xinyi Wang, Chenxi Whitehouse

Keynote Talk

Where it Hurts: Finding Durable Questions While Moving Fast

Noah A. Smith
University of Washington & Allen Institute for Artificial Intelligence
2026-04-07 09:30:00 – Room: **TBD**

Abstract: In a fast-moving field, it can be hard to tell which problems are urgent, which are merely loud, and which are worth building a research life around. This talk considers how researchers can stay responsive to rapid change without letting the field's volatility set their agenda. I will discuss tools for identifying which questions remain meaningful across shifts in methods, data, benchmarks, and institutions. The goal is a practical vocabulary for finding direction when everything seems to be moving at once.

Bio: Noah A. Smith is the inaugural Vice Provost for Artificial Intelligence and Charles and Lisa Simonyi Endowed Chair for Artificial Intelligence and Emerging Technologies at the University of Washington, where he is also a Professor in the Paul G. Allen School of Computer Science & Engineering. He is Senior Director of NLP Research at the Allen Institute for Artificial Intelligence, directs the OLMo open language modeling effort, and leads the NSF- and NVIDIA-supported project "Open Multimodal AI Infrastructure to Accelerate Science." His research spans language and music technologies, multimodal AI, and multifaceted evaluation of AI systems.

Keynote Talk

Does mechanistic interpretability need interventions?

Aaron Mueller
Boston University
2026-04-07 10:50:00 – Room: **TBD**

Abstract: Mechanistic interpretability often treats interventions as the gold standard of evidence, relying on circuit ablations and representation steering to support claims about how models actually work. But are interventions really sufficient, or even necessary, for making mechanistic claims? In this debate-style talk, we trace the history that led the field to embrace interventions, and argue that the answer to both questions is a contentious no.We start by showing that interventions alone are not sufficient to explain model behavior by highlighting cases where causal methods can produce misleading or outright spurious explanations. Then, we debate whether interventions are necessary at all, exploring how alternative notions of causality and carefully designed behavioral evidence may also support strong mechanistic claims without directly intervening on a model. We conclude by discussing future directions for mechanistic interpretability, and how we can draw inspiration from other scientific disciplines to ask what should count as a good explanation.

Bio: Aaron Mueller is an Assistant Professor of Computer Science and, by courtesy, of Data Science at Boston University. His research centers on developing interpretability and evaluation methods inspired by causal and linguistic principles, and applying these to precisely control and improve the generalization of language technologies. He completed his Ph.D. at Johns Hopkins University.

Keynote Talk

Does mechanistic interpretability need interventions?

Tiago Pimentel
ETH Zurich
2026-04-07 10:50:00 – Room: **TBD**

Abstract: Mechanistic interpretability often treats interventions as the gold standard of evidence, relying on circuit ablations and representation steering to support claims about how models actually work. But are interventions really sufficient, or even necessary, for making mechanistic claims? In this debate-style talk, we trace the history that led the field to embrace interventions, and argue that the answer to both questions is a contentious no.We start by showing that interventions alone are not sufficient to explain model behavior by highlighting cases where causal methods can produce misleading or outright spurious explanations. Then, we debate whether interventions are necessary at all, exploring how alternative notions of causality and carefully designed behavioral evidence may also support strong mechanistic claims without directly intervening on a model. We conclude by discussing future directions for mechanistic interpretability, and how we can draw inspiration from other scientific disciplines to ask what should count as a good explanation.

Bio: Tiago Pimentel is a Postdoctoral Researcher at ETH Zürich, working in machine learning interpretability and psycholinguistics. His long-term goal is to understand how humans and machines process language. To this end, his research adopts an interdisciplinary approach, leveraging information theory and causality to study the mechanisms behind model behaviour and human cognition.

Table of Contents

Program

Saturday, July 4, 2026

09:00 - 09:10 *Opening Remarks*

09:10 - 10:30 *Paper Presentations*

From Natural Language to Certified Geometry Proofs: A Survey of LLM-Augmented Verification and Neuro-Symbolic Theorem Proving
Ioannis Tzachristas and Georgios Tzachristas

Open Problems Solved by LLMs? A Survey of Verifiable Mathematical Discovery
Ioannis Tzachristas, Georgios Tzachristas and Aifen Sui

Beyond Hallucination: Reframing LLM Quality Assessment as Task-Output Alignment
Andrew Hoblitzell

Challenging the Myth: A Research Arc on LLMs as Human Simulacra
Simon Münker, Achim Rettinger and Damian Trilling

Towards Trustworthy AI-Mediated Communication Across Languages and Cultures
Dayeon Ki

Challenging Quadratic Attention - A Holistic View On the Rise of Alternative Language Model Architectures
Alexander M. Fichtl, Jeremias Bohn, Josefin Kelber, Edoardo Mosca and Georg Groh

10:30 - 11:00 *Break*

11:00 - 12:30 *Paper Presentations*

Why Low-Resource NLP Needs More Than Cross-Lingual Transfer: Lessons Learned from Luxembourgish
Fred Philippy, Siwen Guo, Jacques Klein and Tegawendé F. Bissyandé

Building Arabic NLP from the Ground Up: Twenty Years of Lessons, Failures, and Open Problems
Wajdi Zaghouani

Saturday, July 4, 2026 (continued)

Speaking of Language: Reflections on Metalanguage Research in NLP
Nathan Schneider and Antonios Anastasopoulos

Harnessing the Latent Space: From Steering Vectors to Model Calibrators for Control and Trust
Nishant Subramani

Language Models as Measurement Apparatus for Culture
Kent K. Chang

Memorisation Meets Compositionality in Natural Language Processing
Verna Dankers

12:30 - 12:45 *Closing Remarks*

From Natural Language to Certified Geometry Proofs: A Survey of LLM-Augmented Verification and Neuro-Symbolic Theorem Proving

Ioannis Tzachristas [1,2]*, **Georgios Tzachristas** [1,3]*
[1]Huawei European Research Institute
[2]Technical University of Munich, Germany
[3]National Technical University of Athens, Greece

Abstract

Large Language Models (LLMs) can produce convincing geometric arguments, yet their outputs are not reliable enough to be treated as proofs without independent verification. In parallel, symbolic geometry tools (e.g. automated theorem provers in dynamic geometry systems) offer strong rigor guarantees, but require formalized inputs and can struggle with problem formalization, auxiliary construction, and proof presentation. This survey synthesizes work at the intersection of these lines: *hybrid LLM–symbolic systems for geometry* that (i) translate natural language and diagrams into formal constraints, (ii) search for solution plans and proof steps using learned or heuristic methods, and (iii) verify the resulting steps using symbolic provers or proof assistants. We propose a taxonomy organized around (a) the role of the LLM in the pipeline (parser, strategist, prover, critic), (b) the target proof artifact (answer-only, informal proof, semi-formal step trace, or kernel-checked formal proof), and (c) the verification backend (numeric testing, algebraic provers, synthetic provers, and proof-assistant kernels). We review representative systems in NLP and AI (e.g. GeoS, Inter-GPS, FormalGeo, AlphaGeometry, AutoGPS, and recent heuristic-only deductive solvers), and connect them to broader neurosymbolic paradigms for *faithful* reasoning (e.g. SatLM, LINC, and autoformalization). Finally, we outline evaluation protocols emphasizing step-level soundness and robustness, and we discuss open problems in multimodal formalization, handling of non-degeneracy conditions, human-readable certified proofs, and reproducibility.

1 Introduction

Geometry sits at the intersection of language, vision, and formal reasoning. Many geometry problems are described in natural language and supported by diagrams, but the desired solution is often a *proof*—a chain of logically valid steps. Classical NLP systems for geometry (e.g. GeoS (Seo et al., 2015)) and later interpretable pipelines (e.g. Inter-GPS (Lu et al., 2021), GeoQA (Chen et al., 2021)) illustrate the core challenge: correct solutions require a consistent interpretation of text, diagram, and domain axioms.

Recent progress in neural theorem proving and LLM-based reasoning has renewed interest in turning informal mathematical reasoning into verifiable proof artifacts (Wu et al., 2022; Li et al., 2024). In Euclidean geometry, this momentum is amplified by strong symbolic engines and new neuro-symbolic solvers that reach Olympiad-level performance (Trinh et al., 2024; Chervonyi et al., 2025; Duan et al., 2025). However, the *verification gap* remains: LLMs can hallucinate steps, omit conditions, and produce plausible-but-invalid proofs, while symbolic provers require careful formalization and may return results that are hard to interpret pedagogically (e.g. algebraic certificates and non-degeneracy conditions).

This survey focuses on **verification-oriented** geometry pipelines that integrate LLMs with symbolic tools. We treat verification broadly, ranging from (i) *symbolic checkers* embedded in dynamic geometry systems such as GeoGebra (Kovács and Sólyom-Gecse, 2016; Botana et al., 2015) to (ii) *formal proof assistants* (Lean/Coq/Isabelle) and their kernels. We argue that modern geometry automation should be evaluated not only by answer accuracy but also by the *soundness and auditability* of its intermediate reasoning steps. We contribute:

1. A **taxonomy** of LLM–symbolic geometry systems grounded in roles, representations, and verification backends.
2. A **comparative review** of symbolic geometry provers (algebraic and synthetic), formal proof assistants, and dynamic geometry environments used for automated verification.

*Equal contribution.
Correspondence: ioannis.tzachristas@tum.de

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 1–9
July 4, 2026

3. A **survey of neuro-symbolic and multimodal systems** for geometry problem solving, including recent Olympiad-level solvers and formalized geometry frameworks.
4. A set of **evaluation guidelines** emphasizing step-level validity, robustness, and reproducibility.

2 Problem Setting and Terminology

2.1 Inputs, outputs, and levels of rigor

Geometry automation spans multiple input modalities and proof artifacts:

Inputs. (i) *Formal* constructions (e.g. coordinates or a domain-specific language), (ii) *text-only* problem statements, (iii) *text + diagram* (raster or vector).

Outputs. (i) a final *answer* (numeric or multiple-choice), (ii) an *informal proof* in natural language, (iii) a *semi-formal* step trace with explicit theorem applications, (iv) a *formal proof* checked by a proof assistant kernel.

Verification targets. A crucial distinction is whether the system produces (or can be converted into) an artifact that a *trusted checker* can validate. We distinguish:

- **Empirical validation:** random instantiation, numeric sampling, or bounded "exact check".
- **Symbolic validation:** algebraic methods (Wu/Gröbner) and semi-algebraic/synthetic methods (area/full-angle/coherent logic).
- **Kernel validation:** proof assistant kernel checking (Lean/Coq/Isabelle/HOL Light/Metamath).

2.2 A canonical hybrid pipeline

Figure 2 sketches the pipeline common to many hybrid systems: (1) *formalization* from text/diagram to constraints, (2) *search* for a proof plan or step sequence, (3) *verification* of each step in a symbolic environment, (4) optionally *natural language rendering* for human consumption.

3 A Taxonomy of LLM–Symbolic Geometry Verification

We propose a three-axis taxonomy:

Axis A: Role of the LLM.

- **LLM as parser:** translate natural language (and/or diagrams) into formal constraints or a DSL (cf. autoformalization (Wu et al., 2022)).
- **LLM as strategist:** propose lemmas, theorem applications, or auxiliary constructions to guide symbolic search (common in neuro-symbolic provers).
- **LLM as prover:** output formal proof scripts (Lean/Coq) directly, checked by a kernel (e.g. LLM provers built on LeanDojo (Yang et al., 2023)).
- **LLM as critic:** score, rerank, or refine candidate steps based on verifier feedback (e.g. solver-in-the-loop refinement (Ye et al., 2023; Olausson et al., 2023)).

Axis B: Proof artifact. Answer-only → informal proof → semi-formal trace → kernel-checked proof.

Axis C: Verification backend. Empirical checking → symbolic algebraic/synthetic provers → proof-assistant kernels.

This taxonomy helps clarify trade-offs: kernel-checked proofs maximize trust but are hard to produce; algebraic provers scale but return conditions and certificates that may be pedagogically opaque; empirical checks are easy but unsound as proof.

4 Symbolic Verifiers for Euclidean Geometry

This section reviews the symbolic tools commonly used to validate geometric claims.

4.1 Algebraic methods

Algebraic geometry theorem proving translates geometric predicates into polynomial equations and uses elimination (e.g. Wu's method or Gröbner bases) to prove that hypotheses imply the conclusion, often generating *non-degeneracy conditions* (NDGs) (Marić et al., 2012). Algebraic methods are powerful and fast in many settings, but their outputs may be less interpretable than classical synthetic proofs.

4.2 Synthetic and semi-algebraic methods

Synthetic approaches aim to produce human-readable proofs. The *area method* is a prominent semi-algebraic technique that produces concise, readable proofs for constructive geometry (Janičić et al., 2012). Tools such as GCLC integrate visualization with theorem proving and can support multiple methods, including the area method (Janičić, 2006).

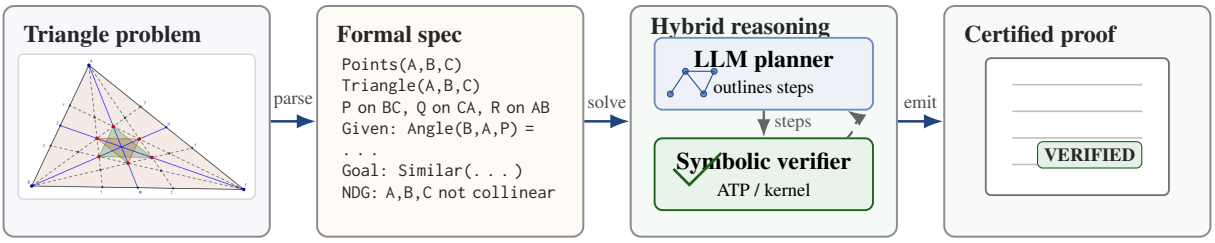


Figure 1: A high-level view of verified geometry reasoning: a natural-language problem with a diagram is translated into a formal specification, solved with LLM-guided symbolic reasoning, and emitted as a certified proof artifact.

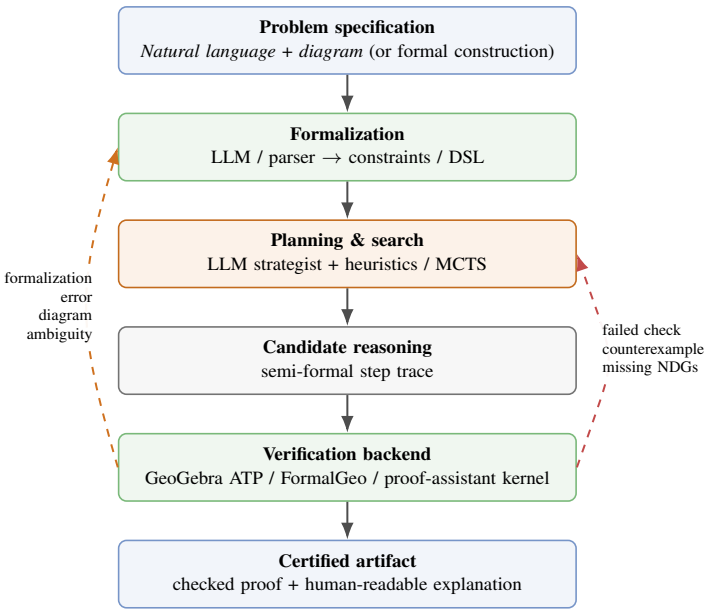


Figure 2: A canonical LLM–symbolic workflow for geometry: propose, check, and repair.

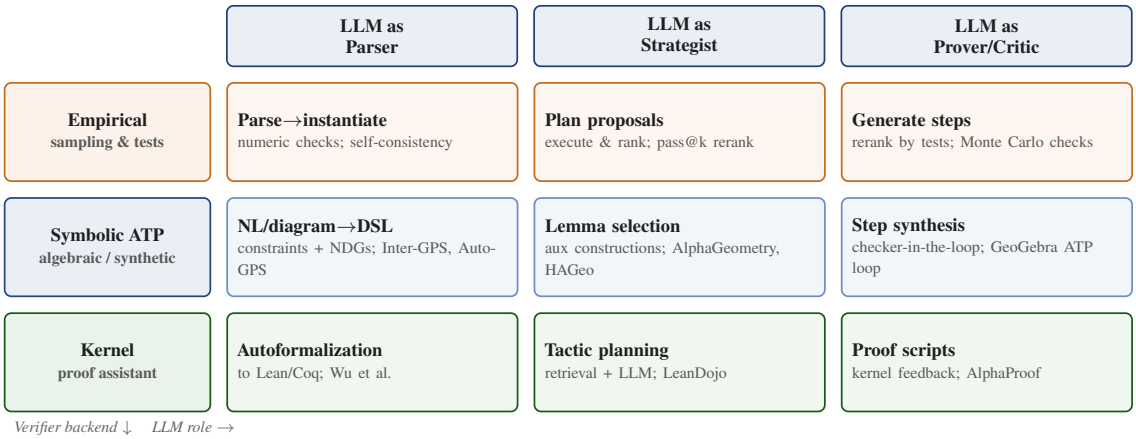


Figure 3: A compact taxonomy of LLM–symbolic geometry systems across LLM roles and verification backends, with representative examples.

4.3 Dynamic geometry environments and embedded provers

Dynamic geometry systems (DGS) provide interactive construction and visualization; some also embed proving functionality. GeoGebra has incorporated automated theorem proving features and a portfolio of provers, including Gröbner-basis-based proving and connections to external provers (e.g. OpenGeoProver for Wu/area) (Botana et al., 2015; Kovács and Sólyom-Gecse, 2016). Recent work

also explores automated *discovery* of geometric properties within GeoGebra constructions (Kovács and Yu, 2022).

4.4 Formal proof assistants

Proof assistants provide the strongest correctness guarantees via small trusted kernels, at the cost of higher formalization effort. For geometry, there is a long line of work on formalizing Euclidean axioms (e.g. Tarski/Hilbert) and connecting automation tactics to kernels. Maric et al. (Marić et al., 2012) describe Isabelle/HOL formalization efforts aimed at bridging algebraic methods and synthetic geometry, enabling verified automation tactics.

5 Representative Geometry Solvers and LLM-Augmented Systems

Table 1 summarizes representative systems along our taxonomy axes. Some entries are not LLM systems by themselves; we include them because they serve as formalizers, verifiers, search environments, or benchmarks inside LLM-augmented geometry pipelines. Figure 4 provides an additional visual summary of these roles.

6 Historical Timeline and Milestones

Figure 5 summarizes a few influential milestones that shaped the modern landscape of verified and neuro-symbolic geometry reasoning, spanning early multimodal QA, embedded automated theorem proving in dynamic geometry software, formalized geometry environments, and Olympiad-level neuro-symbolic solvers.

6.1 Multimodal problem understanding: text and diagram

NLP-oriented geometry systems often start with the *formalization bottleneck*: extracting entities, relations, and constraints from language and diagrams. GeoS (Seo et al., 2015) pioneered combining text and diagram interpretation for SAT geometry. Inter-GPS (Lu et al., 2021) later emphasized *interpretable* symbolic reasoning with a formal language representation, while GeoQA (Chen et al., 2021) provided a benchmark with annotated programs for multimodal numerical reasoning. Recent systems such as AutoGPS aim to jointly learn formalization and deductive reasoning with tight feedback loops between the modules (Ping et al., 2025).

6.2 Formalized geometry environments for verifiable traces

FormalGeo proposes a consistent formal plane geometry system and datasets that support traceable, verifiable solutions (Zhang et al., 2023). Within such environments, learning can focus on theorem selection and search policy while the environment enforces logical validity. FGeo-DRL builds an RL+MCTS agent that operates in the FormalGeo environment and yields readable, verifiable deductive solutions (Zou et al., 2024).

6.3 Olympiad-level provers and auxiliary construction

AlphaGeometry introduced a neuro-symbolic solver trained on synthetic data that outputs proof-like derivations verified by a symbolic engine (Trinh et al., 2024). AlphaGeometry2 reports expanded language coverage and improved search on IMO geometry sets, and it was part of a system that reached silver-medal standard on IMO-2024 (Chervonyi et al., 2025; Google DeepMind, 2024). Complementary to learned components, purely heuristic deductive approaches have also shown strong performance: HAGeo proposes efficient auxiliary constructions and reports high solving rates on Olympiad benchmarks without neural inference (Duan et al., 2025).

6.4 General neurosymbolic patterns for faithful reasoning

Although not geometry-specific, several ACL-relevant paradigms inform verified geometry pipelines. Program-aided LMs (PAL) offload execution to interpreters (Gao et al., 2022); SatLM translates problems into declarative constraints and uses SAT solving (Ye et al., 2023); LINC translates premises and conclusions into first-order logic and delegates deduction to logic provers (Olausson et al., 2023). Autoformalization systems translate informal mathematics into formal statements for proof assistants (e.g. Isabelle/HOL) (Wu et al., 2022). These frameworks highlight a recurring motif: use LLMs for *semantic parsing and proposal*, but use symbolic engines for *sound inference*.

7 Datasets, Benchmarks, and Evaluation

7.1 Geometry datasets

Table 2 lists major datasets spanning text, diagrams, and formal proof traces.

System	Input	Output	Verifier	Notes / LLM Role
GeoS (Seo et al., 2015)	Text + raster diagram	Answer (SAT-style)	Geometric solver (symbolic)	Early end-to-end pipeline; optimization-based parsing and diagram interpretation.
Inter-GPS (Lu et al., 2021)	Text + diagram	Answer + interpretable steps	Symbolic rule-based reasoning	Neural perception + formal language + theorem-driven symbolic reasoning.
GeoQA (Chen et al., 2021)	Text + diagram	Answer + program	Program executor	Dataset + neural solvers; emphasizes multimodal numerical reasoning.
FormalGeo (Zhang et al., 2023)	Formal language (often derived from text)	Stepwise proof trace	Formal system checker	Formalized predicate/theorem library enabling traceable, verifiable solutions.
FGeo-DRL (Zou et al., 2024)	FormalGeo environment	Stepwise proof trace	FormalGeo checker	RL + MCTS for theorem selection and search in a formal environment.
AutoGPS (Ping et al., 2025)	Text + diagram	Minimal stepwise proof	Deductive symbolic reasoner	Multimodal formalizer + deductive reasoner; emphasizes stepwise coherence.
AlphaGeometry (Trinh et al., 2024)	Domain-specific language	Proof (synthetic-style)	Symbolic engine	Neural + symbolic; trained on synthetic theorems; no human demonstrations.
AlphaGeometry2 (Chervonyi et al., 2025)	Extended DSL / partial NL	Proof	Symbolic engine	Expanded language coverage and improved search; used in IMO-2024 silver system.
HAGeo (Duan et al., 2025)	Formal geometry benchmark	Proof / derivation	Deductive engine (no NN)	Heuristic auxiliary constructions; strong performance without neural inference.
GeoGebra ATP (Botana et al., 2015; Kovács and Sólyom-Gecse, 2016)	Interactive construction	True/False + NDGs	Portfolio provers	DGS interface for algebraic/synthetic proving; useful as step checker in hybrid workflows.

Table 1: Representative geometry systems and their verification backends.

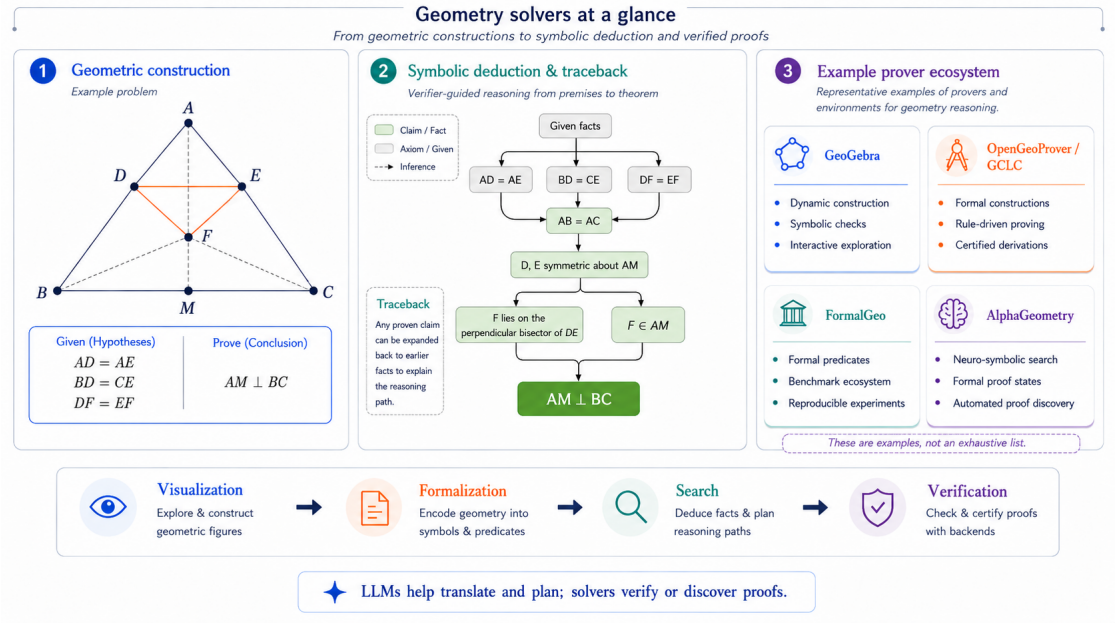


Figure 4: Additional at-a-glance comparison of representative geometry tools and solver families. This visual complements Table 1 by summarizing common stages in LLM-augmented geometry workflows and by showing a representative prover ecosystem; the tools in the right panel are examples rather than an exhaustive list.

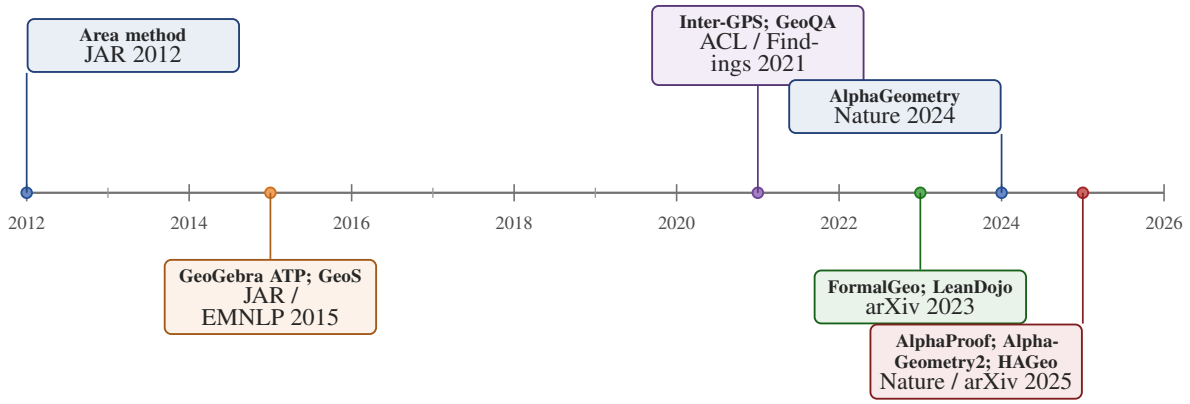


Figure 5: A non-exhaustive timeline of key milestones in multimodal geometry problem solving and verified/neuro-symbolic theorem proving.

Dataset	Modality	Notes
GeoS (Seo et al., 2015)	Text+diagram	SAT-style geometry QA; early end-to-end benchmark combining textual and diagrammatic parsing.
GeoQA (Chen et al., 2021)	Text+diagram	Annotated programs for geometric question answering; emphasizes multimodal numerical reasoning.
Geometry3K (Lu et al., 2021)	Text+diagram	Large benchmark used in Inter-GPS and follow-up work; supports interpretable symbolic reasoning.
FormalGeo7K / IMO (Zhang et al., 2023)	Formal	Formalized predicates and theorem libraries with stepwise proof traces enabling verifiable deduction.
IMO-30 (Trinh et al., 2024)	Formal	Standard Olympiad geometry evaluation subset used for neuro-symbolic benchmarking.
HAGeo-409 (Duan et al., 2025)	Formal	Expanded benchmark with human-assessed difficulty levels and emphasis on auxiliary constructions.

Table 2: Selected datasets and benchmarks for geometry reasoning and verification.

7.2 Metrics beyond answer accuracy

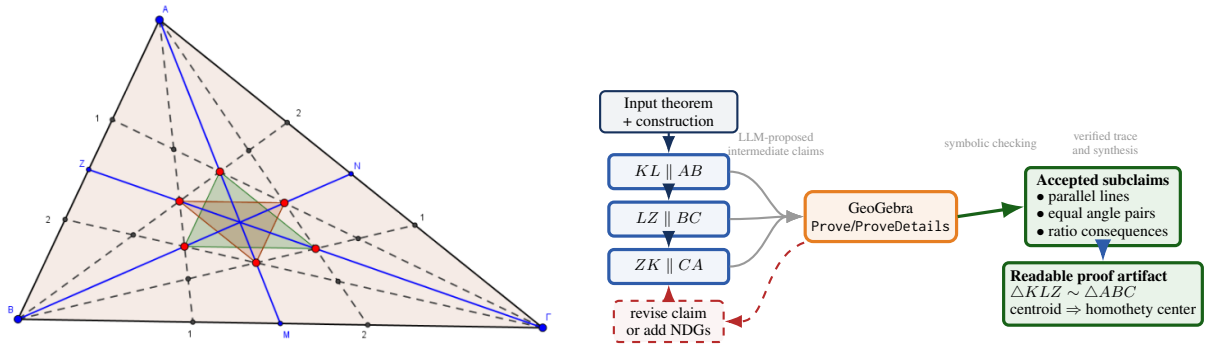
For verification-oriented systems, answer accuracy is insufficient. The checklist below consolidates practices already common in trace-based geometry and verifier-in-the-loop work: executable proof traces, explicit theorem applications, NDG reporting, and reproducibility of verifier settings (Zhang et al., 2023; Zou et al., 2024; Ping et al., 2025; Botana et al., 2015; Trinh et al., 2024). We recommend reporting:

- **Step validity rate:** fraction of generated steps accepted by the verifier.
- **Proof completeness:** whether a full chain from hypotheses to goal is produced.
- **NDG handling:** whether non-degeneracy conditions are made explicit and interpretable.
- **Minimality and readability:** proof length, lemma reuse, and human evaluation.
- **Robustness:** invariance to paraphrases, diagram perturbations, or irrelevant distractors.
- **Reproducibility:** open code/data, deterministic seeds, and clear verifier settings.

8 Case Study: Verifying a Triangle Trisection Generalization with Tool-Augmented LLMs

A representative use-case is to treat an LLM as a *proof planner* that proposes a structured outline, then validate each step with a symbolic geometry prover. For instance, the case study of Tzachristas and Tzachristas (2026) uses a triangle-side trisection configuration to connect natural-language planning, GeoGebra queries, and verified subclaims. The implementation of the case-study workflow was inspired by agentic workflow frameworks such as Hermes Agent and OpenClaw, which organize task execution around persistent agents, tool use, memory, and reusable skills (Nous Research, 2026; OpenClaw Contributors, 2026). Figure 6 adds a visual companion to the textual case-study description.

Concretely, the pipeline input is a theorem statement plus a GeoGebra construction for $\triangle ABC$



(a) Triangle-side trisection configuration from the MICOM case study.

(b) Original schematic of the generate–check–repair loop for verifier-guided proof construction.

Figure 6: Additional case-study visualization. Left: the triangle-side trisection construction discussed by Tzachristas and Tzachristas (2026). Right: an original, redrawn schematic showing how LLM-proposed subclaims are checked with GeoGebra and assembled into a readable proof artifact; this preserves the general deduction-and-traceback idea without reusing the source-uncertain image.

with side/intersection points K, L, Z and the target that the constructed inner triangle is similar and homothetic to ABC (with the centroid as homothecy center). The intermediate structures are verifier-addressable Boolean claims, such as $KL \parallel AB$, $LZ \parallel BC$, and $ZK \parallel CA$, issued through commands of the form `Prove(AreParallel(...))` and optionally expanded with `ProveDetails` / NDG output. The output is a checked trace of accepted subgoals plus a human-readable proof deriving similarity and the homothecy claim. In such a workflow:

1. The LLM proposes a sequence of intermediate claims (collinearity, parallelism, ratios, or Ceva/Menelaus-style equalities).
2. Each claim is translated into a form accepted by a verifier (e.g. GeoGebra's `Prove` /`ProveDetails` commands).
3. Failed steps trigger refinement: the LLM revises the claim or introduces missing NDG conditions.
4. A final verified step trace is rendered into human-readable proof text.

This "generate–check–repair" loop aligns naturally with neurosymbolic paradigms in NLP, and can be evaluated via step validity, completeness, and interpretability.

9 Open Challenges and Future Directions

Multimodal autoformalization. Moving from text+diagram to a formal specification remains brittle; robust datasets with aligned language, diagram, and formal constraints are scarce despite progress in GeoS, Inter-GPS, GeoQA, and AutoGPS-style pipelines (Seo et al., 2015; Lu et al., 2021; Chen et al., 2021; Ping et al., 2025).

Auxiliary construction and search control. Olympiad geometry often hinges on creative constructions. Balancing learned heuristics, symbolic search, and interpretability remains an open problem in AlphaGeometry-style and heuristic-only systems (Trinh et al., 2024; Chervonyi et al., 2025; Duan et al., 2025).

Non-degeneracy conditions (NDGs). Algebraic provers generate NDGs; mapping them to human-friendly geometric conditions and ensuring they are tracked across proof steps is essential for trustworthy proofs (Marić et al., 2012; Botana et al., 2015).

Certified yet readable proofs. Producing *kernel-checked* proofs that are also pedagogically meaningful is an ongoing challenge. Bridging proof-assistant scripts, prover certificates, and classical Euclidean style is a key opportunity (Wu et al., 2022; Yang et al., 2023; Kovács and Sólyom-Gecse, 2016).

Evaluation culture. We encourage the community to report verifier settings, failure modes, and ablations that isolate formalization errors from reasoning errors, following the traceability emphasis of FormalGeo/FGeo-DRL and the reproducible benchmark style of recent solvers (Zhang et al., 2023; Zou et al., 2024; Duan et al., 2025).

10 Conclusion

Hybrid LLM–symbolic systems offer a promising path from natural language and diagrams to verified geometry proofs. The strongest systems tightly couple learned proposal mechanisms with symbolic or kernel-level verification, enabling traceable derivations and reducing hallucinations. This survey provided a taxonomy of such systems, reviewed symbolic verifiers and geometry datasets, and argued for evaluation protocols that prioritize step-level correctness and reproducibility.

Limitations

This survey emphasizes verification-oriented pipelines and may omit purely neural answer-only systems when they do not expose verifiable intermediate artifacts. The field is moving rapidly; despite including recent work up to early 2026 in our bibliography, some contemporaneous results may be missing.

Ethics Statement

We do not anticipate direct harmful applications from surveyed methods. However, educational deployments should clearly communicate the difference between plausible explanations and verified proofs, and should avoid over-reliance on unverified LLM output. Releasing datasets should respect copyright constraints for sourced problem statements and diagrams.

Acknowledgements

The authors thank the open research community for making datasets, code, and preprints publicly available. We acknowledge the use of automated writing tools for language editing and clarity improvements during the preparation of this manuscript. All technical content, interpretations, and final decisions remain the responsibility of the authors.

References

Francisco Botana, Markus Hohenwarter, Predrag Janičić, Zoltán Kovács, Ivan Petrović, Tomás Recio, and Simon Weitzhofer. 2015. Automated theorem proving in GeoGebra: Current achievements. *Journal of Automated Reasoning*, 55:39–59.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Yuri Chervonyi, Trieu H. Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Junsu Kim, Vikas Verma, Quoc V. Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with AlphaGeometry2. arXiv:2502.03544.

Boyan Duan, Xiao Liang, Shuai Lu, Yaoxiang Wang, Yelong Shen, Kai-Wei Chang, Ying Nian Wu, Mao Yang, Weizhu Chen, and Yeyun Gong. 2025. Gold-medal-level olympiad geometry solving with efficient heuristic auxiliary constructions. arXiv:2512.00097.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. PAL: Program-aided language models. arXiv:2211.10435.

Google DeepMind. 2024. Ai achieves silver-medal standard solving international mathematical olympiad problems. Blog post.

Predrag Janičić. 2006. GCLC—a tool for constructive euclidean geometry and more than that. In *International Congress on Mathematical Software (ICMS)*, pages 58–73.

Predrag Janičić, Julien Narboux, and Pedro Quaresma. 2012. The area method. *Journal of Automated Reasoning*, 48(4):489–532.

Zoltán Kovács and Csilla Sólyom-Gecse. 2016. Geogebra tools with proof capabilities. *arXiv preprint arXiv:1603.01228*.

Zoltán Kovács and Jonathan H. Yu. 2022. Automated discovery of geometrical theorems in GeoGebra. In *Proceedings 10th International Workshop on Theorem Proving Components for Educational Software (THedu'21)*, volume 354 of *Electronic Proceedings in Theoretical Computer Science*, pages 1–12.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. 2024. A survey on deep learning for theorem proving. arXiv:2404.09939.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6774–6786.

Filip Marić, Ivan Petrović, Danijela Petrović, and Predrag Janičić. 2012. Formalization and implementation of algebraic methods in geometry. In *Theorem Proving Components for Educational Software (THedu'11), EPTCS 79*, pages 63–81.

Nous Research. 2026. Hermes agent: The self-improving ai agent built by nous research. https://github.com/NousResearch/hermes-agent. Open-source software repository. Accessed 2026-05-14.

Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. arXiv:2310.15164.

OpenClaw Contributors. 2026. Openclaw: Personal ai assistant. https://github.com/openclaw/openclaw. Open-source software repository. Accessed 2026-05-14.

Bowen Ping, Minnan Luo, Zhuohang Dang, Chenxi Wang, and Chengyou Jia. 2025. Autogps: Automated geometry problem solving via multimodal formalization and deductive reasoning. arXiv:2505.23381.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, Thang Luong, et al. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Ioannis Tzachristas and Georgios Tzachristas. 2026. Proofing techniques in geometry using LLMs and GeoGebra: A case study of the triangle-side trisection theorem. In *7th International Congress on Mathematics (MICOM 2026)*, Thessaloniki, Greece. Mathematical Society of South-Eastern Europe (MASSEE). Presentation slides.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. In *Advances in Neural Information Processing Systems*.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. Leandojo: Theorem proving with retrieval-augmented language models. arXiv:2306.15626.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023. SatLM: Satisfiability-aided language models using declarative prompting. In *Advances in Neural Information Processing Systems*.

Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Qike Huang, Xiaoxiao Jin, Yanjun Guo, Chenyang Mao, Yang Li, Zhe Zhu, Dengfeng Yue, Fangzhen Zhu, Yifan Wang, Yiwen Huang, Runan Wang, Cheng Qin, Zhenbing Zeng, Shaorong Xie, Xiangfeng Luo, and Tuo Leng. 2023. Formalgeo: An extensible formalized framework for olympiad geometric problem solving. arXiv:2310.18021.

Jia Zou, Xiaokai Zhang, Yiming He, Na Zhu, and Tuo Leng. 2024. Fgeo-drl: Deductive reasoning for geometric problems through deep reinforcement learning. arXiv:2402.09051.

Open Problems Solved by LLMs?
A Survey of Verifiable Mathematical Discovery

Ioannis Tzachristas [1,2*], **Georgios Tzachristas** [1,3*], **Aifen Sui** [1†]
[1]Huawei European Research Institute
[2]Technical University of Munich, Germany
[3]National Technical University of Athens, Greece

Abstract

Recent years have produced a small but rapidly growing set of results where Large Language Models (LLMs)—usually embedded in a search-and-verification loop—advance the state of the art on problems previously regarded as "open" in the pragmatic sense of lacking a best-known construction, bound, or proof certificate. This paper surveys that emerging line of work with a Big Picture emphasis: *what makes these successes possible, what should count as "solved", and what design patterns generalize?* We (i) propose an evidence ladder for interpreting "LLM solved an open problem" claims, (ii) map mathematical subfields by difficulty dimensions that matter for LLM-based discovery, (iii) curate a timeline of key breakthroughs leading to verifiable discovery systems, and (iv) synthesize the techniques and frameworks—tool use, retrieval, search, and verification—that repeatedly appear in successful case studies. We give particular attention to formal-methods backends common in security and verification contexts, including Linear Temporal Logic (LTL) and Satisfiability Modulo Theories (SMT) solvers, as scalable middle-layer verifiers between lightweight tests and proof assistants. We close with an evaluation and reproducibility checklist aimed at making the next wave of claims easier to trust, reproduce, and build upon, while separating peer-reviewed or certificate-backed results from fast-moving community reports that are useful signals but not yet stable evidence.

1 Introduction

The phrase *"LLMs solve open problems"* is simultaneously exciting and misleading. Exciting, because in a few settings LLM-driven systems have produced artifacts that are *objectively* better than prior best-known ones: new extremal combinatorial constructions, improved numerical bounds, or certified inequalities. Misleading, because "open problem" conflates many different notions of difficulty, and because LLMs almost never act alone. Most credible successes are not *one-shot* "proof writing"; they are *closed-loop discovery* systems that turn an LLM into a proposal generator whose outputs are filtered by an external evaluator.

Mathematics is a particularly revealing arena for this phenomenon. Recent perspective work on generative modeling for mathematical discovery makes a complementary systems-level case for treating learned models as proposal generators in checkable search loops (Ellenberg et al., 2025). Unlike many NLP tasks, math often comes with (or can be engineered to have) strong correctness tests: a candidate construction can be checked against constraints; a numerical bound can be verified by code; a formal proof can be validated by a proof assistant. This makes math a natural testbed for the broader scientific question: *when do generative models become reliable discovery tools?*

Survey scope. We focus on LLM-driven systems that (a) claim progress on problems where the status quo was "best known" or unknown in the literature or community benchmarks, and (b) provide an evaluator or certificate that can in principle be checked independently. We intentionally deemphasize purely anecdotal "ChatGPT proved X" stories unless they culminate in an externally verifiable artifact.

Contributions.

- **Evidence ladder.** We define tiers of evidence for "solved" and connect them to reproducibility expectations.
- **Hardness map.** We analyze which dimensions make math hard for LLM systems and argue that *verifiability* is the dominant factor in current successes.

*Equal contribution.
†Correspondence: aifen.sui@huawei.com.

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 10–21

- **Breakthrough timeline.** We summarize key milestones from transformers to tool use to evolutionary search that culminate in verifiable discovery.
- **Technique taxonomy.** We distill recurring frameworks—LLM+search+verifier, retrieval over formal libraries, test-time adaptation—and outline how they interact.
- **Community signals.** We record how non-paper announcements, problem wikis, and social-media discussions should be handled without confusing attention with verification.

How the axes fit together. The figures and taxonomies below are intended as coordinates of one reference architecture, not as independent lists. Figure 1 asks how strong the evidence is; Figure 3 asks whether the target domain admits a cheap verifier; Figure 2 locates that verifier by rigor and cost; and Figure 5 shows how generation, search, and checking interact. The timeline in Figure 4 is therefore best read as a sequence of improvements to components of the same loop, with milestone citations and component labels made explicit in Tables 4–5.

2 What should count as "an open problem solved by an LLM"?

A central aspect of this topic is definitional. In everyday mathematical practice, an "open problem" can mean anything from "nobody knows the exact answer" to "the best known bound might be improvable" to "no proof exists." Moreover, the role of the LLM can range from minor assistance (e.g., drafting exposition) to being the main generator of candidates in an automated loop.

2.1 A practical evidence ladder

Figure 1 summarizes a spectrum of claims. The ladder does *not* rank mathematical importance; it ranks how directly a reader can verify that the system achieved a new result.

We operationalize the ladder as tiers (Table 1). The goal is not to police language, but to make papers easier to compare and reproduce.

2.2 Community-reported claims

A practical complication is that influential claims often appear first on Mathstodon, GitHub wikis, LinkedIn, Hacker News, personal pages, or shared chat transcripts, before a conventional paper exists. We treat these as *community signals*: useful for discovering what problems people are discussing, but not sufficient evidence by themselves. For the community examples below, we used the Erdős wiki and Problem #728 forum as problem-status sources, Mathstodon and Hacker News as discussion fora, and LinkedIn posts as examples of social amplification rather than as independent validators (Terence Tao and contributors, 2026; Erdős Problems community, 2026; Tao, 2026; Hacker News contributors, 2026; Boland, 2026; Hoefler, 2026). Recent examples include the Erdős-problems community wiki, which explicitly labels AI contributions as full, partial, or incorrect; the public discussion of Erdős Problem #728 and related factorial-divisibility questions; and Knuth's "Claude's Cycles" note on decomposing a directed toroidal grid into Hamiltonian cycles (Terence Tao and contributors, 2026; Tao, 2026; Sothanaphan, 2026; Knuth, 2026; Morrison and contributors, 2026). A parallel, more curated example is the GPT-5.2 learning-curve monotonicity case, where a public company post pointed readers to a technical write-up on an open statistical-learning question (OpenAI, 2025). LinkedIn posts and other reposts amplified these examples quickly, but their role in this survey is pointers to the problem and artifact, not independent validation (Boland, 2026; Hoefler, 2026).

2.3 Why verification dominates

Across surveyed systems, the strongest results are those where verification is cheap relative to generation. This echoes earlier work on *verifier-based* scaling (e.g., generating many candidates and selecting with a verifier) in mathematical reasoning benchmarks such as GSM8K and MATH (Cobbe et al., 2021; Hendrycks et al., 2021). In discovery settings, the same idea becomes a closed-loop optimization: generate, evaluate, mutate, repeat.

2.4 Verifier spectrum: rigor *vs.* throughput

Even when a problem is "verifiable," verifiers differ sharply in (i) *soundness/rigor* and (ii) *evaluation cost per candidate*. This trade-off strongly shapes which mathematical areas see early progress: systems gravitate toward regimes where they can test *many* candidates per unit compute while retaining meaningful correctness guarantees.

Figure 2 summarizes a practical landscape. In practice, the main "sweet spot" for present-day open-problem progress is typically *certificates and deterministic checkers* (Tier B/C): they are much more rigorous than heuristic scoring, and far

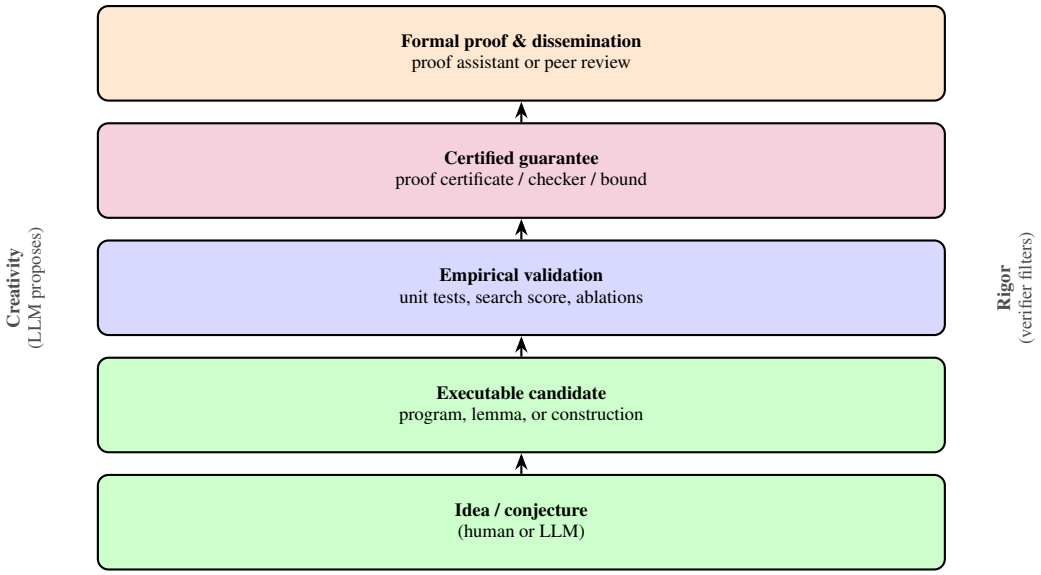


Figure 1: An evidence ladder for interpreting "LLM solved an open problem" claims. The most trustworthy claims reach the upper rungs via formal verification or independently checkable certificates.

Tier	What the paper provides
A (formal)	A formal proof checked by a proof assistant (e.g., Lean/Coq/Isabelle), or a proof-carrying artifact where verification is fully automated.
B (certified)	A machine-checkable certificate (e.g., explicit construction + deterministic checker, inequality certificate, or reproducible code proving a bound).
C (reproducible best-known)	A new best-known result backed by an open evaluator, strong baselines, and ablations; verification may still rely on extensive computation but is independently runnable.
D (suggestive)	Plausible conjectures, heuristics, or partial progress without a verifiable certificate; valuable, but not "solved."

Table 1: Evidence tiers for "open problem" claims. This survey focuses on A–C.

cheaper than full proof-assistant search.

3 Which area of mathematics is "hardest"?

The question "which area of math is the hardest to solve" is underspecified: hardest for humans, for automated provers, or for LLM-based systems? Here we answer it in a way that is actionable for system design: *hardest for current LLM-centric discovery loops*.

3.1 Difficulty dimensions that matter for LLM systems

One useful way to decompose this is:

- **Verifiability:** Is there a cheap, unambiguous checker? (unit tests, constraints, proof assistant)
- **Formalization barrier:** Can the object be represented in code or a formal language without huge overhead?
- **Search topology:** Does improvement require exploring an enormous combinatorial space with sparse rewards?
- **Abstraction depth:** Do solutions require introducing new concepts, definitions, or multi-lemma scaffolding?
- **Data availability:** Is there enough training signal (text, code, formal libraries) aligned with the target domain?

Figure 3 maps common mathematical areas along two dominant axes: verifiability and long-horizon abstraction/search. The takeaway is that what looks "hard" to humans can be comparatively

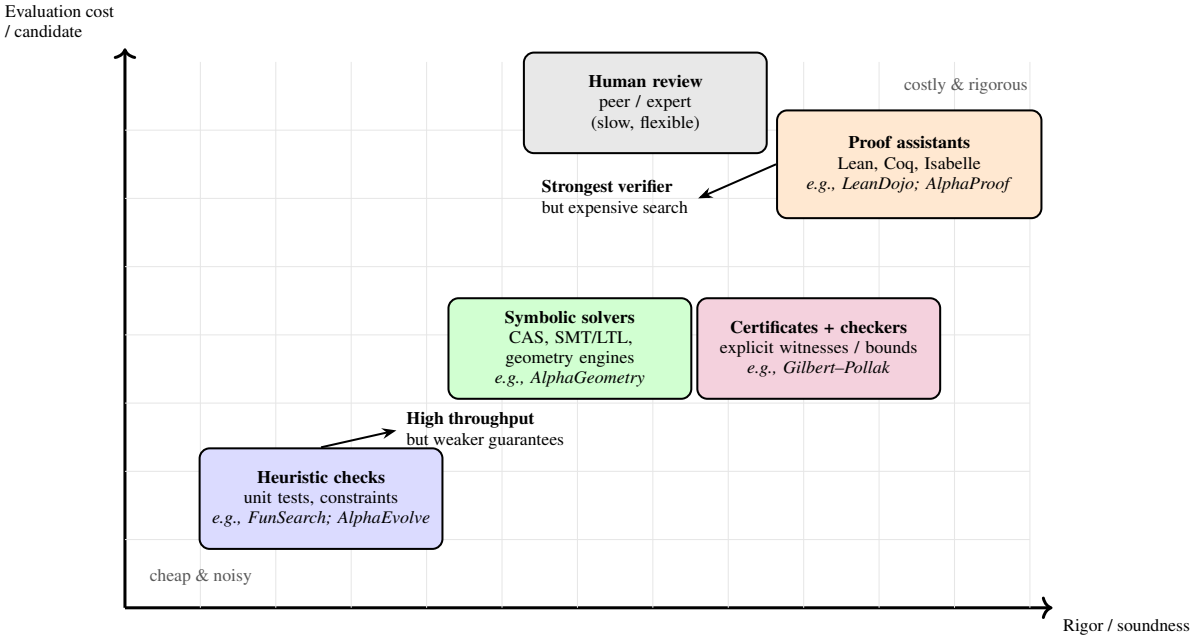


Figure 2: Verifier regimes plotted by approximate evaluation cost and rigor. Different "open-problem" case studies land in different parts of the space; the most scalable successes tend to combine *cheap checking* with *strong guarantees*.

Community signal	**Problem discussed and how we use it**
Erdős wiki / forum / Mathstodon	Problem #728 and related number-theoretic problems; upgrade only when a Lean file, checked write-up, or problem-page status supports the claim.
Knuth page / GitHub repository	Hamiltonian-cycle decomposition of a directed toroidal grid; a public construction plus linked Lean verification is treated differently from a bare claim.
LinkedIn / HN amplification	Useful for noticing fast-moving claims (e.g., Erdős #728, "Claude's Cycles"), but remains Tier D unless it points to a checker, proof, or citable artifact.

Table 2: How non-paper community signals are incorporated without lowering the evidence standard.

accessible to an LLM+verifier loop if the problem admits an objective checker.

3.2 A defensible answer

Under this lens, the most difficult areas for current LLM-based systems are those combining: (i) low verifiability (or extremely expensive checking), (ii) high abstraction depth, and (iii) weak alignment between natural language descriptions and executable representations. This includes large parts of analysis (e.g., PDE regularity), arithmetic geometry, and deep parts of algebraic number theory. By contrast, extremal combinatorics, constructive finite geometry, and some inequality/bounding problems are *comparatively* accessible because they can be posed as "find an object" with a deterministic checker.

4 A timeline of breakthroughs towards verifiable discovery

The emergence of credible "LLM solves open problem" results did not happen at once. It is the product of several strands: transformer scaling (Vaswani et al., 2017; Brown et al., 2020), better prompting and decoding for multi-step reasoning (Wei et al., 2022; Wang et al., 2023), math-specialized LMs (e.g., Minerva) (Lewkowycz et al., 2022), tool-use paradigms (Yao et al., 2023; Schick et al., 2023), and mature verification infrastructure (proof assistants and benchmark tooling).

How to read the timeline. The color of a milestone indicates which part of the closed loop it primarily strengthens: model capability, tool/verifier infrastructure, or the discovery loop itself. For example, chain-of-thought and self-consistency strengthen proposal generation and sampling; ReAct, Toolformer, LeanDojo, and SMT-style backends strengthen the executor/verifier interface; FunSearch, AlphaEvolve, and the Gilbert–Pollak work instantiate the full generate–check–search pattern. This is the link between the timeline, the verifier spectrum, and the framework diagram.

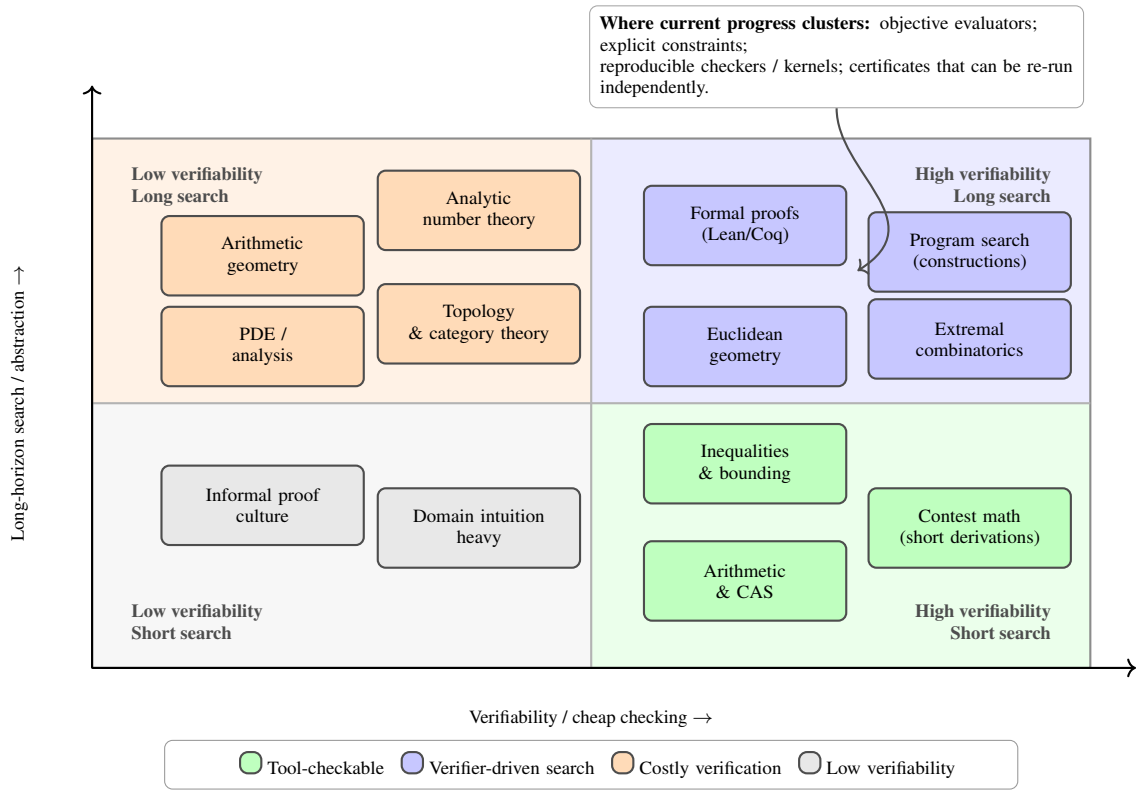


Figure 3: A qualitative hardness map for LLM-based discovery. The regime with *high verifiability* and *long-horizon search* (top-right) is where many credible "open-problem" improvements have appeared so far.

From reasoning to discovery. Chain-of-thought prompting and sampling-based decoding improved multi-step reasoning on math benchmarks (Wei et al., 2022; Wang et al., 2023), and specialized pretraining further improved mathematical competence (Lewkowycz et al., 2022). But benchmark reasoning is still far from open-ended discovery. The shift to discovery required (a) making the target object executable (code, proof assistant, or certificate), and (b) engineering a feedback loop so the model can iterate.

Specialized mathematical LMs. In parallel, several efforts trained or adapted LLMs specifically for mathematics. Minerva demonstrated that large-scale pretraining and math-focused data can substantially improve quantitative reasoning (Lewkowycz et al., 2022). Open models such as Llemma illustrate a complementary direction: releasing weights, training data mixtures, and tooling to enable reproducible research on mathematical LMs (Azerbayev et al., 2023b).

Formal theorem proving accelerates (and supplies stronger verifiers). Recent progress in formal reasoning systems highlights an important point for this survey: proof assistants are not only an evaluation tool but increasingly a *platform* for closed-loop training and search. HyperTree Proof Search introduced a transformer policy coupled with a structured proof search algorithm for neural theorem proving (Lample et al., 2022). Large-scale synthetic-data efforts such as DeepSeek-Prover show how to build formal training corpora in Lean 4 (Xin et al., 2024). AlphaProof demonstrates reinforcement-learning-based formal reasoning at a medal level on IMO problems (paired with AlphaGeometry2) (Hubert et al., 2026). While these systems are not usually presented as "solving open research conjectures," they materially expand the feasibility of Tier A claims.

5 Key techniques and frameworks

Across the literature, LLMs rarely "solve" research problems in a single shot. The most credible progress comes from *closed-loop systems* that turn an LLM into a proposal generator and use an external evaluator to filter, score, and iterate. This section distills the recurring building blocks.

5.1 Core design: generate–check–search

Most successful systems implement some variant of:

> *Generate many candidates; automatically evaluate them; use feedback to*

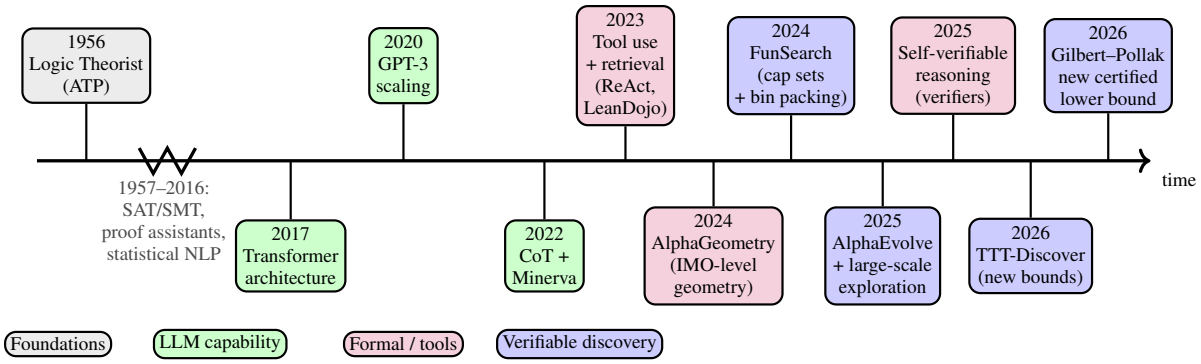


Figure 4: A compressed timeline of milestones leading to verifiable LLM-driven mathematical discovery. "Discovery" systems typically combine an LLM proposal mechanism with automated evaluation loops; Tables 4–5 attach citations and component labels to the milestones.

guide further generation.

At the simplest end, this looks like best-of-N selection with a verifier (Cobbe et al., 2021). At the discovery end, the "verifier" is an executable objective (constraints, tests, certificates, or proof kernels) and the outer loop is an explicit search algorithm.

5.2 Executors and tool use

Tool use makes candidate artifacts *executable*:

- **Program execution / simulators** provide precise feedback (scores, counterexamples) and are central to program-search discovery (e.g., FunSearch) (Romera-Paredes et al., 2024).
- **Symbolic engines** (CAS/SMT/geometry solvers) can act as fast verifiers and structured executors (e.g., AlphaGeometry) (Trinh et al., 2024).
- **Proof assistants** provide the strongest verifiers, enabling fully checkable Tier A claims, but introduce a formalization barrier; recent work increasingly treats them as a platform for learning and search (e.g., LeanDojo; AlphaProof) (Yang et al., 2023; Hubert et al., 2026).

General tool-use paradigms (prompted or learned) such as ReAct and Toolformer help connect LLM reasoning to external actions (Yao et al., 2023; Schick et al., 2023).

5.3 Formal-methods backends: LTL and SMT

Linear Temporal Logic (LTL) is relevant when the candidate object is not a single expression but a trace-producing process: a protocol, strategy, search controller, or transition system. Safety properties ("bad states never occur") and liveness properties ("a desired event eventually occurs") can be stated in LTL; bounded model checking then reduces finite-horizon LTL obligations to SAT/SMT instances and returns counterexamples that are directly useful as feedback to an LLM loop (Pnueli, 1977; Biere et al., 1999).

Satisfiability Modulo Theories (SMT) solvers generalize SAT with theories such as arithmetic, bit-vectors, arrays, and uninterpreted functions, making them a natural middle layer between unit tests and proof assistants (Barrett and Tinelli, 2018; de Moura and Bjørner, 2008). In the framework of Figure 5, an LLM can propose an invariant, lemma schema, constraint encoding, or program fragment; an SMT/LTL backend can then (i) reject it with a model or counterexample, (ii) certify bounded instances, or (iii) simplify the remaining proof obligation before a proof assistant checks the final theorem. This is why SMT/LTL are especially relevant to verifiable discovery: they turn informal mathematical intent into structured, relatively cheap obligations with actionable feedback.

5.4 Search controllers

When naive sampling is insufficient, systems add structure:

- **Evolutionary search** over programs or code edits (FunSearch; AlphaEvolve) (Romera-Paredes et al., 2024; Novikov et al., 2025).
- **Tree search** over structured partial solutions (e.g., proof states in formal math) (Lample et al., 2022).
- **(Test-time) adaptation / RL** when the evaluator provides dense feedback in a specific environment (e.g., AlphaProof; TTT-Discover) (Hubert et al., 2026; Yuksekgonul et al., 2026).

System / paper	Domain	What improved?	Tier	Verifier / artifact
FunSearch (Romera-Paredes et al., 2024)	Constructions	New best-known constructions on checkable problems	C	Program evaluator as fitness; executable code artifacts.
AlphaGeometry (Trinh et al., 2024)	Geometry	IMO-level geometry solving (checkable proofs)	B/C	Symbolic engine + neural guidance; proof traces checkable by solver.
AlphaEvolve (Novikov et al., 2025)	Mixed	Objective-driven rediscovery + improvements	C	Task-specific evaluators; tracked code edits and scores.
Gilbert–Pollak bound (Ke et al., 2026)	Optimization / geom.	New *certified* lower bound (Steiner ratio)	B	Certificate-checking pipeline; independently verifiable bound.
Erdős Problem #728 (Sothanaphan, 2026)	Number theory	Formalized resolution of a factorial-divisibility problem	A	Lean proof attributed to GPT-5.2 Pro + Aristotle; informal write-up and community tracking.
Claude's Cycles (Knuth, 2026)	Combinatorics	Hamiltonian-cycle decomposition for a directed toroidal grid	A/B	Public construction plus linked Lean verification and follow-up variants.

Table 3: Representative case studies illustrating how LLMs (in systems) can yield verifiable progress.

5.5 Retrieval over mathematical libraries

A repeated bottleneck—especially in formal settings—is *premise selection*: retrieving relevant lemmas and definitions. Open infrastructure such as ProofNet (informal↔formal pairs) and LeanDojo (interaction + corpora + benchmarks) makes retrieval and proof-search research more reproducible (Azerbayev et al., 2023a; Yang et al., 2023).

5.6 A reference architecture

Figure 5 summarizes a reusable "discovery loop." Different projects vary in their executors and verifiers, but the overall control flow is stable.

6 Representative case studies

Table 3 lists representative exemplars that reach evidence tiers A–C. The common thread is that the "result" is an independently checkable artifact: code, a certificate, or a formal proof object. For the Gilbert–Pollak entry, we distinguish the classical Steiner-minimal-tree conjecture from the newer verifier-backed lower-bound claim (Gilbert and Pollak, 1968; Ke et al., 2026).

7 Best-practices checklist for verifiable discovery claims

To make "open-problem" claims comparable and reproducible, we recommend reporting:

- **What was open:** precise statement of the target (problem family, n-range, constraints) and the prior best-known baseline.
- **Verifier/evaluator:** complete spec plus code (or proof kernel) needed to check candidates; tests against known instances.
- **Artifacts:** best candidate programs/proofs/certificates and scripts to reproduce the reported numbers.
- **Compute + search budget:** number of evaluations, LLM calls, wall-clock, and search hyperparameters.
- **Robustness + novelty:** reruns with different seeds; adversarial tests; leakage/overlap checks when relevant.
- **Human-in-the-loop disclosure:** manual steps required to reach the final artifact.
- **Community-claim trail:** for social-media or problem-wiki announcements, include the canonical problem page, transcript/checker link, and current status (full/partial/incorrect/pending).

8 Open challenges

Even in the verifiable regime, major challenges remain:

- **Beyond cheap verifiers:** most of mathematics lacks a fast checker; progress may require new in-

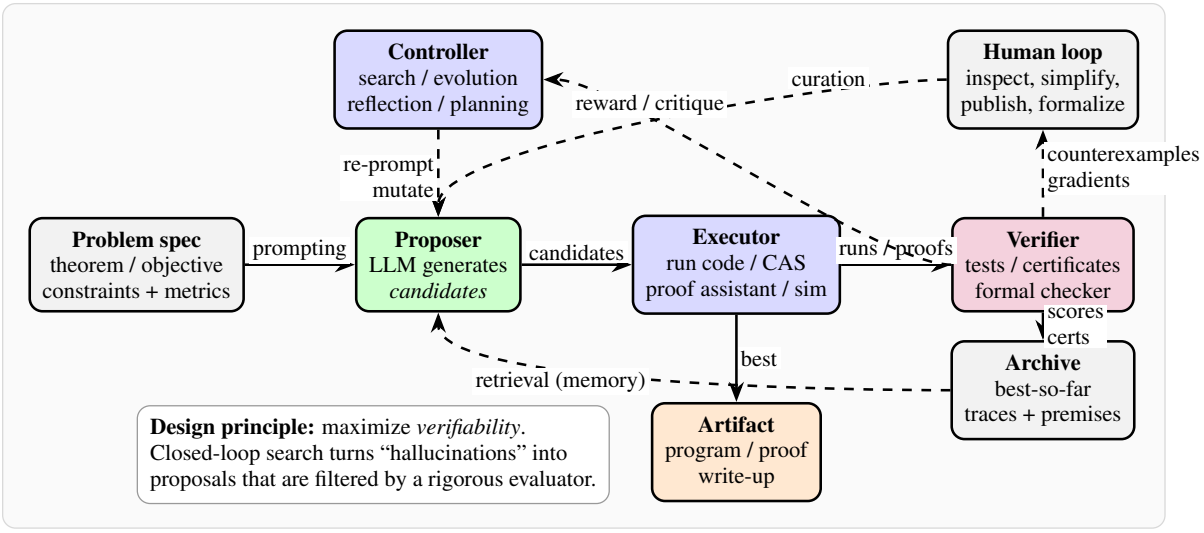


Figure 5: A generic closed-loop framework for verifiable LLM-driven discovery. The strongest "open problem" results appear when the verifier is strong *and* cheap, and when the archive (memory) enables rapid iteration.

termediate representations or hybrid human+AI workflows.

- **From finite-n to theorems:** turning best-known constructions/bounds into general statements (and proofs) is still hard.
- **Novelty and significance:** correctness is not the same as importance; we need better novelty tracking and human-facing inspection tools.
- **Community norms:** clearer standards for attribution, failure reporting, and maintaining open repositories of problems with evaluators.

9 Conclusion

Current systems remain limited, particularly outside domains with strong evaluators. Where strong evaluators exist, however, they can be powerful engines for exploring large search spaces and producing independently checkable improvements. Taken together, these case studies suggest that progress is driven less by "better prose reasoning" and more by *systems design*: executable representations, strong verifiers, retrieval over libraries, and structured search.

Limitations

This survey is a snapshot of a fast-moving area and should not be read as an exhaustive catalog of every community-reported claim. We deliberately emphasize cases with public artifacts, checkers, certificates, or formal proofs, which means that less-verifiable but potentially important mathematical assistance is underrepresented. The evidence ladder also compresses many distinctions: a formally checked result can still depend on problem formalization choices, and a reproducible computational result can still be sensitive to implementation details or search budgets. Finally, several recent examples are tracked through community pages, technical notes, and preprints whose status may evolve.

References

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023a. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. *CoRR*, abs/2302.12433.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023b. Llemma: An open language model for mathematics. *CoRR*, abs/2310.10631.

Clark Barrett and Cesare Tinelli. 2018. Satisfiability modulo theories. In Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith, and Roderick Bloem, editors, *Handbook of Model Checking*, pages 305–343. Springer.

Armin Biere, Alessandro Cimatti, Edmund M. Clarke, and Yunshan Zhu. 1999. Symbolic model checking without BDDs. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 1579 of *Lecture Notes in Computer Science*, pages 193–207. Springer.

Joseph Boland. 2026. ChatGPT 5.2 solves Erdos #728 with Terence Tao. LinkedIn post. Accessed May 12, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Leonardo de Moura and Nikolaj Bjørner. 2008. Z3: An efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 4963 of *Lecture Notes in Computer Science*, pages 337–340. Springer.

Jordan S. Ellenberg, Cristofero S. Fraser-Taliente, Thomas R. Harvey, Karan Srivastava, and Andrew V. Sutherland. 2025. Generative modeling for mathematical discovery. *CoRR*, abs/2503.11061.

Erdős Problems community. 2026. 728 discussion thread. Erdős Problems forum thread. Accessed May 12, 2026.

Bogdan Georgiev, Javier Gómez-Serrano, Terence Tao, and Adam Zsolt Wagner. 2025. Mathematical exploration and discovery at scale. *CoRR*, abs/2511.02864.

Edgar N. Gilbert and Henry O. Pollak. 1968. Steiner minimal trees. *SIAM Journal on Applied Mathematics*, 16(1):1–29.

Hacker News contributors. 2026. Erdős problem #728 was solved more or less autonomously by GPT-5.2 Pro. Hacker News discussion thread. Accessed May 12, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874.

Torsten Hoefler. 2026. Knuth's AI breakthrough: Claude Opus 4.6 solved open math problem. LinkedIn post. Accessed May 12, 2026.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, et al. 2026. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651:607–613.

Yisi Ke, Tianyu Huang, Yankai Shu, Di He, Jingchu Gai, and Liwei Wang. 2026. Towards solving the Gilbert–Pollak conjecture via large language models. *CoRR*, abs/2601.22365.

Donald E. Knuth. 2026. Claude's Cycles. Stanford Computer Science technical note. Revised Apr. 14, 2026; accessed May 12, 2026.

Guillaume Lample, Marie-Anne Lachaux, Thibaut Lavril, Xavier Martinet, Amaury Hayat, Gabriel Ebner, Aurélien Rodriguez, and Timothée Lacroix. 2022. HyperTree Proof Search for neural theorem proving. *CoRR*, abs/2205.11491.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *CoRR*, abs/2206.14858.

Kim Morrison and contributors. 2026. KnuthClaudeLean: Lean 4 formalization of Knuth's "Claude's Cycles". GitHub repository. Accessed May 12, 2026.

Alexander Novikov, Ngan Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *CoRR*, abs/2506.13131.

OpenAI. 2025. Advancing science and math with GPT-5.2. OpenAI blog. Accessed May 12, 2026.

Amir Pnueli. 1977. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 46–57.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36.

Nat Sothanaphan. 2026. Resolution of Erdős problem #728: A writeup of Aristotle's Lean proof. *CoRR*, abs/2601.07421.

Terence Tao. 2026. Recently, the application of AI tools to Erdős problems passed a milestone. Mathstodon post. Accessed May 12, 2026.

Terence Tao and contributors. 2026. AI contributions to Erdős problems. GitHub wiki. Edited May 10, 2026; accessed May 12, 2026.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. DeepSeek-Prover: Advancing theorem proving in LLMs through large-scale synthetic data. *CoRR*, abs/2405.14333.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Mert Yuksekgonul, Daniel Koceja, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, and Yu Sun. 2026. Learning to discover at test time. *CoRR*, abs/2601.16175.

A Supplementary timeline and explanatory material

The main paper provides a compressed timeline (Figure 4). For completeness, this appendix collects supplementary material in one place: (i) a multi-track visualization of milestones (Figure 6), (ii) a detailed timetable split across two eras (Tables 4 and 5), and (iii) an explanatory diagram summarizing common failure modes in discovery loops and practical mitigations (Figure 7).

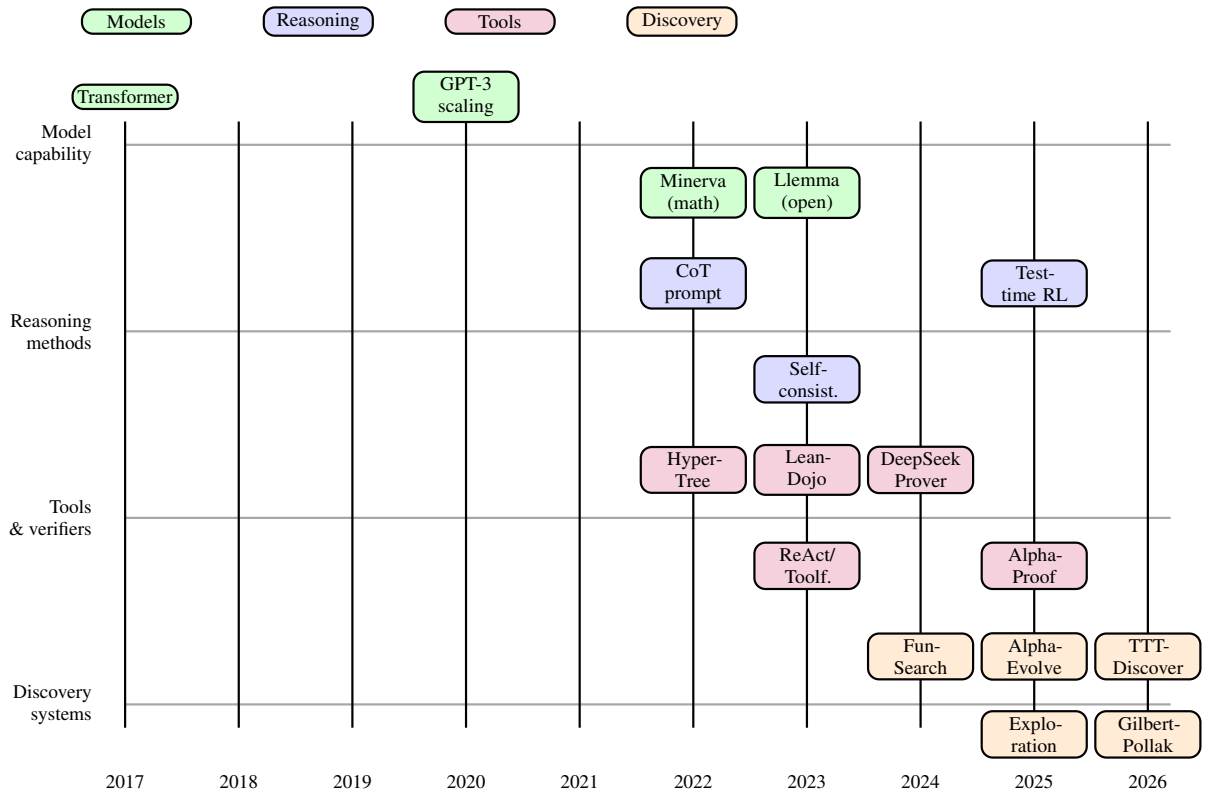


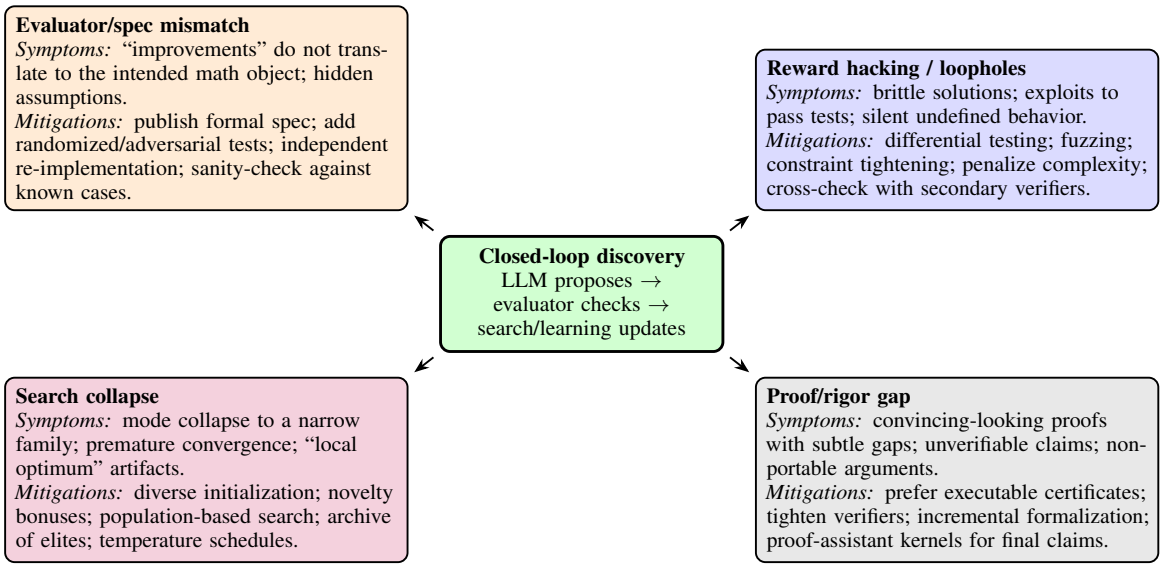
Figure 6: A multi-track timetable of milestones. Each track emphasizes a different ingredient needed for verifiable discovery: model capability, reasoning methods, tool/verifier integration, and discovery systems.

Year	Milestone	Component	Relevance to verifiable discovery
1956	Logic Theorist (ATP)	Foundations	Early template: symbolic search + correctness rules; foreshadows "generate + check" workflows.
1977	Linear Temporal Logic (Pnueli, 1977)	Formal methods	Specifies trace properties such as safety and liveness, which matter when LLM-generated artifacts are protocols or transition systems.
1999	Bounded model checking (Biere et al., 1999)	Verification	Reduces bounded LTL checking to SAT, producing counterexamples that can guide search.
2008/2018	SMT solvers (de Moura and Bjørner, 2008; Barrett and Tinelli, 2018)	Verification	Solvers over arithmetic, bit-vectors, arrays, and uninterpreted functions provide a cheap middle layer between tests and proof assistants.
2017	Transformer (Vaswani et al., 2017)	Models	Scaling-friendly architecture underpinning modern LLMs and in-context reasoning.
2020	GPT-3 scaling (Brown et al., 2020)	Models	Demonstrates few-shot learning and broad competence that later supports mathematical coding/reasoning.
2021	MATH dataset (Hendrycks et al., 2021)	Benchmarks	High-signal benchmark for multi-step symbolic reasoning; drives solver and verifier training.
2021	Training verifiers (Cobbe et al., 2021)	Verification	Establishes verifier-style scaling: sample many candidates, filter with learned/engineered verifiers.
2022	Chain-of-thought prompting (Wei et al., 2022)	Reasoning	Improves long-horizon reasoning via intermediate steps; foundation for later self-critique / reflection loops.
2022	Self-consistency (Wang et al., 2023)	Decoding	Sampling-based aggregation turns stochastic generation into a selection problem (proto-verifier scaling).
2022	Minerva (Lewkowycz et al., 2022)	Math LMs	Math-focused training increases quantitative competence, a key ingredient for program-level discovery.
2022	HyperTree Proof Search (Lample et al., 2022)	Formal search	Structured proof-state search coupled with learning, bridging theorem proving and modern ML.

Table 4: Timetable (foundations): milestones that enabled LLM-based math discovery and verifiable reasoning.

Year	Milestone	Component	Relevance to verifiable discovery
2023	ReAct (Yao et al., 2023)	Tool use	Couples reasoning with actions (tool calls), enabling executable feedback loops.
2023	Toolformer (Schick et al., 2023)	Tool learning	Shows how tool usage can be learned, not only prompted; supports scalable executor integration.
2023	ProofNet (Azerbayev et al., 2023a)	Autoformalization	Paired informal/formal data for evaluating statement translation and proof generation.
2023	LeanDojo (Yang et al., 2023)	Infrastructure	Standardizes interaction with Lean and retrieval; makes premise selection + proof search reproducible.
2023/2024	FunSearch (Romera-Paredes et al., 2024)	Discovery	LLM-guided evolutionary program search surpasses best-known constructions on checkable problems.
2024	AlphaGeometry (Trinh et al., 2024)	Neuro-symbolic	Strong symbolic engine paired with learned guidance solves hard geometry reliably.
2024	DeepSeek-Prover (Xin et al., 2024)	Formal data	Large-scale Lean 4 proof data, improving feasibility of RL/search in formal environments.
2025	AlphaProof (Hubert et al., 2026)	Formal RL	Medal-level formal reasoning; highlights strict verifiers + RL in formal math.
2025	AlphaEvolve (Novikov et al., 2025)	Discovery	Evolves codebases with evaluators; applies beyond math to algorithm discovery and optimization.
2025	Mathematical exploration at scale (Georgiev et al., 2025)	Discovery	Validates exploration at scale (successes + failures) to support scientific norms.
2026	TTT-Discover (Yuksekgonul et al., 2026)	Test-time RL	Learns at test time to optimize one environment; complements evolutionary approaches.
2026	Gilbert–Pollak lower bound (Ke et al., 2026)	Certified progress	Certificate-checking pipeline pushes a certified bound on a long-stagnant problem.
2026	Erdős Problem #728 (Sothanaphan, 2026; Terence Tao and contributors, 2026)	Formal/community	Illustrates the path from community report to Lean-formalized artifact and curated status tracking.
2026	Claude's Cycles (Knuth, 2026)	Formal/community	Public problem note, construction, and linked Lean verification for a Hamiltonian-cycle decomposition problem.

Table 5: Timetable (verifiable discovery era): milestones where LLMs are embedded in evaluator- and certificate-driven loops.



Survey takeaway: scalable discovery requires both *good feedback signals* and *defenses* against optimization pathologies.

Figure 7: Common failure modes in LLM-based discovery loops and practical mitigations.

Beyond Hallucination: Reframing LLM Quality Assessment as Task-Output Alignment

Michael Olaolu Arowolo[1] **Andrew Hoblitzell**[2]
[1]Xavier University of Louisiana [2]Purdue University
marowolo@xula.edu ahoblitz@purdue.edu

Abstract

Hallucination detection systems often operate under a flawed assumption: that any deviation from factual grounding is problematic, regardless of task context, modality, or cultural setting. A joke and a fabricated medical citation can look identical to a hallucination detector; only one is the problem. Through analysis of computational humor as a case study, we show that identical model behaviors warrant different evaluations depending on context. We propose reframing hallucination detection as task-output alignment assessment, organized along three axes: factual grounding, novelty, and risk tolerance. The reframing has implications for how the community evaluates multi-task LLMs and treats the boundary between creative and factual generation.

1 Introduction: Hallucination, Problem or Feature?

Ask ChatGPT for a philosophy joke and you might get: "a philosopher who orders a beer made of pure reason; the bartender serves him nothing, because it doesn't exist." It shows the classic hallucination signatures: a fabricated entity, a semantic leap, a low-probability continuation. The same patterns appear when models fabricate medical citations, as in the 2023 *Mata v. Avianca* sanction.[1] Bard's launch demo that February misattributed to JWST the first picture of a planet outside our solar system. Alphabet's market cap dropped about $100B that day. These behaviors can have radically different consequences.

There is a clear paradox here. Years of work have gone into hallucination detectors. Their job is to flag deviations from factual grounding. Survey papers have built task-specific taxonomies (Ji et al., 2023; Huang et al., 2025). In practice, the detectors still treat any deviation as uniformly problematic. Same detector, same confidence, for a poem metaphor and a fabricated medical fact. That cannot be right.

The field uses "hallucination" under a unifying assumption: that very different model behaviors map to the same kind of error. Through several case studies, we show that the surface-level overlap is misleading. The phenomena, fabrication in medical QA, invention in humor and fiction, incongruity in multimodal memes, and culturally situated exaggeration in multilingual storytelling, share signatures but require different evaluation frameworks.

This is a position paper. We argue that what the field calls "hallucination" is best understood as task-output misalignment, and that the unified framing is the source of the evaluation problems we identify.

1.1 Our Contribution

Hallucination is treated by the field as a monolithic category. We argue this is the wrong frame. Our alternative is task-output alignment assessment, with grounding, novelty, and risk tolerance as the dimensions we have found most useful:

1. Factual grounding requirements (low for creative writing, high for medical QA)
2. Novelty requirements (low for information retrieval, high for brainstorming)
3. Risk tolerance (low for safety-critical applications, higher for entertainment)

Why humor as the running case? Information-theoretic accounts show entropy and ambiguity predict humor (Westbury et al., 2016; Kao et al., 2016); however, what hallucination detectors flag as model uncertainty often is the creative space. Recent datasets including the New Yorker Caption Contest (Hessel et al., 2023), multilingual JOKER (Ermakova et al., 2022, 2023), and code-mixed humor (Khandelwal et al., 2018) show that grounding

[1]Two New York attorneys submitted a federal brief citing *Varghese v. China Southern Airlines* and several other cases that ChatGPT had invented; the judge imposed a joint $5,000 sanction on the attorneys and their firm (S.D.N.Y., June 2023).

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 22–30

requirements vary by modality and culture. And humor theory has long separated intentional semantic violations from accidental ones (Attardo, 2020; Loakman et al., 2025), which we take as the right frame for reasoning about "beneficial hallucination."

Section 2 identifies three critical problems with current hallucination framing. Section 3 proposes our task-output alignment framework. Section 4 discusses implications for responsible AI deployment.

The path forward is to swap universal hallucination minimization for task-conditional alignment.

2 Misframing: Three Critical Problems

2.1 Definitional Incoherence

Hallucination literature disagrees on what hallucination is, how to operationalise it, and how to measure it. The disagreement is substantive, not terminological.

Hallucination surveys often operate under a shared assumption: that the phenomenon they are taxonomizing is a single thing. Through cross-survey comparison, we show that the assumption fails. We propose treating the proliferation of categories not as a labeling problem but as a sign that hallucination is not a natural category (Ji et al., 2023; Maynez et al., 2020; Huang et al., 2025).

Take a concrete case. A creative writing system handed a "magical realism" prompt produces a dragon story. Under the Ji et al. extrinsic definition, the dragon is a fabricated entity and the output is severely hallucinated. Under a faithfulness-to-intent definition, the same output is exactly what the user asked for. Under a self-inconsistency definition, the verdict depends on whether the dragon's color shifts between paragraphs. One output, three verdicts.

The measurement evidence is consistent with our reading. Maynez et al. (2020) report inter-annotator agreement at $\kappa = 0.67$–0.73 on hallucination presence across systems. Among sentences unanimously judged non-factual, agreement falls to $\kappa = 0.39$ on which kind of non-factuality (Pagnoni et al., 2021). The disagreement is not whether the sentence is wrong; it is which kind of wrong. Detection methods on the same benchmark span a wide range of AUC-PR scores (Manakul et al., 2023). We read this as disagreement about the concept itself, not noise in the measurement.

The literature's failure to converge is evidence that "hallucination" is not a natural category. It is a catch-all for output characteristics misaligned with task requirements, and the requirements vary by application. A unified theory of hallucination is, we suspect, out of reach: it would require unifying different quality criteria under one label. Recent work using zero-shot knowledge probes to elicit and inspect hallucination patterns (Lee et al., 2024; Farquhar et al., 2024; Kuhn et al., 2023) sharpens the diagnostic tools available, but the underlying definitional ambiguity remains upstream of any detector.

2.2 Context Trap

Hallucination detectors often operate under a context-independent assumption: that the same criterion applies across tasks, modalities, cultures, and languages. Through deployment evidence, we show that the assumption breaks for creative writing, multi-modal generation, and translation. We propose context-conditioned evaluation, where the same model behavior is scored differently depending on the task it was asked to perform.

Take this output: *"Dr. Elena Vasquez, a neural interface researcher at Stanford's NeuroFutures Lab, demonstrated her brain-computer translation system that converts thoughts directly into synthesized speech."* For medical QA, this is dangerous fabrication. For creative writing, it is exactly the desired output. Current detectors flag both identically.

The New Yorker Caption Contest (Hessel et al., 2023) forces the issue. Consider a cartoon of a person standing in a field of giant pencils with the caption *"The writers' strike is really taking root."* Under text-only evaluation, the caption is severely hallucinated, since writers do not plant pencils. Multi-modal evaluation reads it as a successful visual-linguistic joke. The image and the text disagree in the right way. We argue this dataset is full of cases where humor depends on the modalities failing to ground each other.

The JOKER shared tasks (Ermakova et al., 2022, 2023) make a similar point in translation. The English pun *"Time flies like an arrow; fruit flies like a banana"* cannot be rendered word-for-word into French; the syntactic ambiguity that drives it does not survive the trip. A faithful translation has to invent new structure. We argue that what a hallucination detector flags as an unfaithful translation is, in this case, the only translation that preserves the joke.

Code-mixed settings complicate the picture further. Humour datasets for English-Hindi code-mixed tweets (Khandelwal et al., 2018) show that strict grounding to a single language misses the communicative intent. We argue code-mixed humour requires novelty in language-switching plus cultural grounding in both languages at once.

Low-resource languages face a worse double bind. Models hallucinate more frequently in low-resource translation directions (Guerreiro et al., 2023). ChatGPT achieves only 41% sentence-level accuracy on lemma disambiguation for Erzya, an endangered Uralic language, even with dictionary augmentation (Hämäläinen, 2024). We argue some "hallucination" in endangered-language contexts, like neologisms or grammatical extensions, looks more like language stewardship than error (Zhang et al., 2022).

Current systems treat context as irrelevant. Our critique is targeted: we are arguing against source-grounded factuality detectors built for summarization and QA, now applied as general-purpose quality filters across creative, multi-modal, and multilingual tasks. Task-specific faithfulness metrics that already condition on task requirements fall outside our scope.

2.3 Confusing Failure with Feature

Hallucination detection often operates under a third assumption: that creative model behaviors, including uncertainty, novelty, and semantic deviation, are failures to minimize. Through the computational humor literature, we show that the same surface signals are creative mechanisms in appropriate contexts. We argue these behaviors should be treated as task-conditional: features for some applications, defects for others.

Post-2020 computational humour research gives us the cleanest counter-example. Work on generating and explaining humour remains sparse (Loakman et al., 2025), although recent papers argue for turning hallucination into creativity by drawing on the divergent and convergent phases described in the cognitive-creativity literature (Jiang et al., 2024). The surface signatures hallucination detectors flag as errors are, in these accounts, the creative machinery itself:

Information-theoretic literature makes the connection precise. Shannon entropy of letter combinations predicts perceived funniness of non-words (Westbury et al., 2016). Puns work by holding two near-equally-likely meanings in tension, with ambiguity and distinctiveness as the operationalisation (Kao et al., 2016). Translate this to LLM generation. A confident next-token prediction usually kills the joke. We argue the surprise, the low-probability swerve, is the humour itself; what a traditional reading calls model uncertainty is the resource a creative task needs.

Humor theory gives us frameworks for when norm violations succeed as humor. Attardo separates bona fide from non-bona-fide communication (Attardo, 2020). Benign violation theory shows that violations perceived as simultaneously threatening and benign get judged as humorous (Warren and McGraw, 2016). The same surface categories that identify errors in factual contexts identify successful creative mechanisms in humor:

- *"The rock was getting tired"* Error: hallucinated attributes; Humor: personification
- *"She downloaded the sunset"* Error: impossible action; Humor: domain blending
- *"Gravity works sideways"* Error: physics violation; Fiction: worldbuilding premise

The Unfun task (Horvitz et al., 2024) gives us perhaps the cleanest experimental evidence. The task is to remove humor from jokes, leaving minimal contrastive pairs. Models, it turns out, are good at this. What is striking is the mechanism: they eliminate humor by reducing hallucination signatures. Original: *"A SQL query walks into a bar, walks up to two tables and asks, 'Can I join you?"'* Unfunned: *"A database query searches for information from two data sources and requests to combine them."*

The Unfunning process eliminates anthropomorphization (hallucinated agency), unexpected semantic connections (hallucinated social scenario), technical metaphor (semantic deviation), and ambiguity at "join" (dual meaning). The unfunned version has lower hallucination scores. It also has zero humor.

Creative humor generation calls for leap-of-thought reasoning across semantically distant concepts (Zhong et al., 2024) and for multi-step association pipelines (Tikhonov and Shtykovskiy, 2024). Standard autoregressive LLMs may be structurally hostile to genuine surprise (Franceschelli and Musolesi, 2025). We expect the creative behaviors our framework wants to preserve will require generation strategies beyond next-token prediction.

Other creative domains tell us much the same story. AI-augmented brainstorming improves group ideation precisely by injecting unexpected concept combinations (Shaer et al., 2024), although

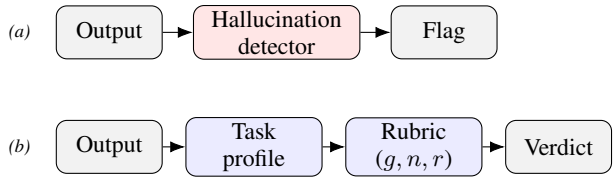


Figure 1: Schematic contrast between (a) current universal hallucination detection and (b) the proposed task-output alignment pipeline, which first identifies a task profile and then applies a grounding-novelty-risk rubric.

Task	Ground.	Novel.	Risk Tol.
Medical QA	High	Low	Low
Humor Gen.	Low	High	High
News Summary	High	Low	Low
Creative Fiction	Low	High	High
Code Gen.	Med	Med	Med
Brainstorming	Low	High	High

Table 1: Illustrative task positioning in alignment space. Ratings are qualitative ordinal bands, not empirical measurements; placements assume a typical instance of each task.

unconstrained generation does risk homogenising outputs across users (Anderson et al., 2024). Studies of LLM creative writing find that evaluation has to split artistic merit from factual accuracy (Chakrabarty et al., 2024). The lesson is consistent: brainstorming, fiction, and humour reward the very semantic leaps that factual detectors penalise. We argue creative applications are penalised, in current setups, for exactly the behaviours they were asked to produce.

3 Reframing: Task-Output Alignment Assessment

We propose replacing hallucination detection with **task-output alignment assessment**: outputs are judged against the requirements of the task, modality, cultural context, and user intent, not against a universal criterion. Figure 1 contrasts the two pipelines at a glance.

What is new beyond saying evaluation should be multidimensional? Two things. First, we argue "hallucination" itself is load-bearing in the field's discourse and detector ecosystem, and we propose retiring the term, not refining it. Second, we operationalize the alternative through three specific axes with anchored ordinal rating criteria (§3), tied to existing rubric-based evaluation work (Hashemi et al., 2024). Most multidimensional evaluation work assumes the field has decided what it is measuring. On hallucination, it has not.

3.1 The Three-Dimensional Framework

We evaluate model outputs along three axes. The first axis is factual grounding: how tightly should outputs be constrained by verifiable reality? Medical diagnosis, legal advice, and news sit at the high end (90–100%). Historical fiction and science communication sit in a middle band (40–70%). Humor, speculative fiction, and brainstorming sit at the low end (0–30%).

The second axis is novelty: how much should outputs introduce new ideas, entities, or connections? Creative writing, humor, and brainstorming demand high novelty (70–100%). Advertising and educational analogies sit in the middle (30–60%). Information retrieval, translation, and summarization sit at the low end (0–20%). The third is risk tolerance: what are the consequences of misalignment?

- Low tolerance: Safety-critical (health harm, financial loss)
- Medium tolerance: Productivity tools (frustration, wasted time)
- High tolerance: Entertainment (user can discard/regenerate)

3.2 Mapping Tasks in Alignment Space

Table 1 positions common LLM applications in this space. Humor and creative writing sit in the opposite corner from medical QA. They share low grounding, high novelty, and high risk tolerance. Translation and educational analogies share high grounding but split on novelty. We argue risk tolerance moderates evaluation strictness independent of grounding and novelty.

Within-task heterogeneity is large. Consider code generation. Autopilot software needs the safety profile of medical QA, while a Discord bot tolerates much more risk. Translation of literary fiction asks for higher novelty than translation of technical manuals. The placements in our table are illustrative; we argue the variation within a task is often as large as the variation between tasks.

An LLM generates: *"The neural pathway lit up like a Fourth of July fireworks show."* Under medical QA alignment, this is misaligned. The metaphor introduces unverifiable imagery where precision is required. Under science-communication alignment, the same output is well-aligned. The metaphor aids comprehension while conveying the core phenomenon accurately. We argue the framework

makes these divergent judgments explicit and principled, where a single hallucination score collapses them.

For each task-output pair, annotators assign ordinal ratings on each axis using anchored bands: *high* (output must match external truth and remain self-consistent), *medium* (output must be plausible and self-consistent but need not match external truth exactly), and *low* (output is judged on task fitness rather than on external truth). They then check whether the output falls within acceptable bounds for that task profile. We build this rubric-based approach on top of recent work in multidimensional, calibrated text evaluation (Hashemi et al., 2024).

The three axes are not orthogonal in practice. Pushing grounding higher narrows the room for novelty, since outputs that must match external reality cannot freely introduce new entities. Raising risk tolerance can buy back some novelty in return. Annotators using the rubric should commit to a position on each axis and justify it, rather than aggregate into a single hallucination score that obscures the trade-offs.

3.3 Operationalizing: The Creativity Dial

For systems that operate across multiple tasks, we propose a creativity dial: explicit control mechanisms that calibrate output characteristics to task requirements:

- Medical QA: constrained decoding, high confidence thresholds, source attribution
- Humor: controlled entropy targets, semantic-distance optimization
- Translation: Minimum Bayes Risk decoding tuned to task-appropriate quality metrics (Kumar and Byrne, 2004)

Evaluation has to be layered. Linguistic coherence is always required. World-knowledge consistency is task-dependent; it is required for medical applications and not for humor. Novelty should be penalized in retrieval tasks and rewarded in creative ones.

We argue creativity is not unconstrained hallucination. It is controlled semantic deviation, calibrated to what the task asks for

3.4 Cross-Cultural and Multimodal Extensions

What counts as appropriate grounding varies by language and culture (Liu et al., 2025; Hershcovich et al., 2022). Through several case studies, including Arabic *saj'* (Elzohbi and Zhao, 2025), Chinese *chengyu* (Fu et al., 2025; Zheng et al., 2019), and code-mixed humor (Khandelwal et al., 2018), we show the variation is large. We propose calibrating grounding and novelty requirements locally to each language and modality, not globally.

Multi-modal humor research shows that modalities contribute unequally; combining acoustic, visual, and textual signals improves humor detection in TED talks (Hasan et al., 2019). We extend this point. Memes likely need low text grounding but high visual-text incongruity. Image captioning needs high visual grounding and low novelty. Creative visual storytelling sits at medium visual grounding with high narrative novelty.

4 Implications

4.1 For Multi-Task LLMs

Current foundation models handle diverse tasks with a single quality criterion. Through the proposed alignment framework, we show that one criterion is the wrong abstraction. We propose task classifiers that activate appropriate evaluation criteria, confidence calibration layers conditioned on task type.

Consider the prompt "Write me a joke about databases." A working system classifies this as a creative task with low grounding, high novelty, and high risk tolerance. It activates humor-specific evaluation, generates with creative decoding, and judges the output on coherence and surprise rather than factual accuracy.

Most current systems lack this kind of task-conditional switching. They apply similar evaluation regardless of whether the user is asking for a joke or for medical advice. We argue that is the bug.

4.2 For Responsible AI Deployment

Regulators have started to converge on a context-specific approach to AI risk. Both the EU AI Act (European Parliament and Council of the European Union, 2024) and NIST's AI Risk Management Framework (National Institute of Standards and Technology, 2023) explicitly call for risk characterization that varies by deployment context. We argue this maps onto our alignment framework. Strict grounding, source citation, and human oversight make sense in medical and legal applications. Creative applications can run looser. The 2024 Canadian decision in *Moffatt v. Air Canada*, where the airline was held liable for its chatbot's invented

bereavement-fare policy, illustrates the deployment side of the same point.[2] Loosening hallucination constraints in appropriate places does not compromise safety in critical ones (Weidinger et al., 2022); conflating the two does.

4.3 For Computational Creativity Research

Computational creativity research has the same measurement problem we are pointing at. Through humor, brainstorming, and creative writing studies, we show that creative quality and factual accuracy live on different axes. We propose evaluating creative outputs against task-specific creative requirements, including semantic distance, surprise, and novelty, not against world-knowledge consistency.

Uncertainty becomes a resource rather than a problem; high-entropy regions are where creative work happens. Humor theory offers ready-made frameworks for distinguishing intentional from accidental norm violations. Cultural and linguistic variation shapes what counts as creative versus nonsensical, which means evaluation has to be local.

4.4 Research Agenda

A handful of problems linger. The three axes aren't necessarily the sole dimensions; whether further ones are required, and which, remains an empirical matter we haven't resolved. Models must also discern, from context alone, what species of output the user expects. The contrast between "tell me a joke" and "give me medical advice" is straightforward, yet a great many real-world prompts lie somewhere in between. Calibration work (Kuhn et al., 2023; Farquhar et al., 2024) to date has clustered around factual tasks.

5 Conclusion

We argue the NLP community has been chasing the wrong target. Universal hallucination detection assumes any semantic departure from factual grounding is uniformly bad, regardless of task, modality, or culture. Through the computational humour literature, we show the assumption falls apart. Our proposal is task-output alignment assessment in lieu of universal detection, with the same model behaviour judged differently in different tasks.

Computational humor gives us the cleanest case study. The signatures hallucination detectors flag as errors, high entropy and ambiguity (Westbury et al., 2016), semantic norm violations (Attardo, 2020), novel entity introduction, and unexpected connections (Zhong et al., 2024), are the same mechanisms that generate successful jokes, fiction, and brainstorming output. We have early evidence from AI-augmented brainstorming (Shaer et al., 2024) and creative writing evaluation (Chakrabarty et al., 2024) that the framework generalises beyond humour, though we are not yet claiming it has been proven to.

We propose task-output alignment assessment in place of universal hallucination detection. Tasks are positioned in an alignment space defined by grounding, novelty, and risk tolerance. The same model behavior receives different evaluations in different cells of that space. Medical chatbots and creative writing assistants should not share evaluation metrics.

6 Limitations

Our argument rests rather heavily on computational humour as the main evidence base. That domain has its peculiarities. We haven't tested whether the same case holds for scientific summarisation or legal drafting. The three-axis framework is a first cut. A working version will probably need more axes, and we haven't done the empirical work to say which.

Ethical Considerations

Reframing LLM quality assessment as task-output alignment carries dual-use implications. On the positive side, alignment-centric evaluation surfaces failure modes that hallucination-focused metrics miss, including outputs that are factually correct but pragmatically misaligned with user intent in high-stakes domains like medical triage, legal drafting, and education. On the negative side, any reframing risks being adopted as marketing language without corresponding rigor; vendors could claim "aligned" outputs without disclosing the underlying evaluation methodology.

[2] Civil Resolution Tribunal, BC, 14 Feb 2024; CAD$812.02 in damages.

References

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*, pages 413–425, Chicago, IL, USA. ACM.

Salvatore Attardo. 2020. *The Linguistics of Humor: An Introduction*. Oxford University Press.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA. ACM.

Mohamad Elzohbi and Richard Zhao. 2025. Tahḏīb: A rhythm-aware phrase insertion for classical Arabic poetry composition. In *Proceedings of the Third Arabic Natural Language Processing Conference*, pages 194–202, Suzhou, China. Association for Computational Linguistics.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of JOKER – CLEF-2023 track on automatic wordplay analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2023)*, volume 14163 of *Lecture Notes in Computer Science*, pages 397–415. Springer.

Liana Ermakova, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Claudine Borg, Élise Mathurin, Gaëlle Le Corre, Sílvia Araújo, Radia Hannachi, Julien Boccou, Albin Digue, Aurianne Damoy, and Benoît Jeanjean. 2022. Overview of JOKER@CLEF 2022: Automatic wordplay and humour translation workshop. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, pages 447–469. Springer.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act). Official Journal of the European Union, OJ L, 2024/1689, 12.7.2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Giorgio Franceschelli and Mirco Musolesi. 2025. On the creativity of large language models. *AI & Society*, 40(5):3785–3795.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Mika Hämäläinen. 2024. DAG: Dictionary-augmented generation for disambiguation of sentences in endangered Uralic languages using ChatGPT. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 36–40, Helsinki, Finland. Association for Computational Linguistics.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from The New Yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):42:1–42:55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cognitive Science*, 40(5):1270–1285.

Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor detection in English-Hindi code-mixed social media content: Corpus and baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1203–1207, Miyazaki, Japan. European Language Resources Association.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 169–176.

Seongmin Lee, Hsiang Hsu, and Chun-Fu Chen. 2024. LLM hallucination reasoning with zero-shot knowledge test. In *Proceedings of the NeurIPS 2024 Workshop on Socially Responsible Language Modelling Research (SoLaR)*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? An overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference (INLG 2025)*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.

National Institute of Standards and Technology. 2023. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829. Association for Computational Linguistics.

Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L. Kun, and Hagit Ben Shoshan. 2024. AI-augmented brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA. ACM.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC 2024)*.

Caleb Warren and A. Peter McGraw. 2016. Differentiating what is humorous from what is not. *Journal of Personality and Social Psychology*, 110(3):407–430.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 214–229, Seoul, Republic of Korea. ACM.

Chris Westbury, Cyrus Shaoul, Gail Moroschan, and Michael Ramscar. 2016. Telling the world's least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language*, 86:141–156.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13246–13257.

Challenging the Myth: A Research Arc on LLMs as Human Simulacra

Simon Münker
Trier University
muenker@uni-trier.de

Achim Rettinger
Trier University
rettinger@uni-trier.de

Damian Trilling
Vrije Universiteit Amsterdam
d.c.trilling@vu.nl

Abstract

When Large Language Models (LLMs) combined with prompt-based approaches as human simulacra emerged, they promised revolutionary shortcuts. Models trained on vast internet corpora may replicate human behavior and communication through text-based alignment. The initial optimism of the NLP community positioned LLMs as universal human proxies capable of replacing participants in surveys, generating authentic social media content, and simulating diverse cultural perspectives. We systematically dismantle this *"myth of universal generalization"* and document a shift toward methodological rigor. Our research reveals fundamental limitations: LLMs exhibit inhuman response patterns in psychometric assessments and produce detectable synthetic content. We analyze the difference between superficial linguistic fluency and genuine human-like representation, and reframe the current paradigm from asking *"can LLMs replace humans?"* to *"under what validated conditions might LLMs serve as useful research components in social sciences?"* Our work shows how interconnected research efforts challenge foundational assumptions and establishes best practices for deploying LLMs as human simulacra.

1 Introduction: A Myth and Its Spiral

Science, as Popper (2014) observed, must begin with myths and with the criticism of myths. The myth examined in this work is seductive: Large Language Models (LLMs), trained on the aggregate output of human civilization (Brown et al., 2020), can serve as reliable proxies for human participants in social science research. Model developers and early adopters promoted this claim (Argyle et al., 2023; Park et al., 2023; Teubner et al., 2023), and a steadily growing group of researchers deploy LLMs as annotators (Pavlovic and Poesio, 2024), survey respondents (Adilazuarda et al., 2024; Mohammadi et al., 2025), and social media agents (Törnberg et al., 2023; Chuang et al., 2024; Larooij and Törnberg, 2025a), often without validating whether model outputs genuinely resemble human behavior beyond surface plausibility (Larooij and Törnberg, 2025a,b; Wang et al., 2025).

LLMs offer unprecedented scalability (Bisbee et al., 2024; de Wynter, 2025; Yu et al., 2025), responses from thousands of *"participants"* in hours rather than months (Bisbee et al., 2024). They eliminate ethical complications of human subject research and promise perfect experimental control (Grossmann et al., 2023). It is less well-understood, though, whether these approaches are *valid*. A critical methodological gap (Tjuatja et al., 2024) separates what LLMs have been demonstrated to do from what researchers assume they can do and that gap is most consequential when the assumption concerns human-likeness (Salles et al., 2020). Benchmark performance and linguistic fluency are not surrogates for structural alignment with human behavior (Agnew et al., 2024; Wang et al., 2025; Yu et al., 2025). Model developers routinely report impressive performance on standardized benchmarks (Wang et al., 2018, 2019) and sometimes claim *"superhuman performance"* on specific tasks (Bubeck et al., 2023), claims that concern task-solving ability, not human-likeness. By definition, superhuman performance is no longer human-like. While computational social science is concerned with learning about human social behavior, all one might be able to learn when deploying LLMs instead of human study participants is about how LLMs *"behave"* (Shao et al., 2023). Thus, we require not only empirical counter evidence but new metrics capable of making invisible failures visible: a diagnostic vocabulary adequate to the depth of the problem.

1.1 Central Hypothesis and Research Questions

Our work constitutes a systematic falsification. We test the following specific null hypothesis and op-

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 31–44

erationalize the core components as follows:

> *Off-the-shelf LLMs with minimal prompt engineering show quantitatively indistinguishable, human-like performance in digital behavioral tasks.*

1. **Off-the-shelf LLMs** denotes publicly available, instruction-tuned open-source models used without task-specific fine-tuning (Touvron et al., 2023; Dubey et al., 2024; Yang et al., 2024; Jiang et al., 2023), the most common deployment mode in social-science applications (Alizadeh et al., 2025; Møller and Aiello, 2024).

2. **Minimal prompt engineering** denotes researcher-specified language prompts, persona descriptions, task instructions, without iterative optimization against outcome-specific test sets (Liu et al., 2023; White et al., 2023).

3. **Quantitatively Indistinguishable Human-like Performance** requires that LLM outputs align with empirical human baselines at the level of distributions, internal response structures, and statistical effect sizes, not only means and variances (Tjuatja et al., 2024; Shu et al., 2024).

4. **Digital Behavioral Tasks** encompass two complementary domains central to social science applications: psychometric questionnaire responses (Demszky et al., 2023; Ye et al., 2025) and social media content generation (Larooij and Törnberg, 2025a; Ng and Carley, 2025).

We falsify this hypothesis through systematic counterexamples rather than a single experiment and formulate the following three research questions that operationalize our hypothesis.

RQ_1 Do LLMs represent the internal structure of psychological constructs observed in textual questionnaires in ways that align with the response patterns of human populations?

RQ_2 Can LLMs generate realistic social media content and replicate authentic patterns of human interaction based on history-based modeling?

RQ_3 Can LLMs serve as human simulacra through prompt-based approaches, or does effective alignment and ecological validity require data-driven adaptation?

1.2 Structure of the Research Arc

We address these questions through progressively more sophisticated methods: from zero-shot text classification (Münker et al., 2025) to psychometric fingerprinting (Münker, 2025b), from informal content ratings to multi-dimensional linguistic authenticity metrics (Münker et al., 2026). Our research arc is not a single experiment but a spiral (Jones, 1994), each study revealing a failure, each failure motivating a more precise diagnostic, each diagnostic generating an insight that the prior vocabulary could not express.

The arc originates from two preliminary failures (Section 2) that raised questions that required methodological answers. First, how does one quantify the misalignment between synthetic and human content? And second, what evaluation framework distinguishes genuine alignment from surface plausibility? The psychometric strand (Section 3) pursues the first question through three progressively finer lenses, mean comparisons, variance analysis, and inter-item correlation fingerprinting, each revealing limitations invisible to its predecessor. The social-agent strand (Section 4) pursues the second question by formalizing empirical realism and introducing multi-dimensional linguistic authenticity metrics. The two strands converge on RQ_3 (Sections 5–6): the evidence across both domains independently demonstrates that prompting fails for structurally different reasons, and that fine-tuning, while necessary, is not sufficient. We show that the absence of a validation culture (Taubenfeld et al., 2024; Qi et al., 2025), the tendency to deploy LLMs as human proxies under the assumption of capability rather than the demonstration of it, is not a peripheral, but rather a central methodological flaw.

2 The Research Context and Its Catalyst

Our work emerged from the TWON project (Twin of an Online Social Network) project, which aimed to build realistic simulations of social media platforms to study democratic discourse (Gao et al., 2024; Rossetti et al., 2024; Münker and Rettinger, 2025). The motivating question was practical: can LLM agents, prompted to behave like real users, populate a digital twin with ecologically valid behavior? This question is within a broader literature that has moved rapidly to deploy LLMs as social science instruments (Thapa et al., 2025; Grossmann et al., 2023), as annotators for complex datasets, as

automated survey respondents (Argyle et al., 2023; Bisbee et al., 2024), and as generative agents in social simulations (Park et al., 2023; Törnberg et al., 2023), often without validating whether the outputs genuinely resemble human behavior beyond surface plausibility (Larooij and Törnberg, 2025b; Agnew et al., 2024; Wang et al., 2025). Two preliminary experiments negatively answered the TWON motivating question, and, crucially, generated the research questions that structured the section that follows.

Failure 1: LLMs as Annotators. We applied zero-shot prompt-based classification to German political tweets (Münker et al., 2025). The results were sobering. Models fabricated categories outside of provided taxonomies, produced different classifications for identical inputs across repetitions, and did not show a consistent relationship between prompt sophistication and performance (Jang et al., 2023). Detailed annotation guidelines sometimes improved large models while confusing smaller ones; task-name prompts occasionally outperformed elaborate handbooks. These failures were diagnostically important: they demonstrated not only poor performance but also unstable and opaque behavior (Ollion et al., 2024; Münker and Sartori, 2026), properties that disqualify a system as a scientific instrument, regardless of average precision. Even when LLMs perform well on average, performance varies substantially across models and prompts with no reliable way to anticipate which combination will succeed, and aggregate metrics can obscure poor coverage of minority classes (Stolwijk et al., 2025).

Failure 2: LLMs as Social Media Users. Parallel experiments compared GPT-3.5-turbo (Achiam et al., 2023) and Mistral-7B (Jiang et al., 2023) generating political social-media posts across English, German, and Dutch for conservative, liberal, and alt-right personas (Hershcovich et al., 2022). Two patterns stood out. First, a dramatic language asymmetry: English content rated much higher in perceived authenticity by native speakers than Dutch content, despite claims of multilingual capabilities (Hershcovich et al., 2022). A native Dutch reviewer found the generated content US-centric, discussing US political figures in a European context. Second, a systematic ideological bias (Münker, 2025c): liberal personas achieved the highest authenticity ratings, while conservative-prompted models often expressed moderate or progressive viewpoints (Rozado, 2023; Rutinowski et al., 2024). An additional pattern emerged around content idealization: the generated posts featured complete sentences, logical transitions, and grammatical correctness that far exceeded the typical platform norms of abbreviations, typos, and emoji (Duncan, 2024), an excessive polish that immediately marks content as synthetic.

These failures were productive because they were *specific*. They raised questions that demanded methodological answers: how does one quantify the misalignment between synthetic and human content? What evaluation framework distinguishes genuine alignment from surface plausibility (Larooij and Törnberg, 2025b)? The rest of our work is, in essence, an attempt to build those frameworks, and the answer to RQ_3 begins here: if prompting cannot even sustain ideological consistency or match a language's informal register, it cannot serve as the foundation for valid human simulation.

3 Psychometrics: From Means to Structure

Our first research question (RQ_1) asked whether LLMs represent the internal structure of psychological constructs in ways that align with human populations. We investigate this through three progressively finer lenses, each exposing limitations the prior level could not detect.

3.1 Mean Comparisons: Necessary but Insufficient

The first study (Münker, 2025c) used the Moral Foundations Questionnaire (MFQ) (Graham et al., 2009, 2011), repeatedly surveying seven open-source models (7B-176B parameters) prompted to respond as conservative, moderate or liberal individuals. The finding was clear: models failed to reproduce the ideological patterns observed in human populations (Hatemi et al., 2019; Hutchinson et al., 2020; Abid et al., 2021; Liu et al., 2022). Conservative-prompted models did not align with conservative human baselines; the variance between repetitions was enormous (0.030–0.425, depending on the model), far exceeding human intra-individual variability (Tjuatja et al., 2024).

The lesson: Mean comparisons reveal systematic bias but hide inconsistency. A model could average to the correct mean while producing wildly scattered individual responses, a problem invisible

to the standard *"does the LLM agree with humans on average?"* design (Bisbee et al., 2024; Petrov et al., 2024; Lee et al., 2025).

3.2 Variance Analysis: Necessary but Still Insufficient

Extending to the MFQ-2 across 19 cultural contexts (Münker, 2025a), we found a deeper problem: LLMs systematically homogenize moral diversity (Anderson et al., 2024; Priyanshu and Vijay, 2024). Average alignment was better for European contexts (Belgium: mean distance 1.321) than non-Western ones (Japan: 2.970), reflecting Western training data biases (Ryan et al., 2024; Myung et al., 2024). However, to extend our analysis beyond mean-only comparison, we utilized an ANOVA which revealed that Mistral 7B produced statistically indistinguishable responses across cultural personas for 34 of 36 questionnaire items, effectively generating the same output regardless of specified cultural background.

A surprising finding: model size did not reliably improve performance. Qwen 2.5 7B outperformed its 72B counterpart (mean distances 0.817 vs. 1.143), while Mistral showed the opposite pattern. This inconsistency would recur throughout the following studies and eventually motivated a structural argument: the limitation is not capacity, but the training objective (de Wynter, 2025). Models learn to produce fluent text, not psychologically valid responses (Bender et al., 2021).

The lesson: Variance analysis catches homogenization, but still evaluates output item-by-item. It cannot detect whether the relationships between items, the factor structure that defines a psychological construct, are preserved (Nunnally, 1975).

3.3 Fingerprinting: The Missing Dimension

The third study (Münker, 2025b) introduced a novel methodology treating the inter-item correlation matrix of questionnaire responses as a *"fingerprint"* of how a model internally organizes psychological constructs (Pearson, 1901; Cronbach, 1951). Using the Humor Styles Questionnaire (Martin et al., 2003) and 1,000 independent response sets from six LLMs, we constructed these fingerprints and compared them to human baselines.

Human response groups showed high *"fingerprint"* similarity (0.776–0.891; mean 0.823), reflecting the robust psychological constructs underlying humor preferences. The LLM fingerprints showed near-zero similarity to with human patterns (mean 0.026), orthogonal relationships that indicate fundamentally different organizational principles. Exploratory Graph Analysis (Golino and Epskamp, 2017) confirmed that no tested model recovered the theoretically expected four-factor structure of the HSQ; instead producing 2-8 idiosyncratic communities. Cronbach's α (Cronbach, 1951) ranged from 0.008 to 0.617 across models and dimensions, compared to 0.790–0.841 for humans.

The surprising insight: cross-family model similarities often exceeded within-family similarities, suggesting that factors beyond architectural lineage (presumably training data and instruction-tuning procedures (Sparrenberg et al., 2024)) dominate the organization of psychological constructs. This architectural independence challenges the implicit assumption that model families share psychological representations (Sandhan et al., 2025).

The lesson: The failure is not noise around approximately correct means; it is structural. LLMs organize psychological constructs according to qualitatively different principles from human cognition (Ren et al., 2025). This has direct implications for RQ_3: if the deficit is structural rather than superficial, prompt engineering, which operates at the surface level, is inherently insufficient.

4 Social Agents: From Static to Dynamic

The second research question (RQ_2) shifted from controlled psychometric settings to open-ended social media content generation. Two studies (Münker et al., 2026; Münker et al., 2026) closed the empirical arc by extending the misalignment argument from Likert-scale responses to free-text communication.

4.1 Formalizing Empirical Realism

A foundational contribution (Münker et al., 2026) was methodological: formalizing what it means for an LLM agent to behave realistically in a social network context. Prior simulation work typically assumed validity based on surface plausibility; we introduced mathematical definitions for user-level behavior and platform mechanics, along with quantifiable loss functions to measure *empirical realism*, the distance between simulated and observed behavior. This formalization made validity claims falsifiable rather than anecdotal and untestable.

Instantiating the framework on a dataset of German and English political discourse from X, we compared prompt-based and fine-tuned approaches

on three tasks: generating original posts, generating replies, and predicting reply likelihood. The language asymmetry observed in Section 2 was replicated and quantified. Fine-tuned English reply generation achieved substantial BLEU scores (0.239) and strong embedding similarity (distance 1.427), with TweetEval correlations of 0.377-0.586. Fine-tuned German models failed dramatically (BLEU: 0.021; embedding distance: 2.891), with high variance indicating unstable performance. The reply likelihood task showed an English fine-tuned F1 of 0.978 vs. German 0.703.

The lesson: Benchmark performance in English does not transfer even to major European languages with substantial training data. Universal generalization claims cannot be taken at face value.

4.2 Multi-Dimensional Authenticity Detection

The final study (Münker et al., 2026) introduced a multi-dimensional evaluation framework that combines quantitative linguistic features, morphosyntactic analysis, semantic classification, and embedding-based clustering to assess where synthetic content diverges from human communication. Fine-tuned models consistently outperformed prompt-based approaches in all feature types, confirming the superiority of data-driven adaptation. However, both remained detectably synthetic. The classifier with the highest performance that combined tf-idf, fastText embeddings, and extracted features achieved macro F1 of 0.7301 (German) and 0.6972 (English).

The lesson: Traditional tf-idf representations proved remarkably effective for detecting prompt-based content (German: 0.8510; English: 0.8000), outperforming modern neural embeddings. This suggests that naively generated content exhibits surface-level lexical regularities so systematic that bag-of-words features suffice for detection. The excessive polish observed in preliminary experiments (complete sentences, grammatical correctness, formal transitions) (Münker et al., 2026) leaves a lexical fingerprint as distinctive as the psychometric fingerprint (Münker, 2025b) at the correlation level.

5 Insights: What the Spiral Reveals

Telling this story as a connected arc, rather than as eight independent papers, surfaces insights difficult to glean from any individual contribution.

The Prompting Insufficiency, Formally. Each phase of the research provides independent evidence that prompting LLMs cannot overcome training-induced limitations. In psychometrics: prompting fails to produce stable, culturally distinct, or structurally valid responses. In content generation: prompting produces easily detectable lexical signatures and ideological homogenization. The convergence across domains and methods, from Cronbach's α (Cronbach, 1951) to BLEU scores (Papineni et al., 2002) to tf-idf classifiers (Ramos et al., 2003), constitutes stronger evidence than any single experiment. Crucially, each study reveals a *different mechanism*: instability (high inter-repetition variance), homogenization (ANOVA indistinguishability across cultural personas), structural misalignment (fingerprinting orthogonality), and lexical regularities (tf-idf superiority over neural embeddings). Together, they suggest that prompting fails for multiple, compounding reasons that are unlikely to be resolved by further prompt engineering alone (Liu et al., 2023; Møller and Aiello, 2024).

Scale as a Red Herring. The most consistent cross-study finding is that model size does not reliably improve performance on socially-grounded tasks. Qwen 2.5 7B outperformed its 72B counterpart in cultural diversity representation; no tested model, regardless of size, recovered the expected structure of the HSQ factor; size-performance correlations were inconsistent between tasks and languages. The implication goes deeper than a negative result: current scaling approaches (Brown et al., 2020) optimize for benchmark performance and linguistic fluency, not for psychological validity or cultural fidelity (Adilazuarda et al., 2024). The failure mode is not an insufficient capacity but a misspecified training objective (de Wynter, 2025).

Fine-Tuning: Necessary but Not Sufficient. Data-driven adaptation through supervised fine-tuning consistently outperforms prompting and should be treated as the minimum viable approach for deployment (Alizadeh et al., 2025; Møller and Aiello, 2024). But fine-tuned models remain detectably synthetic through multi-dimensional analysis. Fine-tuning reduces the most visible symptoms of misalignment without addressing the underlying cause (Lin, 2024). This distinction matters for how researchers frame validity claims: competitive task metrics and genuine behavioral fidelity are not the same thing, and treating them as equivalent is precisely the conflation that produced the myth this

work dismantles (Larooij and Törnberg, 2025b).

Not all Languages are Equal. The English-German-Dutch performance hierarchy, observed in preliminary experiments and quantified in multiple studies, reveals that the claims of LLM capability are implicitly English-centric (Hershcovich et al., 2022; Ryan et al., 2024). Researchers who deploy LLMs validated on English data in other languages conduct invalid experiments without knowing it (Heseltine, 2025). This is not a peripheral concern for multilingual NLP; it is a fundamental threat to the validity of any computational social science research using LLMs in non-English contexts, that is to say, most of the world.

The Absence of a Validation Culture. Across the literature that this arc responds to, the dominant pattern is deployment without calibration. LLMs are used as human proxies under the assumption of capability rather than the demonstration of it (Taubenfeld et al., 2024; Qi et al., 2025; Balluff et al., 2026). Our most practically consequential contribution is not any single metric or finding, but the argument that this assumption is unjustified and that the field requires a norm of mandatory domain-specific validation before each new deployment (Larooij and Törnberg, 2025b). As with any scientific instrument (Popper, 2014), the question is not whether the tool is impressive, but whether it has been calibrated for the task at hand (Grimmer and Stewart, 2013).

6 Lessons Learned & Future Work

Start with the metric. The methodological arc moved from surface comparisons (means) toward structural ones (fingerprinting, multi-dimensional detection). In retrospect, beginning with the detailed metrics would have been more efficient, but the superficial metrics were necessary to establish that means and variances were insufficient (Norris and Lecavalier, 2010). The lesson for future researchers: define what *"valid alignment"* means before collecting data. Without a pre-specified structural criterion, apparent success may reflect statistical coincidence or implicit prompt optimization against an undeclared test set (Miller et al., 2021).

Report negative results explicitly. Several findings here are null results or failures that carry scientific value: the absence of size-performance correlation; the failure of cultural persona prompting; the inability of any tested model to recover factor structures. These are as informative as positive findings (de Wynter, 2025), but face publication pressures that reward only the latter. The research arc format is precisely the venue where such results can be foregrounded rather than buried in appendices.

The evaluation protocol problem compounds over time. Iterative prompt refinement without independent evaluation is methodologically equivalent to tuning hyperparameters on the test set (Ollion et al., 2024). Every study here was designed with pre-specified evaluation criteria and human baselines collected independently of the prompting process. This design discipline was costly but essential: without it, any observed alignment could reflect prompt optimization rather than model capability (Koh et al., 2021).

6.1 The Road Ahead: EVAS

We propose the *Ecological Validation of Artificial Simulacra* (EVAS) agenda as a framework for advancing empirically grounded evaluation of LLMs as behavioral agents. The psychometric fingerprinting methodology (Münker, 2025b) is modular: any Likert-scale instrument can be fingerprinted and compared across models and human populations. The multi-dimensional authenticity framework (Münker et al., 2026) provides a toolkit for moving beyond single-metric content evaluation. These are not endpoints but scaffolding for a validation culture the field currently lacks. Three extensions are most urgent.

Multi-Turn and Longitudinal Protocols: All psychometric studies here use single administrations; authentic human behavior is dynamic and context-sensitive (Park et al., 2023). Human behavioral consistency, or inconsistency, evolves across exchanges. Turn-based correlation analysis, analogous to fingerprinting but applied across conversation history, could track behavioral consistency in ways static questionnaires cannot, directly extending RQ_1 into dynamic interaction contexts (Sandhan et al., 2025).

Mechanistic Interpretability: Fingerprinting reveals that LLMs organize psychological constructs differently from humans; it does not reveal why. Circuit-level analysis (Dunefsky et al., 2024) may locate the architectural origins of structural misalignment, a prerequisite for designing training procedures that reduce it rather than masking it

through surface-level adaptation. Understanding the mechanism is necessary to know whether the fix lies in training data, instruction tuning, or architecture (Shen et al., 2024).

Expanded Linguistic and Cultural Coverage: This work covers 19 cultural contexts and two languages. The gaps, non-Latin script languages, Indigenous language communities, and cultural contexts underrepresented in digital text corpora (Myung et al., 2024; Adilazuarda et al., 2024), are precisely those where failures are most likely to be severe and least likely to be detected by researchers working in high-resource language contexts. Any universal claim about LLM human-simulacrum capability should be treated as unwarranted until coverage of these blind spots exists.

7 Conclusion

Box's dictum, all models are wrong, but some are useful, applies here, but only if applied with care. The myth of universal generalization is empirically falsified across multiple independent lines of evidence. LLMs cannot reliably serve as human simulacra through minimal prompt engineering; this claim fails across representation of political ideology, cross-cultural moral diversity, psychological factor structure, and social media communication. Each failure is documented not as a single anomaly, but through systematic, quantified evaluation against pre-specified human baselines. Each research question receives a negative answer on the null hypothesis:

RQ_1 LLMs do not represent the internal structure of psychological constructs in ways that align with human response patterns. The failure is structural rather than superficial, not noise around approximately correct means, but qualitatively different organizational principles revealed by inter-item correlation fingerprinting (Münker, 2025b) and Exploratory Graph Analysis (Golino and Epskamp, 2017). No tested model, regardless of architecture or parameter count, recovered the established factor structure of either the Moral Foundations Questionnaire (Graham et al., 2009) or the Humor Styles Questionnaire (Martin et al., 2003).

RQ_2 LLMs cannot reliably generate realistic social media content across languages (Münker et al., 2026). Fine-tuned English models approach human performance on specific metrics, but German models fail dramatically even after fine-tuning, and even the best-performing models remain distinguishable through multi-dimensional linguistic classification (Münker et al., 2026). The persistent detectability emerges from systematic signatures, morphosyntactic patterns, semantic distributions, lexical regularities (Ramos et al., 2003), that characterize generated text across all tested approaches.

RQ_3 Prompt-based approaches are insufficient; data-driven adaptation is necessary but not sufficient. Fine-tuning outperforms prompting on every evaluated task and language (Alizadeh et al., 2025; Møller and Aiello, 2024), establishing it as the minimum viable approach for deployment. But fine-tuning ameliorates rather than eliminates misalignment, and the residual gap is not a matter of insufficient training data or architecture, it reflects the fundamental representational distance between statistical text approximation and embodied, culturally situated human cognition (Shanahan, 2024)

7.1 The Constructive Conclusion

The more constructive conclusion is not that LLMs are useless for social science. It is that the field has lacked a validation culture adequate to distinguish genuine alignment from superficial mimicry (Larooij and Törnberg, 2025b). Individual studies documented failures in annotation stability, ideological bias in content generation, cultural homogenization in moral questionnaires, and detectable linguistic signatures in synthetic text. Read as a connected narrative, these converge on a unified theoretical argument: current LLMs are sophisticated pattern-completion systems whose outputs reflect the statistical regularities of training text, not the embodied, culturally situated, psychologically structured experience of human cognition (Bender et al., 2021; Ren et al., 2025).

Prompting fails for different reasons across domains, instability, homogenization, structural misalignment, lexical regularity, which together indicate that the limitation is not a single fixable bug but a feature of how these systems represent meaning (Dziri et al., 2024). Scaling does not resolve it; fine-tuning ameliorates but does not eliminate it (Lin, 2024). This has direct implications for research design. Because the deficit is structural rather than superficial, interventions that operate only at the surface level, longer prompts, richer persona descriptions, more elaborate few-shot examples (Min et al., 2022), are inherently insufficient. The question is not whether to engineer the prompt more

carefully but whether the validation framework can detect the remaining misalignment.

7.2 Practical Recommendations

Do not trust English benchmarks. Validation on English data does not transfer to other languages or cultural contexts; every deployment context requires its own validation study. The English-German-Dutch performance hierarchy documented in our studies is not an edge case, but a predictable consequence of the imbalance of training data (Hershcovich et al., 2022; Ryan et al., 2024). Researchers who deploy LLMs for non-English social science tasks on the basis of English benchmark scores are conducting invalid experiments without adequate grounds to know it (Heseltine, 2025).

Evaluate with independent, domain-matched test sets. Iterative prompt refinement without independent evaluation is methodologically equivalent to tuning hyperparameters on the test set (Ollion et al., 2024); prompts must be fixed before consulting evaluation data. Evaluation data must be drawn from the same domain, register, and language as the intended deployment setting: a classifier that achieves high precision in formal news text tells us little about behavior in informal social media discourse (Koh et al., 2021). Human baseline data should be collected under conditions that are truly independent of the prompt development process.

Do not trust surface-level output as evidence of deep alignment. High BLEU scores (Papineni et al., 2002) and plausible survey responses are compatible with deep structural misalignment; validation must incorporate structural measures along with surface-level metrics (Ye et al., 2025). As the results of the psychometric fingerprinting demonstrate, a model can produce questionnaire responses that match human means and fall within human variance ranges while organizing the underlying construct according to principles orthogonal to human psychology (Münker, 2025b).

Fine-tune, but validate independently. Fine-tuning should be treated as the minimum viable approach (Møller and Aiello, 2024; Alizadeh et al., 2025), but a fine-tuned model is a new system requiring its own validation pipeline (Koh et al., 2021). Fine-tuning domain-specific data consistently reduced visible symptoms of misalignment in our studies, but did not eliminate them: fine-tuned models remained detectably synthetic in multiple types of features and continued to show structural divergence from human baselines (Lin, 2024).

Acknowledge theoretical humility. Fundamental limitations arising from the lack of embodiment and cultural grounding (Shanahan, 2024) are not engineering problems that can be solved by larger models or better instructions. Our cross-study finding that model size does not reliably improve performance on socially-grounded tasks is not merely a negative empirical result but a signal about the nature of the deficit (de Wynter, 2025): the gap between statistical text approximation and embodied, culturally situated human cognition is unlikely to close through scaling procedures optimized for benchmark fluency (Bender et al., 2021).

Validate individually; there is no universal rule. Human-likeness is context-, language-, and domain-dependent (Adilazuarda et al., 2024); in each new application context, the degree of alignment must be empirically demonstrated, not assumed. A model validated for English survey simulation cannot be assumed to generalize to German social media content generation, and a model validated for political ideology representation cannot be assumed to generalize to humor style or moral foundations (Münker, 2025b,a).

7.3 Closing

The shift our work calls for is ultimately simple to state: from asking *"can LLMs replace humans?"* to asking *"under what validated conditions might LLMs serve as useful research components?"* That second question is harder to answer, requires domain-specific work, and does not generate universal rules. However, it is a scientifically defensible question, and every deployment of LLMs as human proxies should be treated as requiring the same standards applied to any other scientific instrument: calibration, validation, and explicit acknowledgment of limitations. The building block for this validation is where our work in this arc began. It is unfinished, but the direction is clear: validate before deploying, report failures alongside successes, and resist the conflation of impressive benchmark numbers with the far more demanding standard of genuine behavioral fidelity. Thus, in line with Box's observation, we argue that LLMs can be made useful for digital behavioral tasks, even though they remain fundamentally different from the humans they approximate.

Limitations

Our research arc covers eight studies conducted between 2023 and 2026 with numerous secondary literature, and quantitative findings reflect the capability of models from those development cycles. The specific performance gaps documented, between English and German, between prompted and fine-tuned approaches, between LLM and human factor structures, should be interpreted as snapshots rather than permanent verdicts. Future architectures or training procedures may narrow specific gaps, though the theoretical argument, that disembodied statistical text approximation is structurally distinct from embodied human cognition, is unlikely to be resolved by scale alone.

The psychometric strand of our work relies on established instruments (MFQ, MFQ-2, HSQ) with existing human baselines. While this grounding enables principled comparison, it also means our findings are specific to the constructs these instruments measure. Generalizability to other psychological dimensions or questionnaire formats requires separate validation. Similarly, the human baselines for MFQ and MFQ-2 were collected under specific sampling conditions; cross-study comparisons carry the usual caveats about population representativeness.

The social agent strand is limited to two languages (English and German) and one platform type (micro-blogging discourse on X). The dramatic English-German performance asymmetry suggests that findings from high-resource language evaluations should not be extrapolated to other language contexts without independent validation. Non-Latin script languages, low-resource languages, and communities underrepresented in digital text corpora remain unexamined, and these are precisely the contexts where failures are likely to be most severe.

Our detection framework demonstrates that synthetic content is classifiable, but the specific feature combinations and thresholds are calibrated to the dataset collected in 2023. As generation techniques advance, classifiers require retraining to remain valid. Detection performance should therefore be treated as a lower bound on the distinguishability of future synthetic content, not an upper bound.

Finally, our research exclusively examines open-source instruction-tuned models, a deliberate methodological choice ensuring reproducibility. Findings may not transfer directly to proprietary systems with different alignment procedures, though the theoretical and structural arguments we advance do not depend on specific implementations.

Ethical Considerations

Our work investigates the use of LLMs as substitutes for human participants in social science research, a practice with direct consequences for the validity of scientific claims and for the populations those claims are meant to represent. Our primary ethical concern is the harm caused by unvalidated deployment: when LLMs are used as human proxies without calibration, the resulting research may systematically misrepresent the populations it claims to study, particularly non-Western, non-English-speaking, and politically minority communities whose perspectives are demonstrably underrepresented in model outputs.

All human baseline data used in our psychometric studies was drawn from previously published datasets collected under their respective ethical review protocols. No new human subjects data was collected as part of this research arc.

The guardrail vulnerability work documented in our broader research program revealed that open-source models can be prompted to generate extremist and antisemitic content. We followed responsible disclosure norms and do not provide specific jailbreaking prompts in any publication. The findings are reported to motivate stricter validation standards rather than to enable misuse.

We are aware of a tension in this work: by documenting that LLM-generated content remains detectable, we simultaneously provide a benchmark against which evasion can be measured. We consider this tension unavoidable; the detection methods we describe are necessary for scientific validation, and concealing detection capabilities would not meaningfully impede determined adversarial actors while it would harm the research community's ability to assess authenticity.

The EVAS agenda we propose is intended to raise, not lower, the bar for deploying LLMs in sensitive social contexts. Practitioners who use our frameworks to establish domain-specific validation protocols advance a more responsible research practice; those who use benchmark performance as a substitute for validation do not.

Acknowledgments

We thank Nils Schwager, Kai Kugler, Fabio Sartori, Michael Heseltine, Sjoerd Stolwijk, and Simon Werner for our constructive discussions. This study was conducted with a financial contribution from the EU's Horizon Europe Framework (HORIZON-CL2-2022-DEMOCRACY-01-07) under grant agreement number 101095095.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Paul Balluff, Justin Chun-ting Ho, Johannes B Gruber, Sean Palicki, Alexis Palmer, Luca Rossi, Irina Shklovski, and Chung-hong Chan. 2026. Newer, larger, better? a critique of the unreflective llm adoption in communication research. *Political Communication*, pages 1–10.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

George EP Box. 1976. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Adrian de Wynter. 2025. Awes, laws, and flaws from today's llm research. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12834–12854.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniel Duncan. 2024. Does chatgpt have sociolinguistic competence? *Journal of Computer-Assisted Linguistic Research*, 8:51–75.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, and 1 others. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Hudson F Golino and Sacha Epskamp. 2017. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS one*, 12(6):e0174035.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Peter K Hatemi, Charles Crabtree, and Kevin B Smith. 2019. Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, 63(4):788–806.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. Association for Computational Linguistics.

Michael Heseltine. 2025. Comparing large language models for text classification: Model selection across tasks, texts, and languages.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Karen Sparck Jones. 1994. Natural language processing: a historical review. *Current issues in computational linguistics: in honour of Don Walker*, pages 3–16.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, and 1 others. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.

Maik Larooij and Petter Törnberg. 2025a. Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. *arXiv preprint arXiv:2504.03274*.

Maik Larooij and Petter Törnberg. 2025b. Validation is the central challenge for generative social simulation: a critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, and 1 others. 2025. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437.

Haocheng Lin. 2024. Designing domain-specific large language models: The critical role of fine-tuning in public opinion simulation. *arXiv preprint arXiv:2409.19308*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of research in personality*, 37(1):48–75.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Hadi Mohammadi, Yasmeen FSS Meijer, Efthymia Papadopoulou, and Ayoub Bagheri. 2025. Do large language models understand morality across cultures? In *Proceedings of the 2nd LUHME Workshop*, pages 30–39.

Anders Giovanni Møller and Luca Maria Aiello. 2024. Prompt refinement or fine-tuning? best practices for using llms in computational social science tasks. *arXiv preprint arXiv:2408.01346*.

Simon Münker. 2025a. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. In *Proceedings of 0 th Moral and Legal AI Alignment Symposium of the IACAP/AISB Conference*, page 61.

Simon Münker. 2025b. Fingerprinting llms through survey item factor correlation: A case study on humor style questionnaire. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 245–258.

Simon Münker. 2025c. Political bias in llms: Unaligned moral values in agent-centric simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138.

Simon Münker, Kai Kugler, and Achim Rettinger. 2025. Zero-shot prompt-based classification: topic labeling in times of foundation models in german tweets. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 53–63.

Simon Münker and Achim Rettinger. 2025. twony: A micro-simulation of the impact of osn mechanics on the emotionality of online discourse. In *Joint Proceedings of the ESWC 2025 Workshops and Tutorials*.

Simon Münker and Fabio Sartori. 2026. Guardrail vulnerabilities in open-source language models: Implications for democratic discourse and marginalized communities. *Hawaii International Conference on System Sciences (HICSS)*.

Simon Münker, Nils Schwager, and Achim Rettinger. 2026. Don't trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Simon Münker, Nils Schwager, Kai Kugler, Michael Heseltine, and Achim Rettinger. 2026. Next reply prediction x dataset: Linguistic discrepancies in naively generated content. *Preprint*, arXiv:2602.19177.

Lynnette Hui Xian Ng and Kathleen M Carley. 2025. Are llm-powered social media bots realistic? In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 14–23. Springer.

Megan Norris and Luc Lecavalier. 2010. Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of autism and developmental disorders*, 40:8–20.

Jum C Nunnally. 1975. Psychometric theory—25 years ago and now. *Educational Researcher*, 4(10):7–21.

Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary llms for research. *Nature Machine Intelligence*, 6(1):4–5.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 100–110.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*.

Karl Popper. 2014. *Conjectures and refutations: The growth of scientific knowledge*. routledge.

Aman Priyanshu and Supriti Vijay. 2024. The silent curriculum: How does llm monoculture shape educational content and its accessibility? *arXiv preprint arXiv:2407.10371*.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. Representation bias in political sample simulations with large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267.

Juan Ramos and 1 others. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2025. Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001.

Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. 2024. Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024(1):7115633.

Michael Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95.

Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. 2025. Cape: Context-aware personality evaluation framework for large language models. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:10648–10662.

Murray Shanahan. 2024. Simulacra as conscious exotica. *Inquiry*, pages 1–29.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, and 1 others. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281.

Lorenz Sparrenberg, Tobias Schneider, Tobias Deußer, Markus Koppenborg, and Rafet Sifa. 2024. Correcting systematic bias in llm-generated dialogues using big five personality traits. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3061–3069. IEEE.

Sjoerd B Stolwijk, Mark Boukes, Wang Ngai Yeung, Yufang Liao, Simon Münker, Anne C Kroon, and Damian Trilling. 2025. Can we use automated approaches to measure the quality of online political discussion? how to (not) measure interactivity, diversity, rationality, and incivility in online comments to the news. *Communication Methods and Measures*, pages 1–25.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 251–267.

Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. 2023. Welcome to the era of ChatGPT et al. - the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in

survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. In *Proceedings of the 30th Conference on Pattern Languages of Programs*. The Hillside Group.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.

Ziyun Yu, Yiru Zhou, Chen Zhao, and Hongyi Wen. 2025. An analysis of large language models for simulating user responses in surveys. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 242–259.

Towards Trustworthy AI-Mediated Communication Across Languages and Cultures

Dayeon Ki
University of Maryland
dayeonki@umd.edu

Abstract

A socio-technical gap exists between how NLP systems are developed and evaluated and how people use them in practice. To help close this gap, I propose a direction for scientific progress in NLP centered on advancing *trustworthy* AI-mediated communication between humans, using cross-lingual and cross-cultural interaction as a stress test for this goal—settings where common ground is hard-won, miscommunication can go unnoticed, and human users often lack the means to independently evaluate AI outputs. I outline a research agenda emphasizing two complementary requirements spanning both sides of the interaction. On the model side, I study how multilingual systems access and use knowledge across languages, and when they systematically privilege sources in certain languages. On the user side, I design decision-support mechanisms and evaluate how they shape user's reliance on imperfect outputs. Taken together, these results motivate future work for aligning multilingual NLP with real communicative practice, with the goal of building AI systems that more reliably serve diverse communities. This paper summarizes and draws heavily on my PhD thesis proposal.

1 Introduction

Communication across linguistic and cultural boundaries has long been central to human society—enabling trade, diplomacy, and everyday relationships (Pratt, 1991). Throughout history, humans have devised various ways to bridge these divides, including more recent developments in Natural Language Processing (NLP) for contributing multilingual Artificial Intelligence (AI) systems that mediate communication at scale. As these systems have been trained on increasingly diverse multilingual corpora (Conneau and Lample, 2019; Conneau et al., 2020), they have grown from supporting simple Machine Translation (MT) into *general*-purpose models used for a wider range of cross-lingual and cross-cultural support. People now use them in diverse communicative settings: from direct mediation of conversations between speakers without a shared language to more indirect assistance for interpreting unfamiliar cultural practices (Tamkin et al., 2024).

This shift also creates a mismatch with how progress in multilingual NLP is often measured. Despite widespread real-world use, many systems are still optimized and validated primarily via average performance on *decontextualized* benchmarks (Ackerman, 2000), which tend to reward fluent and adequate-sounding outputs (Bender et al., 2021a). Yet uneven capabilities and behaviors across languages remain visible (e.g., MT (Goyal et al., 2022), Question-Answering (QA, Li et al. (2025b))), and these systems are still under-tested in more open-ended *communicative* settings users now attempt, such as language learning and multi-turn interaction. Therefore, it remains difficult to characterize their impact in practice: how users perceive, interpret, and act on system outputs in real-world downstream decisions (Lee and See (2004)).

With this motivation, I argue that advancing trustworthy AI-mediated communication across languages and cultures is needed (§2.1). Here, *trustworthiness* is not achieved by building more capable models alone, nor by placing the burden on users to independently vet system outputs (§2.2). Instead, it requires complementary progress on both sides of the interaction: models that can help users establish common ground across differences in backgrounds and communicative goals (Clark and Schaefer, 1989), and user-facing designs and interaction strategies that support users' informed reliance on inevitably imperfect outputs. This paper investigates work along both threads.

On the model side, I begin with a core requirement for multilingual systems acting as communication mediators: knowledge parity across *any* languages and cultures (§3.1-3.2). The claim

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 45–59

is not simply that systems should achieve high task performance, but that they should *ground* their outputs in evidence in ways that give users equitable access to multilingual knowledge sources (Blasi et al., 2022), without systematically privileging certain languages during generation. To examine this, we develop a framework for measuring how models rely on multilingual evidence in retrieval-augmented generation (RAG) pipelines and empirically show a strong preference for English sources, even when they are irrelevant to user query (§3.3).

On the user side, trustworthy mediation also requires empowering the communication participants (§4.1). Yet using AI reliably is uniquely challenging in cross-lingual settings: users often lack practical ways to independently assess system outputs (e.g., evaluating a translation in a language they do not understand; Mehandru et al. (2023)), and these outputs are typically *not* direct predictions for downstream decisions, instead serving as inputs to many *implicit* judgments (e.g., Is the translation good enough to share with a friend? To translate an official document?) (§4.2). Accordingly, we propose a new decision support and design human-subject studies to compare it with alternatives, examining how each shapes users' decision-making and reliance on MT outputs (§4.3).

Together, these two threads point to a broader takeaway: advancing AI-mediated communication helps bring empirical findings from the model-side analysis to motivate the questions we ask in human studies, and interaction signals from human studies surface what model-side evaluations are missing, which creates a feedback loop between AI system development and real-world communicative practices. I conclude by distilling lessons from both threads and outlining forward-looking research directions toward multilingual AI systems that more reliably serve diverse communities (§5).

2 Background

I first establish the conceptual foundations for AI-mediated cross-lingual and cross-cultural communication (§2.1) and define trustworthiness in this context, framing it as a property jointly shaped by both the AI system and the human user (§2.2).

2.1 Why AI-Mediated Communication?

Communication may seem like a simple exchange of ideas through words that carry meaning (Reddy, 1979), yet it is fundamentally a collaborative process (Grice, 1975; Allwood, 1976; Bohm and Weinberg, 2004). As Clark and Brennan (1991) describe, successful communication requires coordination between interlocutors, both in content and in process, through the continual establishment of *common ground*: a set of mutual beliefs, presuppositions, and shared background knowledge that provide the necessary context for understanding (Stalnaker, 1972; Clark and Schaefer, 1989).

Consequently, when the interlocutors come from different linguistic or cultural backgrounds, establishing this common ground becomes substantially harder (Hall, 1959; Thomas, 1983; Hershcovich et al., 2022). For instance, when a Korean speaker wishes to communicate their plans for making *Songpyeon* at their grandparents' house for the upcoming *Chuseok*[1] with their American friend, several challenges arise. If they do not share a language, they would likely rely on translation tools even to initiate the conversation. Even when they understand the words, the concept of *Songpyeon* or *Chuseok* carry culturally specific associations with no direct counterpart in American culture, which requires the Korean speaker to explain or provide analogies to help their friend understand. In such cases, human interlocutors often construct a "third space," a communicative middle ground between two cultures (Planken, 2005), by, for example, describing *Chuseok* as a "Korean Thanksgiving" or using pragmatic strategies such as paraphrasing or providing brief clarifications (House, 2003).

As more complex cross-lingual and cross-cultural interactions grow common, NLP has researched to build multilingual AI systems as mediators that help people navigate these communicative gaps once bridged by human strategies at scale. However, these systems do more than transmit information, they actively *shape* how meaning is constructed and whose perspectives are amplified, and as they increasingly participate in human sense-making in the real-world, ensuring that AI-mediated cross-lingual and cross-cultural communication is trustworthy becomes an important goal.

2.2 What Constitutes Trustworthiness?

The notion of trustworthiness is not unique to AI; similar concepts appear across diverse disciplines, though their emphases differ. Across these fields, scholars generally agree perceived trustworthiness,

[1] *Chuseok* is the Korean harvest festival, which occurs on the 15th day of the 8th lunar month. *Songpyeon* is a traditional food eaten during *Chuseok*.

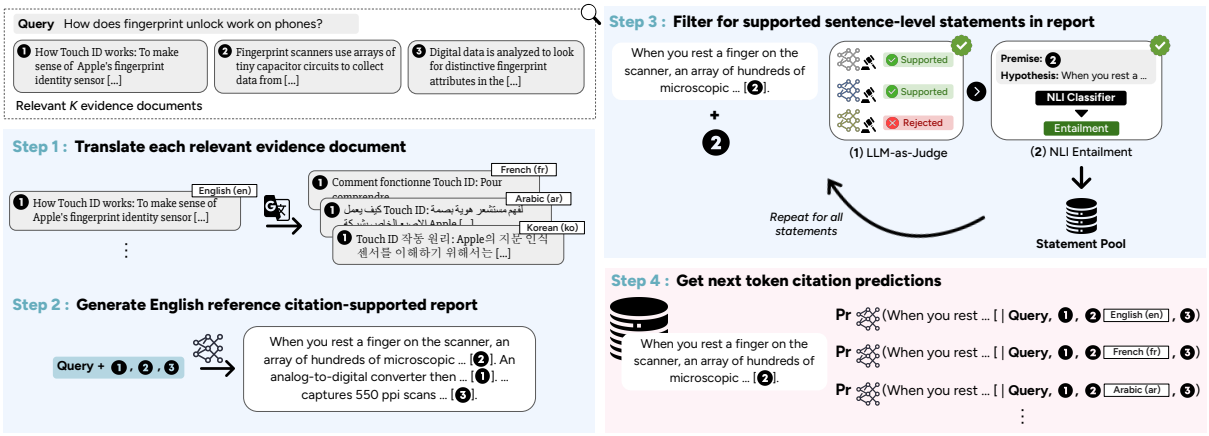


Figure 1: Overview of our approach measuring language preference using model internals. Synthetic data generation: Given an English query and its relevant evidence documents, we translate the documents into target languages and generate reference citation-supported reports (sentence-level statements with citation IDs). Measurement method: We detect language preference when next-token prediction accuracy for the correct citation ID drops as the language of the cited document varies.

whether of a person or a system, is multidimensional, encompassing *competence*, benevolence, and integrity: the ability to perform well, goodwill to act in others' interests, and adherence to ethical principles such as honesty and fairness (Mayer et al., 1995; Mishra, 1996; Rousseau et al., 1998). In sociolinguistics, particularly in linguistic anthropology and intercultural communication, trustworthiness is closely tied to authentic representation—accurately conveying social and cultural meanings—as well as to communicative competence, the ability to use language appropriately in specific social contexts (Gumperz, 1970, 1982; Fox, 1997). In Human-Computer Interaction (HCI), researchers have developed various ways to measure human trust in AI systems, often through trust-related behaviors (Vereschak et al., 2021).[2] One central indicator for trust is *appropriate reliance*, where a user's level of trust matches the system's true capabilities (Lee and See, 2004).

Building on these perspectives, I define trustworthy AI mediators in cross-lingual and cross-cultural communication as systems that (1) possess and appropriately use the knowledge needed to establish common ground between users of different languages and cultures, and (2) support users in making informed decisions from system outputs, even when those are in languages they are not fluent in.

The following sections study each thread in turn. I first lay the groundwork for what hinders and what is required for balanced knowledge coverage across languages and cultures. I then introduce a framework for measuring models' language preferences in evidence selection during RAG (§3). Next, I discuss how user reliance on AI systems has been conceptualized, and propose and evaluate decision supports in helping users engage more reliably with imperfect outputs (§4).

3 Model-Side Thread

I review the prevailing paradigm for understanding knowledge disparities across languages and its limitations (§3.1-3.2), then introduce a framework that probes model internals to examine language preference in multilingual knowledge access (§3.3).

3.1 Understanding Knowledge Disparities

Despite training data becoming increasingly linguistically diverse, large-scale pre-training remains heavily Anglocentric, with English comprising 80–90% of the corpora (Touvron et al., 2023; Grattafiori et al., 2024). As Joshi et al. (2020) point out, current AI systems still disproportionately center on a small subset of languages—often geographically clustered and drawn from a few dominant language families, such as English, Chinese, and Spanish—in both training and evaluation. Meanwhile, many others, such as Zulu or Fijian, remain excluded, creating a typological echo chamber. This persistent "data problem" (Aharoni et al., 2019) is represented well in the language distributions of pre-training corpora for widely used models, which remain heavily skewed toward English

[2] In HCI, "trust" refers to human users' reactions or attitudes toward an AI system, whereas "trustworthiness" concerns the properties of the system itself (Vereschak et al., 2024).

and other high-resource languages (e.g., Chinese).[3]

This disparity in language representation within training data leads to performance gaps across tasks when queries are posed in different languages. Accuracy and perplexity often worsen for certain languages, even when models are evaluated on the same examples translated into different languages (Zhang et al., 2023; Jin et al., 2024; Li et al., 2025b). Beyond performance disparities, uneven language coverage further results in unequal access to knowledge across languages (Bender et al., 2021b; Yu et al., 2022; Feng et al., 2024), wherein information expressed in higher-resource languages becomes more accessible and frequently amplified (Phillipson, 2018). As a result, models exhibit systematic language preference—a tendency to favor certain languages when accessing or eliciting knowledge—which ultimately creates differences in response quality (Boughorbel and Hawasly, 2023), consistency (Dong et al., 2025), and dispute resolution (Li et al., 2024) for users across languages.

3.2 Limitations of Prior Work

Prior work has examined language preference in RAG pipelines: whether models tend to retrieve (Telemala and Suleman, 2022; Yang et al., 2024; Amiraz et al., 2025) or rely on evidence written in certain languages during generation (Park and Lee, 2025; Sharma et al., 2025; Li et al., 2025a). Existing approaches to measure language preference in multilingual RAG (mRAG), however, often fail to capture citation correctness. In short-form RAG, preference has been estimated via information overlap (Sharma et al., 2025) or embedding similarity (Park and Lee, 2025), which do not directly account for correctness. In long-form RAG, where outputs contain in-line citations (Zheng et al., 2025; Xu and Peng, 2025), preference has typically been measured by comparing citation frequencies against the language distribution of retrieved documents. This signal is coarse and confounded by the relevance and informativeness of multilingual sources (C1) and in-line citations are prone to hallucinations (Gao et al., 2023; Zhang et al., 2025), making it unclear whether observed preferences reflect true attribution or spurious citations (C2).

[3]For instance, GPT-3 (Brown et al., 2020) has 92.7% of the training tokens in English; LLaMA-2 (Touvron et al., 2023) and LLaMA-3 (Grattafiori et al., 2024) has 89.7% and approximately 92% of pre-training data in English, respectively.

3.3 Our Approach

Method. In Ki et al. (2026), we address both challenges by proposing a controlled methodology for measuring language preference using model internal metrics. As illustrated in Figure 1, we first construct a synthetic multi-parallel dataset of relevant documents, which allows us to isolate the effect of language while controlling for other factors such as document content and relevance (**Step 1+2**; addresses C1). Citation correctness is then verified through a two-step filtering process (**Step 3**; addresses C2). Next, we compare the accuracy of next token citation predictions (e.g., predicting "2" for document ID 2) while varying the language of the same cited document and keeping other variables fixed, including the language of remaining documents, document positions in the input context, and the query language (**Step 4**). Differences in citation accuracy between languages indicate a preference for the higher-accuracy language.

Experiment Setup. We use ELI5 dataset (Fan et al., 2019) of long-form questions from the Reddit forum "Explain Like I'm Five". We study eight languages representing diverse range of resource levels and number of speakers: Arabic (ar), Bengali (bn), Spanish (es), French (fr), Korean (ko), Russian (ru), Swahili (sw), and Chinese (zh). We use six open-weight models varying in degree of multilinguality: LLAMA-3.1 8B and LLAMA-3.3 70B (Grattafiori et al., 2024), QWEN-3 8B and 14B (Yang et al., 2025), GEMMA-3 27B (Team et al., 2025), and AYA23 8B (Aryabumi et al., 2024).

Results. We address the overarching question: Do models preferentially cite documents in certain languages during long-form mRAG? To further inform building more robust systems, we empirically address two questions: (1) What factors amplify language preference? and (2) Is citation behavior driven more by document relevance or language? Our main findings can be summarized as follows:

- **Evidence of an English preference:** As shown in Table 1, we observe a pronounced tendency to cite English documents (with the highest citation accuracy) when the query is in English across all models. This preference amplifies when the cited document is in a lower-resource language (e.g., Bengali, Swahili).
- **Language outweigh relevance:** We show that models frequently cite English documents even when they are irrelevant to the query as

Language	LLaMA-3.1 8B	Qwen-3 8B	Aya23 8B	Qwen-3 14B	Gemma-3 27B	LLaMA-3.3 70B
English	67.4	62.6	60.0	83.0	86.2	85.9
French	62.9 (-4.49)	48.4 (-14.2)***	48.5 (-11.5)***	76.0 (-7.04)***	79.0 (-7.21)**	77.4 (-8.50)***
Russian	62.1 (-5.30)*	50.4 (-12.2)***	48.1 (-11.9)***	74.8 (-8.17)***	77.1 (-9.12)***	74.5 (-11.4)***
Spanish	62.1 (-5.32)*	51.9 (-10.7)***	49.1 (-10.9)***	77.4 (-5.61)*	80.2 (-6.04)**	76.0 (-9.90)***
Korean	61.7 (-5.68)*	49.7 (-12.9)***	42.2 (-17.8)***	70.3 (-12.7)***	77.5 (-8.71)***	69.2 (-16.7)***
Chinese	59.9 (-7.51)*	49.2 (-13.4)***	46.3 (-13.7)***	73.5 (-9.49)***	75.4 (-10.8)***	74.1 (-11.8)***
Arabic	59.5 (-7.91)**	47.6 (-15.0)***	43.2 (-16.8)***	72.6 (-10.4)***	78.4 (-7.82)***	67.3 (-18.6)***
Bengali	56.6 (-10.8)***	41.3 (-21.3)***	27.2 (-32.8)***	65.4 (-17.6)***	77.9 (-8.33)***	68.8 (-17.1)***
Swahili	53.0 (-14.4)***	30.4 (-32.2)***	22.4 (-37.6)***	54.7 (-28.3)***	74.0 (-12.2)***	67.3 (-18.6)***

Table 1: Citation accuracies (%) by model and language. We present mean accuracy values with the difference to English accuracy in subscript. Pairwise two-sided t-tests with Bonferroni correction are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant. Color indicates the magnitude of accuracy difference: largest, second largest, others. Columns: increasing model size; rows: decreasing accuracy difference (of first model).

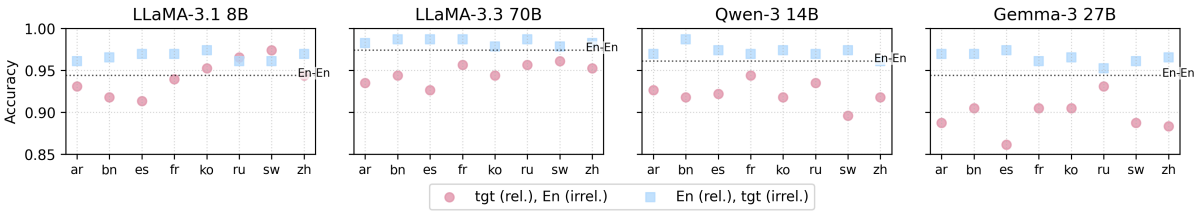


Figure 2: Citation accuracy per model with one relevant (rel.) and one irrelevant evidence document (irrel.) in different languages. We test three different conditions: (1) **En-En**: Both relevant and irrelevant documents are in English; (2) **tgt-En** (●): Relevant document in the target language, irrelevant document in English; (3) **En-tgt** (■): Relevant document in English, irrelevant document in the target language. Models trade off document relevance for language preference.

shown in Figure 2, suggesting that language itself exerts a stronger influence than document relevance in long-form mRAG.

Takeaways. Taken together, our findings show that model internals can expose systematic citation behaviors in mRAG systems, raising inclusivity concerns in multilingual knowledge access, where models not only *favor* certain languages, but may also *trade-off* evidence quality in doing so. These disparities illustrate how imbalanced language representation in training data shapes what knowledge multilingual AI systems access, prioritize, and ultimately use to justify their outputs.

4 User-Side Thread

The model-side thread reveals systematic imperfections in how multilingual AI systems access and use knowledge, but whether and how these impact real users attempting to communicate across languages remains an open question that controlled benchmarks cannot answer alone. Addressing this requires shifting focus from model to user behavior—from *what* the system outputs to *how* users interpret, act on, and are misled by those outputs.

I first review how prior work has measured and designed human interaction with, and reliance on, AI systems (§4.1), and how cross-lingual communication setup poses unique challenges (§4.2). I then propose a new form of decision support and evaluate its impact through human studies (§4.3).

4.1 Human Reliance on AI Systems

People increasingly use AI as decision support in everyday settings, such as video recommendation or search autocompletion (Bunt et al., 2012), and photo organization (Amershi et al., 2019). In these contexts, AI typically plays a supportive role by providing *direct* predictions or explanations in various formats (Bussone et al., 2015; Buçinca et al., 2021; Wang et al., 2022). Such feedback is intended to help users calibrate when and how much to rely on system outputs (Lai et al., 2023).

A growing body of work therefore examines the nature of human reliance on AI systems (Lai et al., 2023), especially in decisions involving risk and uncertainty (Jacovi et al., 2021). To operationalize trust in measurable terms, prior works often study users' reliance behavior (de Fine Licht and Brülde, 2021), commonly defined as the "decision to follow someone's recommendation" (Vereschak et al., 2021). As such, a central challenge in human-AI

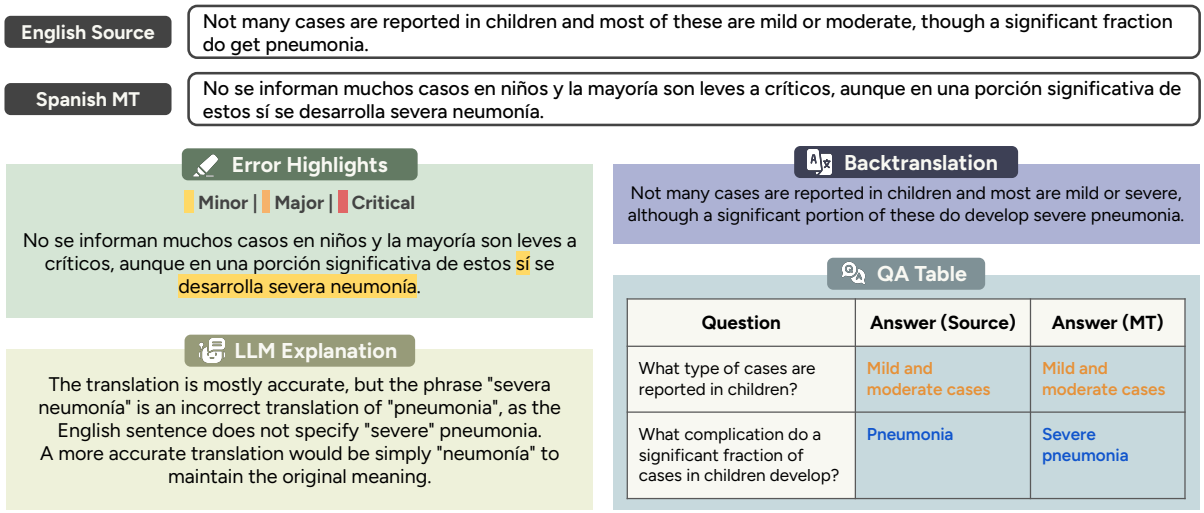


Figure 3: Overview of our tested quality feedback in human study. During the AI-Assisted step, each treatment group participant is presented with an English source, Spanish MT, and one of four randomly assigned quality feedback types. Error Highlights: We adopt a QE system, xCOMET-XXL (Guerreiro et al., 2024), to generate error annotations and display error spans with color-coded legend; LLM Explanation: We use LLAMA-3.3 70B-generated textual explanations of overall MT quality; Backtranslation: We use Google Translate to backtranslate Spanish MT to English; QA Table: We use LLAMA-3.3 70B.

interaction is achieving *appropriate* reliance: helping users accept correct AI advice while rejecting incorrect ones (Eckhardt et al., 2025).

4.2 Limitations of Prior Work

While human reliance on AI systems has been extensively studied in traditional AI-assisted decision-making tasks, extending this concept to cross-lingual communication introduces new challenges. For example, in the context of MT, the role of the AI system (i.e., MT system) takes on a different character since (1) monolingual users often lack the mechanisms to reliably assess MT quality, and (2) the AI prediction (i.e., MT output) is *not* a direct prediction for the user's decision-making task. Given this difference, we focus on quality feedback, which are more generic assessments of MT quality rather than direct recommendations, and ask whether users can rely on such feedback to make more informed decisions.

Various forms of quality feedback have been proposed, including backtranslation (Agrawal et al., 2022), error highlighting that flags problematic spans in MT (Rubino et al., 2021; Briakou et al., 2023), and textual explanations (Fomicheva et al., 2022; Xu et al., 2023). However, such feedback can be hard to interpret and often fails to convey how mistranslations affect real users (C1). Moreover, only a small number of human studies have evaluated how quality feedback influences user decision-making and reliance in MT (Zouhar et al., 2021; Mehandru et al., 2023), where findings remain mixed, and systematic comparisons against newer feedback mechanisms are still lacking (C2).

4.3 Our Approach

Method. In Ki et al. (2025a), we propose a question generation and answering (QG/QA) framework grounded in the idea that a translation is unreliable if key questions about the source yield different answers when derived from the source text or the backtranslated MT. We hypothesize that these QA pairs foreground the *functional* consequences of potential errors, rather than offering a mechanistic account of what is wrong (Lombrozo and Wilkenfeld (2019); addresses C1). This design also aligns with the view of explanations as *social*—facilitating knowledge transfer through interaction, where users can weigh evidence in light of their existing beliefs (Miller, 2019).

We then conduct a between-subjects human study in Ki et al. (2025b) with 91 English-speaking monolingual participants, where they are asked to decide whether Spanish MT outputs are safe to share with a hypothetical Spanish-speaking neighbor (i.e., "Is the Spanish translation good enough to safely share with your Spanish neighbor?"). For each of 20 examples, participants first make a binary shareability judgment (Safe to share as-is/Needs bilingual review before sharing)[4] and self-

[4]We use the notion of shareability to capture not only perceived MT quality but also the potential *risk* of miscommu-

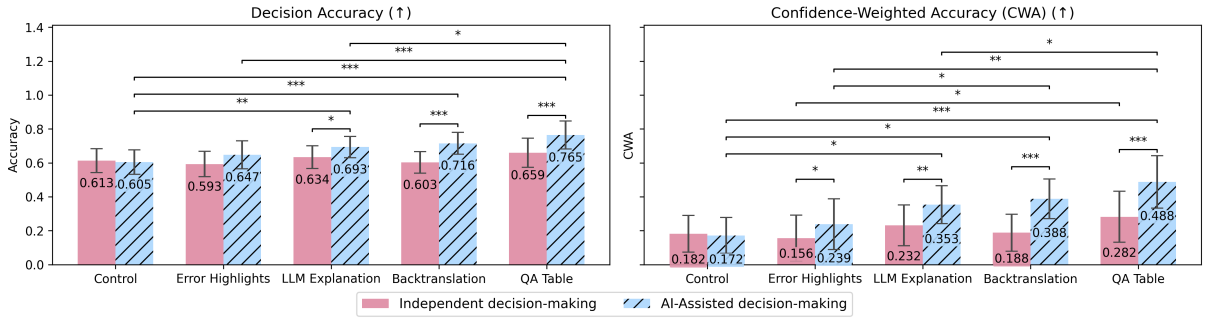


Figure 4: Average decision accuracy (*left*) and CWA (*right*) for each condition. Paired-sample t-tests are performed to compare independent and AI-assisted performance and linear mixed-effects ANOVA with Bonferroni corrections. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant.

report confidence (Independent) and then reassess the same example with a randomly assigned quality feedback (AI-Assisted). We compare four types of quality feedback as shown in Figure 3, grouped by their degree of explicitness: error highlights and LLM explanation provide *explicit* assessments of MT quality, whereas backtranslation and QA table offer *implicit* support by guiding participants to compare MT inputs and outputs (addresses C2).

Experiment Setup. We construct 20 English-Spanish examples in the COVID-19 domain by introducing targeted linguistic perturbations into Spanish MT outputs. We categorize each perturbations as either minor or critical based on the potential real-world impact of the resulting error. We recruit 91 U.S.-based participants who self-identify English as their first, primary, and fluent language. Self-reported monthly MT usage varies: 5 participants (5.49%) never used MT, 24 (26.4%) rarely used it, 32 (35.2%) used it sometimes, 19 (20.9%) often, and 11 (12.1%) used MT almost every day.

For dependent variables, we measure (1) decision accuracy, by comparing participants' shareability judgment to the gold label; (2) confidence-weighted accuracy (CWA), which combines accuracy with self-reported confidence via confidence weighting (Ebel, 1965; Mehandru et al., 2023) to capture whether participants made the correct decision weighted by their confidence in that decision; and (3) switch percentage, the rate at which participants change their decision after viewing AI feedback (Srivastava et al., 2022; He et al., 2023). Following Schemmer et al. (2023), we compute

nication in high-stakes contexts. This framing aligns with how people often make decisions in practice: while the choice to share or not is often made implicitly in real world, our study makes this decision more explicit, yet still allows participants to make their own judgment as they naturally would.

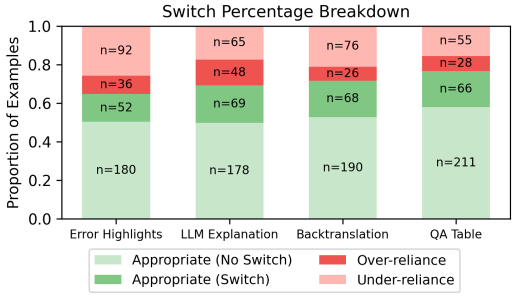


Figure 5: Switch percentage breakdown per quality feedback type. Each shows appropriate, over-, and under-reliance.

three reliance outcomes: (i) over-reliance, switching from a correct to an incorrect decision after feedback; (ii) under-reliance, failing to switch from an incorrect to a correct decision after feedback; and (iii) appropriate reliance, either correcting an incorrect decision after feedback (switch) or maintaining a correct one (no switch).

Results. Our main findings are as follows:

- **Providing any feedback helps:** As shown in Figure 4, all four treatment conditions improve decision accuracy and CWA in the AI-Assisted step relative to the Independent step.
- **QA table yields the largest gains:** Among the four feedback types, QA table shows the strongest and most consistent effects overall.
- **Implicit feedback has lower over-reliance:** As illustrated in Figure 5, both implicit quality feedback (backtranslation and QA table) yield higher appropriate reliance and lower over-reliance than the explicit feedback types. Across conditions, participants are more likely to *maintain* their initial decisions than to switch: under-reliance consistently exceeds over-reliance, and appropriate reliance (no switch) exceeds (switch).

Takeaways. Together, our findings suggest that using QA pairs as quality feedback is a promising direction for supporting cross-lingual communication. More broadly, these results underscore the value of interpretable, *user-driven* feedback mechanisms that help users construct their own functional explanations—reasoning grounded in the goals and consequences of an AI output—to determine how and when to rely on it for safe and effective use (Lombrozo and Wilkenfeld, 2019; Schoeffer et al., 2024), rather than having the system explicitly prescribe what to do.

5 Concluding Thoughts

I propose a direction for scientific progress in NLP centered on advancing AI-mediated communication across languages and cultures. This agenda requires two complementary threads: (1) models that help users establish common ground across differences in background and communicative goals, and (2) interaction strategies that support users' informed reliance on inevitably imperfect outputs.

Lessons Learned. On the model side (§3), the central lesson is that a trustworthy AI mediator requires knowledge parity *in use*—counteracting the imbalanced language representation in training data. The core issue here is not simply whether a model can achieve high multilingual task performance, but whether it provides equitable access to multilingual sources when grounding its claims. By measuring language preference through model internal signals in controlled settings, this thread moves beyond coarse proxies (e.g., citation frequency or surface overlap) toward diagnostics that can isolate when language itself drives evidence selection and when it induces trade-offs with evidence quality. This makes visible a failure mode: even when multilingual knowledge sources exist, systems may still systematically privilege those written in particular languages, shaping which evidence is surfaced and which perspectives are amplified. Methodologically, the broader implication is that evaluating multilingual systems as communication mediators requires attention to the full pipeline, retrieval, selection, generation, and reasoning, to understand not only *what* a model outputs, but *how* it arrives at its predictions.

On the user side (§4), the central lesson is that trustworthy mediation is not achieved by model capability alone; it also requires interaction designs that preserve users' *agency* (Shneiderman, 2022): users need support to assess risks, goals, and consequences for themselves, rather than prescribing decisions. We study this in a realistic cross-lingual communication scenario, where reliable use is uniquely challenging since users often lack practical ways to assess outputs and system outputs do not map directly onto downstream decisions. We propose QA pair-based feedback and show, through human-subject studies, that this interpretable, user-driven support can improve decision accuracy with better-calibrated confidence and promote more appropriate reliance on MT outputs.

Looking Forward. These lessons motivate several future research directions that better aligns multilingual NLP with communicative practice:

- **Evaluate systems in communicative contexts, not just on benchmarks.** For mediator-like systems, we should complement model-centric metrics with user-centered outcomes (e.g., decision accuracy, over/under-reliance, and downstream risk) and explicitly measure communicative failure modes that benchmarks often miss (e.g., misunderstandings, repair behavior, and conversation breakdowns).
- **Understand systems' *reasoning* processes.** A promising next step is to characterize *why* models behave differently across languages, both in evidence use and final outputs, and to translate that understanding into user-facing guidance. Reasoning traces are one candidate bridge: if designed carefully, they could help users retrace and oversee why the model made a particular prediction, and potentially enable them to learn patterns they can extrapolate to future decisions (Holzinger et al., 2023).
- **Build AI literacy for human users.** While developers may be aware of a system's weaknesses, these limitations are rarely communicated to end users. To enable functional, user-driven feedback, interfaces should explicitly communicate the systems' capabilities and blind spots through actionable signals, clarifying what the system can and cannot be expected to do in the current setting.
- **Examine effects on interpersonal dynamics.** Beyond individual decision-making, AI-mediated communication can reshape how people coordinate with one another. Interpersonal adaptation—how interlocutors adjust phrasing, clarify intent, and negotiate meaning—is central to establishing common

ground, and AI mediation may shift when and how this occurs. A key direction is to evaluate mediator systems for their effects on conversational dynamics (e.g., attribution, accountability, perceived effort) and on longer-term interpersonal relationships.

The broader goal is to move multilingual NLP toward systems that reliably help people establish common ground and make reliable decisions across languages and cultures. This work develops methods, measurements, and interaction designs aimed at supporting diverse communities in practice.

Acknowledgments

I would first like to thank my PhD thesis advisor Marine Carpuat, as well as my committee members Kevin Duh, Rachel Rudinger, and Fumeng Yang. Many thanks also to all coauthors on the projects reviewed in this paper, including (in alphabetical order) Daniel Khashabi, Dawn Lawrie, Eugene Yang, and Paul McNamee. Last but not least, thanks to all the members of the CLIP lab at the University of Maryland for their constructive feedback and support in developing this work.

References

Mark S Ackerman. 2000. The intellectual challenge of cscw: the gap between social requirements and technical feasibility. *Human–Computer Interaction*, 15(2-3):179–203.

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via back-translation at the WMT 2022 quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 593–596, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Jens Allwood. 1976. *Linguistic Communication as Action and Cooperation*. Ph.D. thesis, University of Göteborg, Department of Linguistics, Göteborg, Sweden.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan. 2025. The cross-lingual cost: Retrieval biases in RAG over Arabic-English corpora. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 69–83, Suzhou, China. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *Preprint*, arXiv:2405.15032.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021a. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

David Bohm and Rachel A. Weinberg. 2004. *On Dialogue*, 2nd edition. Routledge.

Sabri Boughorbel and Majd Hawasly. 2023. Analyzing multilingual competency of LLMs in multi-turn instruction following: A case study of Arabic. In *Proceedings of ArabicNLP 2023*, pages 128–139, Singapore (Hybrid). Association for Computational Linguistics.

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. 2023. Explaining with contrastive phrasal highlighting: A case study in assisting humans to detect translation differences. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, page 169–178, New York, NY, USA. Association for Computing Machinery.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics*, ICHI '15, page 160–169, USA. IEEE Computer Society.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Karl de Fine Licht and Bengt Brülde. 2021. On Defining "Reliance" and "Trust": Purposes, Conditions of Adequacy, and New Definitions. *Philosophia*, 49:1981–2001.

Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang. 2025. Evaluating and mitigating linguistic discrimination in large language models: perspectives on safety equity and knowledge equity. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

Robert L Ebel. 1965. Confidence weighting and test reliability. *Journal of Educational Measurement*, 2(1):49–57.

Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. 2025. A survey of ai reliance. *ACM Comput. Surv.*, 58(6).

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, Miami, Florida, USA. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

Christine Fox. 1997. The authenticity of intercultural communication. *International Journal of Intercultural Relations*, 21(1):85–103.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste

Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich

Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press, New York.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

John J. Gumperz. 1970. *Sociolinguistics and Communication in Small Groups*. University of California Language-Behavior Research Laboratory, Berkeley, CA. Language-Behavior Research Laboratory Report.

John J. Gumperz. 1982. *Discourse Strategies*. Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge. Originally published 1982; online publication November 2009.

Edward T. Hall. 1959. *The Silent Language*. Doubleday, Garden City, NY.

Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Bettina Finzel, Ute Schmid, and Heimo Mueller. 2023. Toward human-level concept learning: Pattern benchmarking for ai algorithms. *Patterns*, 4(8):100788.

Juliane House. 2003. English as a lingua franca and its influence on discourse norms in other languages. In Gunilla Anderman and Margaret Rogers, editors, *Translation Today: Trends and Perspectives*, pages 168–179. Multilingual Matters.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn Lawrie, and Kevin Duh. 2026. Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG. In *Forty-third International Conference on Machine Learning*.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025a. AskQE: Question Answering as Automatic Evaluation for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17478–17515, Vienna, Austria. Association for Computational Linguistics.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025b. Should I Share this Translation? Evaluating Quality Feedback for User Reliance on Machine Translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12080–12103, Suzhou, China. Association for Computational Linguistics.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1369–1385, New York, NY, USA. Association for Computing Machinery.

John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80. Original work published 2004.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is your, my land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025a. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4215–4241, Vienna, Austria. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025b. Language ranker: a metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Tania Lombrozo and Daniel A. Wilkenfeld. 2019. *Mechanistic versus functional understanding*, chapter 11. New York, NY: Oxford University Press.

Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Aneil K. Mishra. 1996. Organizational responses to crisis: The centrality of trust. In Roderick M. Kramer and Thomas R. Tyler, editors, *Trust in Organizations: Frontiers of Theory and Research*, pages 261–287. Sage Publications, Newbury Park, CA.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating Language Preference of Multilingual RAG Systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5647–5675, Vienna, Austria. Association for Computational Linguistics.

Robert Phillipson. 2018. Linguistic imperialism. *The Encyclopedia of Applied Linguistics*.

Brigitte Planken. 2005. Managing rapport in lingua franca sales negotiations: A comparison of professional and aspiring negotiators. *English for Specific Purposes*, 24(4):381–400.

Mary Louise Pratt. 1991. Arts of the contact zone. *Profession*, pages 33–40. Accessed 25 Oct. 2025.

Michael J. Reddy. 1979. The conduit metaphor. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press, Cambridge.

Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422.

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–18. ACM.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2025. Faux Polyglot: A Study on Information Disparity

in Multilingual Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8090–8107, Albuquerque, New Mexico. Association for Computational Linguistics.

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

Divya K. Srivastava, J. Mason Lilly, and Karen M. Feigh. 2022. Improving human situation awareness in AI-advised decision making. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, pages 1–6.

Robert C. Stalnaker. 1972. Assertion. In Peter Cole, editor, *Pragmatics*, pages 315–332. Brill.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, and Deep Ganguli. 2024. Clio: Privacy-preserving insights into real-world ai use. *Preprint*, arXiv:2412.13678.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Joseph P Telemala and Hussein Suleman. 2022. Language-preference-based re-ranking for multilingual swahili information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 144–152.

Jenny Thomas. 1983. Cross-Cultural Pragmatic Failure. *Applied Linguistics*, 4(2):91–112.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,

Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Jialu Wang, Yang Liu, and Xin Wang. 2022. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland. Association for Computational Linguistics.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *Preprint*, arXiv:2506.12594.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024. Language fairness in multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2487–2491.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China. Association for Computational Linguistics.

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021. Backtranslation feedback improves user confidence in MT, not quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 151–161, Online. Association for Computational Linguistics.

Challenging Quadratic Attention - A Holistic View On the Rise of Alternative Language Model Architectures

Alexander M. Fichtl, Jeremias Bohn, Josefin Kelber, Edoardo Mosca and **Georg Groh**
Social Computing Group
Technical University of Munich
Boltzmannstraße 3, 85748, Garching, Germany
{alexander.fichtl, jeremias.bohn, josefin.kelber, edoardo.mosca, georg.groh}@tum.de

Abstract

Transformers have dominated sequence processing tasks for the past seven years—most notably language modeling. However, the inherent quadratic complexity of their attention mechanism remains a significant bottleneck as context length increases. We review and distill the recent efforts to overcome this bottleneck, including advances in (sub-quadratic) attention variants, recurrent neural networks, state space models, and hybrid architectures. We critically analyze approaches regarding compute and memory complexity, benchmark results, and fundamental limitations to assess whether the dominance of pure-attention transformers may soon be challenged, which we consider possible, particularly in domain-specific and edge-device applications.

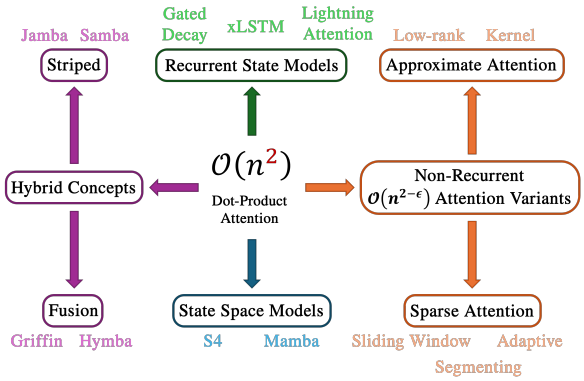


Figure 1: The four types of dot-product attention alternatives identified in our survey, including examples. We further divide hybrid concepts (striped and fusion hybrids), and sub-quadratic attention variants (approximate and sparse attention) each in two major classes.

1 Introduction

The transformer architecture is a foundational breakthrough in *Natural Language Processing* (NLP) (Vaswani et al., 2017), forming the backbone of most *Large Language Models* (LLMs) (Brown et al., 2020) and is a reliable architecture choice for predictable performance scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). Its self-attention mechanism (Bahdanau et al., 2015) projects inputs into *queries* (Q), *keys* (K), and *values* (V), enabling efficient pairwise token interactions:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Despite providing direct $\mathcal{O}(1)$ paths between any pair of tokens, computing the full $n \times n$ attention matrix incurs $\mathcal{O}(n^2)$ time complexity, increasing latency and compute costs as the input length n grows (Vaswani et al., 2017). While efficiency improvements to standard attention mostly focused on caching and memory layout (see Appendix A.5), its core problems have motivated research efforts into sub-quadratic sequence-modeling operators to replace attention, aiming to improve efficiency while retaining strong task performance. These include sub-quadratic attention variants (Katharopoulos et al., 2020), *Recurrent Neural Networks* (RNNs) (Beck et al., 2024), *State Space Models* (SSMs) (Gu and Dao, 2023; Gu et al., 2022), and hybrids thereof (De et al., 2024).

This paper reports on the overarching narrative of developing alternatives to transformers, reviews the most impactful milestones and new primitives, and examines whether the transformer dominance may soon be challenged. Our main contributions are:

(1) A review of the most relevant (sub)-quadratic attention variants, recurrent state models, SSMs, and hybrid architectures. An overview can be found in Figure 1.

(2) A comparative analysis of time and memory complexity for training and inference of sequence-modeling mechanisms, as well as reported benchmark results for SOTA models.

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 60–81

(3) A critical analysis of strengths, tradeoffs, and limitations, with an informed perspective on when and where pure attention-based transformers may be surpassed.

Our methodology is described in Appendix A.1.

2 Related Review Work

While several recent and concurrent works overlap with aspects of our scope, they differ in focus and conclusions. Schneider (2025) discusses hypothetical post-transformer architectures without restricting to sub-quadratic complexity or state-of-the-art performance. Wang et al. (2024c) review approaches for handling longer input sequences, and Tiezzi et al. (2025) examine alternative architectures from a recurrent-processing perspective.

Several surveys provide overviews of efficient transformers and LLMs in general (Tay et al., 2022; Miao et al., 2025; Huang et al., 2024; Wan et al., 2024; Miao et al., 2024; Tang et al., 2024), but often emphasize linear attention variants when considering alternative architectures. Focused surveys address specific subgroups, such as SSMs (Somvanshi et al., 2025; Wang et al., 2024b) and recurrent models (Tiezzi et al., 2024), or specific domains like computer vision (Patro and Agneeswaran, 2025) and time series forecasting (Kim et al., 2025), whereas our emphasis is on NLP tasks and all sub-quadratic alternatives to attention-based models. Existing surveys also frequently give extensive full historical lineages of the discussed models (e.g., Sun et al., 2025), while we focus on practical relevance in recent applications and research. Finally, Strobl et al. (2024) provide an overview of works on transformer expressivity, which relates to our discussion of architectural limitations in Section 8.

3 Non-Recurrent $\mathcal{O}(n^{2-\epsilon})$ Attention

Categorizing sub-quadratic attention alternatives is challenging due to overlapping ideas and mechanisms. We organize them as non-recurrent attention variants, recurrent state models, SSMs, and Hybrids according to their main design motivation, though some fall into several categories.

3.1 Approximate Attention

These mechanisms, including linear attention, reduce computational cost by using approximations such as kernel functions or low-rank factorization. Kernel-based linear attention reformulates self-attention as a linear dot-product in feature space, achieving $\mathcal{O}(n)$ complexity (Katharopoulos et al., 2020; Zhuoran et al., 2021). However, a poorly chosen kernel can result in reduced expressivity. Sequential cumulative summation can also slow inference in causal settings, as seen in the Performer (Choromanski et al., 2020). Low-rank methods—e.g., Linformer (Wang et al., 2020)—similarly achieve $\mathcal{O}(n)$ complexity, but their effectiveness depends on the selected rank. Notably, the Performer performs worse on autoregressive generation than for masked language modeling (MLM), and the Linformer is not applicable to decoder-based modeling in general. The *Attention Free Transformer* (AFT) (Zhai et al., 2021) replaces dot product self-attention by learned position biases. The biases are added to the keys and values, and the result is multiplied by the query element-wise, which is an advantage for large model sizes. Later, Hyena (Poli et al., 2023) generalizes the work of AFT by combining multiplicative element-wise gating with implicit long convolutions. [1]

However, while these variants are important foundational works, they are generally no longer competitive with more recent architectures. Recent competitive variants, such as ReGLA (Lu et al., 2025b), Hedgehog (Zhang et al., 2024), and RoFly (Ro et al., 2025), improve efficiency while also enhancing expressivity. Log-linear attention (Guo et al., 2025) extends linear attention through a logarithmically growing set of hidden states, providing a trade-off between efficiency and expressiveness.

3.2 Sparse Attention

Sparse attention mechanisms focus computation on a subset of the sequence using fixed or learnable patterns. Sparse Transformers (Child et al., 2019) pioneered sparse factorizations of the attention matrix, reducing complexity to $\mathcal{O}(n\sqrt{n})$. Local (sliding window) attention restricts computation to a window around each token and is often paired with global attention, as in Longformer (Beltagy et al., 2020), to regain expressivity by allowing selected tokens to attend globally. Other variants, such as strided or random patterns, are often combined (e.g., Zaheer et al., 2020). While some sparse patterns achieve $\mathcal{O}(n)$ time and memory complexity, they may underperform on tasks requiring fine-grained global dependencies and often require task-

[1] Hyena, however, includes a recurrent operator with a time complexity of $\mathcal{O}(NL \log_2 L)$ (Poli et al., 2023)

specific tuning. Learnable and adaptive sparsity patterns (e.g., Correia et al., 2019) can address these limitations. More recent examples of sparse attention are MoBA (Lu et al., 2025a) or NSA (Yuan et al., 2025): NSA uses hardware-aligned sparse attention kernels and a learned gating mechanism to combine sliding attention, compressed attention for coarse-grained patterns, and selected attention for important token blocks (Yuan et al., 2025). MoBA is inspired by *Mixture-of-Experts* (MoE) systems, divides the context into blocks, and uses a dynamic gating mechanism to selectively route query tokens to the KV blocks most relevant to them (Lu et al., 2025a).

4 Recurrent State Models

Recurrent Neural Networks (RNNs) process sequences by maintaining a fixed-size state updated at each time step, allowing them to model temporal dependencies (Yu et al., 2019). *Long Short-Term Memory* (LSTM) networks (Hochreiter and Schmidhuber, 1997) mitigate the vanishing gradient problem through a complex gating mechanism, while *Gated Recurrent Units* (GRU) (Cho et al., 2014) offer a simpler alternative with similar performance and lower computational cost. RNN variants offer linear autoregressive generation, but suffer from (1) varying degrees of vanishing/exploding gradients, (2) limited training parallelism, and (3) lack of expressivity due to a representation state not scaling with context length (Yu et al., 2019).

While the following models are not RNNs per se, they recurrently update a state in the form of a matrix (in contrast to single vectors in classic RNNs) where the influence of earlier inputs decays over time, thus a formulation as such a model is possible. These successors partly mitigated the limitations named above, starting with Katharopoulos et al. (2020), which introduce linear attention.

4.1 Gated Attention Models

Instead of attending to a full key-value matrix of previous outputs, gates can recombine a fixed context state that tokens attend to with new outputs. The Retentive Network (RetNet) (Sun et al., 2023) builds on linear attention, improving the performance by applying exponential decay to the hidden state before the RNN update. It introduces the retention mechanism for sequence modeling, whose parallel representation (for efficient training) resembles self-attention but replaces the softmax operation by a Hadamard product, data-independent exponential decay, and GroupNorm. This enables a recurrent representation and consequently low-cost $\mathcal{O}(1)$ inference per token. Moreover, a chunkwise recurrent representation combines parallel encoding within chunks with recurrent summarization to efficiently model long sequences with linear complexity.

Successively, Gated Linear Attention (GLA) Yang et al. (2024a) introduces data-dependent gates and the hardware-efficient algorithm FlashLinearAttention. This enabled GLA Transformers to perform competitively with full attention for small-scale language models at the time and came with no significant perplexity degradations when using a model trained on 2000 tokens on sequences longer than 20000 tokens, which demonstrates their effectiveness at length generalization. Since the gating computation per token depends only on the current token and the learnable parameters, it remains parallelizable.

DeltaNet (Schlag et al., 2021; Yang et al., 2024b) is a linear transformer variant that retrieves and updates a value vector associated with each key using an update rule similar to the delta rule. DeltaNet employs a *diagonal plus low-rank* (DPLR) state-update mechanism similar to S4, enabling efficient parallelization across the temporal dimension and significantly improving training efficiency. Gated DeltaNet (Yang et al., 2025b) builds upon this and introduces the gated delta rule, which integrates gating for adaptive memory control. Similarly, Goose (RWKV-7) (Peng et al., 2025), the latest RWKV version, integrates a generalized delta rule, vector-valued gating, in-context learning rates, and a relaxed value replacement rule. The initial RWKV-4 (Peng et al., 2023) uses token shift and builds on AFT (Zhai et al., 2021) by using channel-wise time decay vectors instead of global interaction weights. See Li et al. (2025b) for a detailed overview.

4.2 Lightning Attention (LA2)

Lightning Attention-2 (Qin et al., 2024b) divides attention into intra-block (standard attention) and inter-block (linear attention via kernel tricks) computations. This divide-and-conquer strategy addresses the slow training of causal linear attention—caused by sequential cumulative sums—by combining efficient intra-block processing with fast, kernel-based inter-block calculations. LA2 keeps a fixed-size hidden state and is considered a data-

independent decay variant. While the forward pass of LA2 resembles RetNet's chunk-wise retention algorithm (Sun et al., 2023), LA2 additionally includes the backward pass, incorporates IO-aware optimizations from FlashAttention, and enhances GPU performance through tiling. Both forward and backward passes have time complexity $\mathcal{O}(nd^2)$ (Qin et al., 2024c). Although TransNormerLLM (Qin et al., 2024a) is tailored for Lightning Attention, LA2 itself is model-agnostic. MiniMax-01 (MiniMax et al., 2025) uses LA2 and reports that, under a fixed compute budget, it enables more parameters and tokens, achieving lower loss than standard softmax attention.

Other models that belong in this section, namely the xLSTM (Beck et al., 2024) and HGRN families (Qin et al., 2023), are described in Appendix A.2.

5 State-Space-based Models

State Space Models (SSMs), originally from control theory for modeling dynamic systems via state variables (Kalman, 1960), have emerged as promising sub-quadratic alternatives to transformers. A key aspect is their dual perspective: a recurrent formulation enables $\mathcal{O}(n)$ inference, while a convolutional view allows for $\mathcal{O}(n \log(n))$ training via efficient FFT-based convolutions. Note that the models are listed in this separate subsection, not as a concept distinct from recurrent state models, but for their importance in current research.

5.1 Structured SSMs

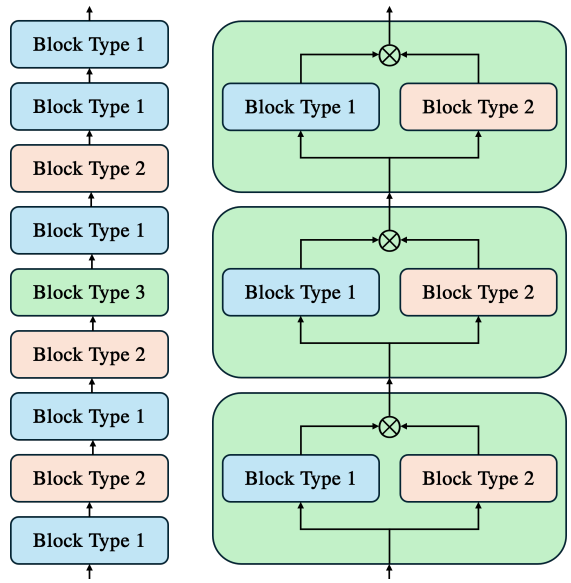
Structured SSMs impose a specific mathematical structure—such as low-rank or diagonal-plus-low-rank forms—on state transition and input matrices, enabling efficient and expressive modeling of long-range dependencies. S4 (Gu et al., 2022) introduces the use of a *Highly Predictive Polynomial Projection Operator* (HiPPO) matrix for initializing the state transition. This approach enables the construction of global convolution kernels that can efficiently encode long-term dependencies. At the time of release, S4 matched the performance of transformers (Gu et al., 2022). S5 (Smith et al., 2023) simplifies and extends S4 by replacing its diagonal block structure with dense matrices. Additionally, S5 leverages an efficient parallel scan, removing the need for S4's convolutional and frequency domain computations and streamlining kernel computation.

5.2 Selective SSMs

Mamba (Gu and Dao, 2023) advances SSMs by replacing fixed transition matrices with input-dependent functions, increasing flexibility and expressivity. Its core is the Mamba block, which combines the ideas of H3 (Fu et al., 2023) and gated MLP blocks by adding a convolution and an SSM to the main branch of the gated MLP. Efficient implementation is achieved via kernel fusion, parallel scan, and recomputation.

Mamba2 (Dao and Gu, 2024) further unifies structured SSMs with attention mechanisms, enabling the application of transformer-style optimizations. It uses modified Mamba blocks for tensor parallelism and introduces the *State Space Dual* (SSD) layer as the inner SSM, which, in its recurrent form, is a selective SSM with single-input single-output structure. This design slightly reduces expressivity but significantly improves training efficiency on modern accelerators.

6 Hybrids



(a) Striped Hybrid (b) Fusion Hybrid

Figure 2: Different types of hybrids. **(a)**: block types using different primitives are connected in series. **(b)**: block types are connected in parallel.

Hybrid architectures combine different primitives—such as SSMs, attention, and RNNs—to leverage their strengths while mitigating the limitations of individual approaches (see Sections 8.1 and 8.2). Such hybrids are usually of a striped (i.e., alternating primitives in series) or

a fusion nature (i.e., primitives are calculated in parallel, combining their outputs). See Figure 2 for reference.

6.1 $\mathcal{O}(n^2)$ Hybrids

SSM + Quadratic Attention Several works demonstrate that combining SSM and attention layers often outperforms using either one alone: For instance, Dao and Gu (2024) show that integrating SSD layers, attention, and MLPs can surpass pure Transformers or Mamba-2. Jamba (Lenz et al., 2025) merges transformer, Mamba, and MoE layers into a striped hybrid, achieving performance comparable to Llama-2 70B or Mixtral, but has 2–7x longer context windows, 3x higher throughput, fewer total parameters (52B, 12B active), and reduced KV cache memory (32GB vs. 4GB for 256K tokens). Other examples are MambaFormer (Park et al., 2024), another striped hybrid, and Hymba (Dong et al., 2025), combining fusion and striped hybrid patterns.

Lightning Attention + Quadratic Attention MiniMax et al. (2025)'s MiniMax-01 series combines lightning attention with an MoE approach. To address LA2's limited retrieval, their Hybrid-lightning architecture replaces LA2 with $\mathcal{O}(n^2)$ attention every eight layers, resulting in a striped hybrid. MiniMax-Text-01 was competitive with SOTA models like GPT-4o or Claude-3.5-Sonnet at the time of release, supporting context windows up to 1M tokens during training and 4M during inference at reasonable cost. However, it still struggles to follow multilevel instructions due to sparse training data.

Gated Attention + Quadratic Attention Kimi Linear (Team et al., 2025) uses Kimi Delta Attention (KDA), a linear attention mechanism that refines the gated delta rule with fine-grained gating. The proposed architecture has three 3 KDA layers for every global attention (MLA) layer.

6.2 $\mathcal{O}(n^{2-\epsilon})$ Hybrids

De et al. (2024) propose the *Real-Gated Linear Recurrent Unit* (RG-LRU), a gated LRU (Orvieto et al., 2023) variant without complex transformations in the recurrence, as these do not improve language modeling in practice. RG-LRU, a fusion hybrid of local attention and linear recurrence, is used for sequence mixing in a recurrent block, replacing MQA. Griffin, using RG-LRU, achieves higher inference throughput and lower latency on long sequences than MQA Transformers (De et al., 2024). On benchmarks, Griffin-3B outperforms Mamba-3B, and Griffin-7B and 14B are competitive with Llama-2 despite using much less training data. Griffin is also used as the base for RecurrentGemma (Botev et al., 2024).

Another notable sub-quadratic hybrid is Samba (Ren et al., 2025), a striped hybrid combining sliding window attention and Mamba/SSM layers.

6.3 Novel Architecture Design Concepts

Memory System Design Recent models increasingly integrate several memory types (Irie et al., 2025; Nunez et al., 2025). Titans (Behrouz et al., 2024) introduce meta in-context neural long-term memory, storing surprising data at test time, and combine core attention-based short-term, neural long-term, and persistent task memory modules.

B'MOJO (Zancato et al., 2024) generalizes transformers and SSMs by blending permanent, short-term, fading, and long-term memories, with a sliding attention mechanism to aggregate information. Both models show good results versus transformers on several benchmarks (see Table 2).

Tailored Architecture Search Thomas et al. (2024)'s STAR framework unifies popular sequence model architectures under the theory of *Linear Input-Varying systems* (LIVs), creating a larger and more structured search space for model design. Given target metrics such as cache size, perplexity, or device latency, STAR uses gradient-free evolutionary algorithms to automatically search the LIV space and generate architectures optimized for several objectives, outperforming highly-tuned transformer and hybrid models on various quality and efficiency frontiers. A recent model realized through STAR (with slight modifications) is the edge model LFM2 (LiquidAI, 2025).

Another example of architectural search is Post-NAS (Gu et al., 2025), which enables an exploration of attention block designs building on pre-trained transformer models.

7 Complexity and Benchmark Analysis

Moving beyond the qualitative analysis in the previous sections, this section focuses on quantitative results and directly compares model architectures in terms of complexity and benchmark performance.

Method	Training			Inference	
	Time	Space	Parallel	Time	Space
FFT-Convolution	$\mathcal{O}(Bnd\log(dn))$	$\mathcal{O}(Bnd)$	Yes	$\mathcal{O}(nd\log(nd))$	$\mathcal{O}(nd)$
RNN	$\mathcal{O}(Bnd^2)$	$\mathcal{O}(Bnd)$	No	$\mathcal{O}(d^2)^2$	$\mathcal{O}(nd)$
Vanilla Transformer	$\mathcal{O}(B(n^2d+nd^2))$	$\mathcal{O}(B(n^2+nd))$	Yes	$\mathcal{O}(n^2d+d^2n)$	$\mathcal{O}(n^2+nd)$
LSH (Reformer)	$\mathcal{O}(Bd^2n\log n)$	$\mathcal{O}(Bn\log n+Bnd)$	Yes	$\mathcal{O}(d^2n\log n)$	$\mathcal{O}(n\log n+nd)$
FAVOR+ (Performer)	$\mathcal{O}(Bnd^2\log d)$	$\mathcal{O}(Bnd\log d+Bd^2\log d)$	Yes	$\mathcal{O}(nd^2\log d)$	$\mathcal{O}(nd\log d+d^2\log d)$
Linear Transformer	$\mathcal{O}(Bnd^2)$	$\mathcal{O}(B(nd+d^2))$	Yes	$\mathcal{O}(nd^2)$	$\mathcal{O}(nd+d^2)$
Lightning Attention	$\mathcal{O}(Bnd^2)$	$\mathcal{O}(B(nd+d^2))$	Yes	$\mathcal{O}(nd^2)$	$\mathcal{O}(nd+d^2)$
Channel-wise AFT (RWKV-4)	$\mathcal{O}(Bnd^2)$	$\mathcal{O}(Bnd)$	Yes	$\mathcal{O}(nd)$	$\mathcal{O}(d)$
Hyena-3	$\mathcal{O}(Bnd\log(dn))$	$\mathcal{O}(Bnd)$	Yes	$\mathcal{O}(nd\log(n+d))$	$\mathcal{O}(nd)$
S4	$\mathcal{O}(Bnd\log(dn))$	$\mathcal{O}(Bnd)$	Yes	$\mathcal{O}(d^2)$	$\mathcal{O}(nd)$
Mamba[3]	$\mathcal{O}(B(nd^2+nd\log(nd)))$	$\mathcal{O}(Bnd)$	Yes	$\mathcal{O}(nd^2+nd\log(nd))$	$\mathcal{O}(nd)$

Table 1: Overview on time & space complexities for training on a single batch and inference of a single token of different sequence-modeling mechanisms. n: sequence length; d: hidden dimension; B: batch size

7.1 Complexity Comparison

We compare the complexities of selected sequence-modeling mechanisms in Table 1. It is important to note that these complexities are sometimes dominated by feed-forward neural networks in the full model, e.g., in S4, which have a time complexity of $\mathcal{O}(nd^2)$. Except for RWKV, which can process a single query at a time at inference, models are lower bounded on memory complexity by storing the sequence in its entirety. Many of these algorithms rely on projections, thus requiring at least $\mathcal{O}(nd^2)$ operations, often serving as an upper bound for time complexity. Another major influence on time complexity is the use of FFT convolutions, as used in SSM-based models for training, which requires $\mathcal{O}(nd\log(dn))$ computational steps, binding the algorithm to log-linear time.

7.2 Benchmark Performance

In Table 2, we provide a performance comparison of previously mentioned sub-quadratic models with recent high-performing quadratic attention models. We chose a frequently used configuration variety: two table sections comparing models with parameter sizes of 0.7-1.5B and 14-70B (total parameter count for MoE models) across eight prominent benchmarks. For the model and benchmark sources, see Appendix A.3. In the 0.7-1.5B range, several edge models compete for the top scores. In particular, Samba and RWKV7-World3 significantly outperform the full attention Llama 3.2 and Qwen2.5 in several instances. In the midrange (14-70B), no pure sub-quadratic models are present anymore; merely the hybrids Griffin and Jamba remain, with only the latter realistically competing with Qwen2.5 and Llama3.1. In the evaluation of frontier (100B+) models, we refer to the LMsys Chatbot Arena (Chiang et al., 2024) instead of a custom-made table. Across all benchmarks (accessed on 2025-10-02), only MiniMax-Text-01 (MiniMax et al., 2025) appears in the top-30 ranking once[4], specifically in the *WebDev* Arena leaderboard. In the top 10, no model is known to be built on alternative architectures.

Note that the benchmark comparison should only be taken as a rough overview. Public leaderboards and benchmarks are highly volatile, and scores reflect only the status at the given timestamp. Moreover, unstandardized benchmarking makes comparing architectures throughout the literature difficult (see Appendix A.3 and Limitations for details).

8 Fundamental Architectural Limitations

Both quadratic attention and sub-quadratic architectures face fundamental limitations that cannot be overcome by scaling parameters or training. In this section, we discuss these inherent restrictions. General limitations of language models (e.g., Wheeler and Jeunen, 2025) are beyond this survey's scope.

8.1 Limitations of Attention

General Theoretical Expressivity The standard transformer forward pass belongs to the log-time uniform TC^0 circuit complexity class (Merrill and Sabharwal, 2023). This fundamentally limits its ability to simulate finite automata or solve

[2] Assuming the sequence has been processed already, only necessary once

[3] We consider an entire Mamba layer here, including projections

[4] There have been new developments between initial and camera-ready submission, some Hybrids actually reached the top 20 of the *Text*, and top 5 of the *WebDev* leaderboard, see Section 9.2.

Model	Benchmark Selection							
0.7-1.5B	Size	**MMLU**	**LMB**	**ARC-E**	**ARC-C**	**Wino.**	**Hella.**	**PIQA**
Titans-MAG	760M	-	41.0	68.2	36.2	52.9	48.9	70.3
Griffin	1B	29.5	-	67.0	36.9	65.2	67.2	**77.4**
Llama3.2*	1B	32.1	63.0	-	-	60.7	63.7	-
GLA	1.3B	-	46.9	57.2	26.6	53.9	49.8	71.8
HGRN2	1.3B	-	49.4	58.1	28.1	52.3	51.8	71.4
Mamba2	1.3B	-	65.7	61.0	33.3	60.9	59.9	73.2
xLSTM[1:0]	1.3B	-	57.8	64.3	32.6	60.6	60.9	74.6
BMoJo-Fading	1.4B	-	45.4	52.3	26.6	53.3	46.0	70.0
RWKV7-World3	1.5B	43.3	**69.5**	78.1	44.5	68.2	**70.8**	77.1
Qwen2.5*	1.5B	**60.9**	63.0	75.5	**54.7**	65.0	67.9	75.8
Samba	1.7B	48.0	-	**79.3**	48.2	**72.9**	49.7	77.1
14-70B	Size	**MMLU**	**BBH**	**GSM8K**	**ARC-C**	**Wino.**	**Hella.**	**HumanEval**
Griffin	14B	49.5	-	-	50.8	74.1	81.4	-
Qwen*	14B	79.7	78.2	90.2	67.3	81.0	84.3	56.7
Jamba	52B	67.40	45.40	59.9	64.40	82.5	87.1	29.30
Mixtral*	56B	70.6	-	60.4	59.7	77.2	84.4	40.2
Llama3.1*	70B	79.5	81.0	95.1	68.8	**85.3**	**88.0**	48.2
Qwen2.5*	72B	**86.1**	**86.3**	**95.8**	**72.4**	83.9	87.6	**59.1**

Table 2: Performance comparison of recent pure quadratic attention LMs (highlighted with *) and subquadratic models of similar size. Best results for each parameter category are marked in **bold**, second-best results are underlined. Model names are in bold or underlined when they scored first or second at least once. Results are accuracy-based and rounded to one decimal point. For sources, see Section A.3

graph connectivity—necessary for state tracking and multi-step reasoning (Merrill and Sabharwal, 2024). In practice, such tasks are tractable for short contexts (e.g., by using transformers of depth $\mathcal{O}(\log C)$ for context length C), but remain infeasible for unbounded inputs under standard complexity assumptions. To scale up these capabilities, the model dimension must grow with the task complexity, as is also highlighted in related work (Hahn, 2020; Sanford et al., 2023).

Allowing intermediate steps, i.e., *Chain of Thought* (CoT) (Wei et al., 2022), increases transformer expressivity w.r.t. the number of steps. Li et al. (2024) show that with T CoT steps, constant-depth transformers with $\mathcal{O}(\log n)$ embeddings can solve any problem solvable by boolean circuits of size T. Additionally, Qiu et al. (2025) prove that prompting is Turing-complete: for any computable function, a finite-size transformer can compute it with an appropriate prompt. However, these enhancements also introduce new drawbacks, as shown by Bavandpour et al. (2025); Peng et al. (2024); Saparov et al. (2025).

Length Generalization Transformers struggle to extrapolate, i.e., to generalize from shorter training context sizes to longer test sequences. In addition to being limited by memory constraints, the transformer architecture has fundamental length-generalization limits caused by positional encodings (Kazemnejad et al., 2023). While transformers without position encodings (NoPE) seem to be an alternative and work for longer sequences than explicit encodings, they still impose a context length limit (Wang et al., 2024a).

Building upon Huang et al. (2025)'s framework to analyze length generalization, Veitsman et al. (2025) show that, if pretraining is done right, certain capabilities w.r.t. length generalization of transformers can be improved, but fundamental limitations persist. For models like SSMs and B'MOJO, the length generalization is instead limited by the capacity of the recurrent state.

For Huang et al. (2025)' s framework or more details on limitations of attention, see Appendix A.4.

8.2 Limitations of Sub-Quadratic Alternatives

Sub-quadratic architectures share some limitations with quadratic attention—e.g., SSMs are also in the complexity class TC^0 (Merrill et al., 2024). Furthermore, these models introduce additional new challenges due to the inherent difficulty of compressing sequence context into a reduced state.

This finite state capacity has strong implications for "lookup table" tasks (e.g., MQAR (Arora et al., 2024a), hop_k (Sanford et al., 2024)), where such information is part of the input, as SSMs cannot recall an arbitrary amount of information previously

seen (Arora et al., 2024b; De et al., 2024; Jelassi et al., 2024), even though recent work (Grazzi et al., 2024) shows that some improvements can be made, as seen in Mamba (Gu and Dao, 2023).

A similar problem occurs in linear RNNs, which are highly sensitive to the context order, making prompt engineering critical—selection and recall become harder as input order varies (Sutskever et al., 2014; Arora et al., 2024c). RNNs require $\Omega(N)$ space for reliable recall (Arora et al., 2024b), and constant-memory models cannot perform associative recall or solve tasks like q-sparse averaging or copying, unlike shallow transformers (Sanford et al., 2024; Jelassi et al., 2024; Wen et al., 2025).

Backurs and Indyk (2018) prove that under SETH (which implies $P \neq NP$), edit distance cannot be computed in subquadratic time, setting a fundamental limit on sequence comparison efficiency for any such architecture. Under the same assumption, Alman and Yu (2025) show that document similarity tasks inherently require quadratic time.

8.3 Implications

The limitations applying to alternative architectures mostly subsume the limitations of transformers. This implies that while sub-quadratic alternatives significantly enhance efficiency and lower computational costs, they do not fundamentally surpass transformers in theoretical expressivity.

9 Discussion

In this section, we synthesize insights from our review to discuss whether sub-quadratic and hybrid alternatives start claiming meaningful territory.

9.1 Current Landscape

Despite the reviewed advances in alternative architectures, at the time of writing, most frontier general-purpose models strongly rely on full attention mechanisms. No model scoring in the top 10 on LLMSys (Chiang et al., 2024) is known to be sub-quadratic or a hybrid, showing that the "Transformer++" remains the default choice when compute is not a limiting factor. We have also seen that full attention is free from many limitations that apply to alternative architectures (Section 8.2), adding to the extent of their superiority.

However, the picture changes for edge models, where compute, memory, and latency are tightly bound, and alternative architectures have gained substantial traction. Especially hybrids, such as Samba (Ren et al., 2025) or RWKV7 (Peng et al., 2025), offer favorable inference properties. They can meet resource constraints by offloading local or intermediate computations to more efficient modules, while maintaining reasonable generalization and global context modeling via attention. For the edge, we also increasingly see differentiated memory modeling with newer models, like Titans (Behrouz et al., 2024) and B'MOJO (Zancato et al., 2024), segmenting memory into short-term, long-term, and permanent storage, assigning specialized mechanisms to each.

In the mid-size regime, hybrids like Jamba (Lenz et al., 2025) show promise, though they remain a minority and do not outperform well-tuned transformers. Their advantages are domain-specific, tied to scenarios where efficiency provides tangible gains. In general, the maturity of transformer infrastructure also makes switching to other architectures costly due to ecosystem inertia (Brem and Nylund, 2024). However, work that enables the conversion of pretrained transformers to alternative architectures without retraining, such as RWKV, starts lowering these barriers.

Regarding the types of hybrids we see, striped and fusion, there is no clear tendency in current research to use one over the other, since this choice highly depends on what primitives are combined. Using full attention in a fusion hybrid comes with no gains in efficiency, while combinations of purely subquadratic primitives can benefit from fusion to balance out their different disadvantages compared to full attention.

Together, these trends signal a shift toward architectural diversity. While transformers remain dominant, alternatives are finding footholds in specific use cases and operational niches.

9.2 Outlook

At the frontier, full attention is likely to remain central for the foreseeable future. Still, even these models may begin incorporating hybrid elements, especially for memory management or task-specific routing. In this sense, we also anticipate model routing and *Mixture of Architectures* (MoA) paradigms to become more relevant. The shift is not toward replacement, but toward building flexible systems from a growing set of specialized primitives, an idea that has already been surfaced by Yu et al. (2025), Varangot-Reille et al. (2025) and Fu et al. (2024), and continues to gain traction.

Several new open source models that interleave sparse and global attention (so $\mathcal{O}(n^2)$ hybrids) were released between the initial and camera-ready submission of our work. As of 2026-05-13, they are in the very top of both the *WebDev* (up to top 5) and *Text* (up to top 20) Chatbot Arena leaderboards, reinforcing our expectations: Some examples are Mimo-v2.5-pro (Xiaomi MiMo Team, 2026) and Gemma-4-31b (Google, 2026) that both use sliding window attention combined with global attention, GLM-5.1 (Zeng et al., 2026), which switches from global to sparse attention after mid-training, and Deepseek-V4 (DeepSeek-AI, 2026), which uses a striped hybrid architecture comprising compressed forms of sparse and global attention.

10 Conclusion

Through our review of recent subquadratic architectures, we have highlighted the most promising alternatives to full attention for sequence modeling in NLP. Our analysis shows that sub-quadratic elements introduce valuable tradeoffs in efficiency and latency, particularly in edge and mid-sized deployments, but do remain fundamentally constrained in generality compared to transformers. We do not expect pure subquadratic architectures on the frontier for the foreseeable future, but see Hybrids catching up fast.

Limitations

As a focused and concise survey, our work comes with several limitations. We restrict our analysis to language models, and therefore, our findings may not generalize to other modalities such as vision, audio, or multimodal systems. Additionally, the performance comparison presented in Table 2 is limited in its language coverage, as it focuses primarily on English. There is also a slight variation in training data and procedure across the benchmark results of the models we report on, which is explained in Section A.3. Finally, while aimed at identifying and synthesizing all relevant literature, researchers with a different focus could consider some missing works more significant.

Acknowledgments

All analysis, research, and ideas are either our own or cited. This work used LLM-based tools for language edits and clarity improvements. This research has been funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) through grant 01IS23069 Software Campus 3.0 (Technical University of Munich) as part of the Software Campus project "Know ELViS".

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Josh Alman and Hantao Yu. 2025. Fundamental limitations on subquadratic alternatives to transformers. In *The Thirteenth International Conference on Learning Representations*.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. 2024a. Zoology: Measuring and improving recall in efficient language models. In *The Twelfth International Conference on Learning Representations*.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré. 2024b. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*.

Simran Arora, Aman Timalsina, Aaryan Singhal, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré. 2024c. Just read twice: closing the recall gap for recurrent language models. In *Workshop on Efficient Systems for Foundation Models II @ ICML2024*.

Arturs Backurs and Piotr Indyk. 2018. Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). *SIAM Journal on Computing*, 47(3):1087–1097.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alireza Amiri Bavandpour, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *Forty-second International Conference on Machine Learning*.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2024. xLSTM: Extended long

short-term memory. In *Advances in Neural Information Processing Systems*, volume 37, pages 107547–107603. Curran Associates, Inc.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. 2024. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, 34(05):7432–7439.

Aleksandar Botev, Soham De, Samuel L. Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Sertan Girgin, Olivier Bachem, Alek Andreev, Kathleen Kenealy, Thomas Mesnard, Cassidy Hardin, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Armand Joulin, Noah Fiedel, Evan Senter, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, David Budden, Arnaud Doucet, Sharad Vikram, Adam Paszke, Trevor Gale, Sebastian Borgeaud, Charlie Chen, Andy Brock, Antonia Paterson, Jenny Brennan, Meg Risdal, Raj Gundluru, Nesh Devanathan, Paul Mooney, Nilay Chauhan, Phil Culliton, Luiz Gustavo Martins, Elisa Bandy, David Huntsperger, Glenn Cameron, Arthur Zucker, Tris Warkentin, Ludovic Peran, Minh Giang, Zoubin Ghahramani, Clément Farabet, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, Yee Whye Teh, and Nando de Frietas. 2024. RecurrentGemma: Moving past transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*.

Alexander Brem and Petra Nylund. 2024. The inertia of dominant designs in technological innovation: An ecosystem view of standardization. *IEEE Transactions on Engineering Management*, 71:2640–2648.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse transformers. *Proceedings of the 2019 Conference on Empirical Meth-*

ods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2174–2184.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359. Curran Associates, Inc.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10041–10071. PMLR.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. 2024. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui

Gu, Zilin Li, and Ziwei Xie. 2024. DeepSeek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. 2025. Hymba: A hybrid-head architecture for small language models. In *The Thirteenth International Conference on Learning Representations*.

Qihang Fan, Huaibo Huang, and Ran He. 2025. Breaking the low-rank dilemma of linear attention. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25271–25280.

Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. 2023. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*.

Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, et al. 2024. Moa: Mixture of sparse attention for automatic large language model compression. *arXiv preprint arXiv:2406.14909*.

Google. 2026. Gemma 4. `https://ai.google.dev/gemma/docs/core`. Version 4.0.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Fe-

ichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Riccardo Grazzi, Julien Niklas Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. 2024. Is mamba capable of in-context learning? In *Proceedings of the Third International Conference on Automated Machine Learning*, volume 256 of *Proceedings of Machine Learning Research*, pages 1/1–26. PMLR.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations*.

Yuxian Gu, Qinghao Hu, Shang Yang, Haocheng Xi, Junyu Chen, Song Han, and Han Cai. 2025. Jet-nemotron: Efficient language model with post neural architecture search. *arXiv preprint arXiv:2508.15884*.

Han Guo, Songlin Yang, Tarushii Goel, Eric P Xing, Tri Dao, and Yoon Kim. 2025. Log-linear attention. *arXiv preprint arXiv:2506.04761*.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171. MIT Press.

Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. 2024. Bridging the divide: Reconsidering softmax and linear attention. In *Advances in Neural Information Processing Systems*, volume 37, pages 79221–79245. Curran Associates, Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *The Ninth International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. 2025. A formal framework for understanding length generalization in transformers. In *The Thirteenth International Conference on Learning Representations*.

Yunpeng Huang, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, Shupeng Li, and Penghao Zhao. 2024. Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*.

Kazuki Irie, Morris Yau, and Samuel J. Gershman. 2025. Blending complementary memory systems in hybrid quadratic-linear transformers. *arXiv preprint arXiv:2506.00744*.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. 2024. Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21502–21521. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. 2023. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 24892–24928. Curran Associates, Inc.

Jongseon Kim, Hyungjoon Kim, HyunGi Kim, Dongjun Lee, and Sungroh Yoon. 2025. A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges. *Artificial Intelligence Review*, 58:216.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. 2025. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*.

Haoyang Li, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. 2025a. A survey on large language model acceleration based on KV cache management.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*.

Zhiyuan Li, Tingyu Xia, Yi Chang, and Yuan Wu. 2025b. A survey of RWKV. *Neurocomputing*, 649:130711.

Zhixuan Lin, Evgenii Nikishin, Xu He, and Aaron Courville. 2025. Forgetting transformer: Softmax

attention with a forget gate. In *International Conference on Learning Representations*, volume 2025, pages 69704–69738.

LiquidAI. 2025. Introducing LFM2: The fastest on-device foundation models on the market | liquid AI. Accessed: 2025-07-24.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, Zhiqi Huang, Huan Yuan, Suting Xu, Xinran Xu, Guokun Lai, Yanru Chen, Huabin Zheng, Junjie Yan, Jianlin Su, Yuxin Wu, Neo Y. Zhang, Zhilin Yang, Xinyu Zhou, Mingxing Zhang, and Jiezhong Qiu. 2025a. Moba: Mixture of block attention for long-context llms.

Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Boxing Chen, and Philippe Langlais. 2025b. ReGLA: Refining gated linear attention. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2884–2898, Albuquerque, New Mexico. Association for Computational Linguistics.

Shi Luohe, Hongyi Zhang, Yao Yao, Zuchao Li, and hai zhao. 2024. Keep the cost down: A review on methods to optimize LLM's KV-cache consumption. In *First Conference on Language Modeling*.

William Merrill, Jackson Petty, and Ashish Sabharwal. 2024. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35492–35506. PMLR.

William Merrill and Ashish Sabharwal. 2023. A logic for expressing log-precision transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 52453–52463. Curran Associates, Inc.

William Merrill and Ashish Sabharwal. 2024. A little depth goes a long way: The expressive power of log-depth transformers. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 2025. Towards efficient generative large language model serving: A survey from algorithms to systems. *ACM Computing Surveys*. Association for Computing Machinery.

Xupeng Miao, Shenhan Zhu, Fangcheng Fu, Ziyu Guo, Zhi Yang, Yaofeng Tu, Zhihao Jia, and Bin Cui. 2024. X-former elucidator: Reviving efficient attention for long context language modeling. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8179–8187. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. 2025. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Elvis Nunez, Luca Zancato, Benjamin Bowman, Aditya Golatkar, Wei Xia, and Stefano Soatto. 2025. Expansion span: Combining fading memory and retrieval in hybrid state space models. In *Proceedings of the International Conference on Neuro-symbolic Systems*, volume 288 of *Proceedings of Machine Learning Research*, pages 570–596. PMLR.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. 2023. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024. Can mamba learn how to learn? A comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39793–39812. PMLR.

Badri Narayana Patro and Vijay Srinivas Agneeswaran. 2025. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling:

Methods, applications, and challenges. *Engineering Applications of Artificial Intelligence*, 159:111279.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. 2024. On limitations of the transformer architecture. In *First Conference on Language Modeling*.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077. Association for Computational Linguistics.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. 2025. RWKV-7 "goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Re. 2023. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28043–28078. PMLR.

Zhen Qin, Dong Li, Weigao Sun, Weixuan Sun, Xuyang Shen, Xiaodong Han, Yunshen Wei, Baohong Lv, Xiao Luo, Yu Qiao, and Yiran Zhong. 2024a. TransNormerLLM: A faster and better large language model with improved TransNormer. *arXiv preprint arXiv:2307.14995*.

Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. 2024b. Lightning Attention-2: A free lunch for handling unlimited sequence lengths in large language models. *arXiv preprint arXiv:2401.04658*.

Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. 2024c. Various lengths, constant speed: Efficient language modeling with lightning attention. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41517–41535. PMLR.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. 2024d. HGRN2: Gated linear RNNs with state expansion. *arXiv preprint arXiv:2404.07904*.

Zhen Qin, Songlin Yang, and Yiran Zhong. 2023. Hierarchically gated recurrent neural network for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 36, pages 33202–33221. Curran Associates, Inc.

Ruizhong Qiu, Zhe Xu, Wenxuan Bao, and Hanghang Tong. 2025. Ask, and it shall be given: On the turing completeness of prompting. In *The Thirteenth International Conference on Learning Representations*.

Liliang Ren, Yang Liu, Yadong Lu, yelong shen, Chen Liang, and Weizhu Chen. 2025. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *The Thirteenth International Conference on Learning Representations*.

Yeonju Ro, Zhenyu Zhang, Souvik Kundu, Zhangyang Wang, and Aditya Akella. 2025. On-the-fly adaptive distillation of transformer to dual-state linear attention. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106. Association for Computing Machinery.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. 2024. Transformers, parallel computation, and logarithmic depth. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43276–43327. PMLR.

Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. 2023. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 36677–36707. Curran Associates, Inc.

Abulhair Saparov, Srushti Ajay Pawar, Shreyas Pimpalgaonkar, Nitish Joshi, Richard Yuanzhe Pang, Vishakh Padmakumar, Mehran Kazemi, Najoung Kim, and He He. 2025. Transformers struggle to learn to search. In *The Thirteenth International Conference on Learning Representations*.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. 2021. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9355–9366. PMLR.

Johannes Schneider. 2025. What comes after transformers? a selective survey connecting ideas in Deep LearningGPT. In *Agents and Artificial Intelligence: 16th International Conference, ICAART 2024, Rome, Italy, February 24–26, 2024, Revised Selected Papers, Part II*, page 55–82, Berlin, Heidelberg. Springer-Verlag.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems*, volume 37, pages 68658–68685. Curran Associates, Inc.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. 2023. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*.

Shriyank Somvanshi, Md Monzurul Islam, Mahmuda Sultana Mimi, Sazzad Bin Bashar Polock, Gaurab Chhetri, and Subasish Das. 2025. From S4 to mamba: A comprehensive survey on structured state space models. *arXiv preprint arXiv:2503.18970*.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. 2024. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561. MIT Press.

Weigao Sun, Jiaxi Hu, Yucheng Zhou, Jusen Du, Disen Lan, Kexin Wang, Tong Zhu, Xiaoye Qu, Yu Zhang, Xiaoyu Mo, et al. 2025. Speed always wins: A survey on efficient architectures for large language models. *arXiv preprint arXiv:2508.09834*.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhijun Tu, Kai Han, Hailin Hu, and Dacheng Tao. 2024. A survey on transformer compression. *arXiv preprint arXiv:2402.05964*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *The Ninth International Conference on Learning Representations*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6). Association for Computing Machinery.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. 2025. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*.

Armin W. Thomas, Rom Parnichkun, Alexander Amini, Stefano Massaroli, and Michael Poli. 2024. STAR: Synthesis of tailored architectures. *arXiv preprint arXiv:2411.17800*.

Matteo Tiezzi, Michele Casoni, Alessandro Betti, Tommaso Guidi, Marco Gori, and Stefano Melacci. 2024. On the resurgence of recurrent models for long sequences – survey and research opportunities in the transformer era. *arXiv preprint arXiv:2402.08132*.

Matteo Tiezzi, Michele Casoni, Alessandro Betti, Tommaso Guidi, Marco Gori, and Stefano Melacci. 2025. Back to recurrent processing at the crossroad of transformers and state-space models. *Nature Machine Intelligence*, 7(5):678–688. Nature Publishing Group.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Clovis Varangot-Reille, Christophe Bouvard, Antoine Gourru, Mathieu Ciancone, Marion Schaeffer, and François Jacquenet. 2025. Doing more with less–implementing routing strategies in large language model-based systems: An extended survey. *arXiv preprint arXiv:2502.00409*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yana Veitsman, Mayank Jobanputra, Yash Sarrof, Aleksandra Bakalova, Vera Demberg, Ellie Pavlick, and Michael Hahn. 2025. Born a transformer – always a transformer? *arXiv preprint arXiv:2505.21785*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient large language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. 2024a. Length generalization of causal transformers without position encoding. In *Findings of the Association for Computational Linguistics: ACL 2024*,

pages 14024–14040, Bangkok, Thailand. Association for Computational Linguistics.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Xiao Wang, Shiao Wang, Yuhe Ding, Yuehang Li, Wentao Wu, Yao Rong, Weizhe Kong, Ju Huang, Shihao Li, Haoxiang Yang, Ziwen Wang, Bo Jiang, Chenglong Li, Yaowei Wang, Yonghong Tian, and Jin Tang. 2024b. State space model for new-generation network alternative to transformers: A survey. *arXiv preprint arXiv:2404.09516*.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024c. Beyond the limits: A survey of techniques to extend the context length in large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8299–8307. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. 2025. RNNs are not transformers (yet): The key bottleneck on in-context retrieval. In *The Thirteenth International Conference on Learning Representations*.

Schaun Wheeler and Olivier Jeunen. 2025. Procedural memory is not all you need: Bridging cognitive gaps in LLM-based agents. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, page 360–364, New York, NY, USA. Association for Computing Machinery.

Xiaomi MiMo Team. 2026. Mimo-v2.5-pro. https://huggingface.co/collections/XiaomiMiMo/mimo-v25.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025a. Qwen2.5 technical report. *arXiv preprint arXiv:2407.10671*.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2025b. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2024a. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56501–56523. PMLR.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2024b. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, pages 115491–115522. Curran Associates, Inc.

Shibo Yu, Mohammad Goudarzi, and Adel Nadjaran Toosi. 2025. Efficient routing of inference requests across LLM instances in cloud-edge computing. *arXiv preprint arXiv:2507.15553*.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 31(7):1235–1270.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. 2025. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23078–23097, Vienna, Austria. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Luca Zancato, Arjun Seshadri, Yonatan Dukler, Aditya Golatkar, Yantao Shen, Benjamin Bowman, Matthew Trager, Alessandro Achille, and Stefano Soatto. 2024. B'MOJO: Hybrid state space realizations of foundation models with eidetic and fading memory. In *Advances in Neural Information Processing Systems*, volume 37, pages 130433–130462. Curran Associates, Inc.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Shuangfei Zhai, Walter A. Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and J. Susskind. 2021. An attention free transformer. *arXiv preprint arXiv:2105.14103*.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. 2024. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *The Twelfth International Conference on Learning Representations*.

Shen Zhuoran, Zhang Mingyuan, Zhao Haiyu, Yi Shuai, and Li Hongsheng. 2021. Efficient attention: Attention with linear complexities. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3530–3538.

A Appendix

A.1 Survey Methodology

Our survey followed a two-fold methodology: First, to determine which alternative model architectures to include, we began with a set of seed papers drawn from recent articles in the field, namely Wang et al. (2024c), Gu and Dao (2023), Sun et al. (2023), and Tay et al. (2022). From this base, we employed a backward and forward snowballing strategy: we examined the references cited within these seed papers (backward snowballing) as well as subsequent papers that cited them (forward snowballing). This iterative process enabled us to trace the development and recurrence of specific architectural primitives over time and across various research communities. Architectures that consistently reappeared in recent high-impact publications were included in the main body of our review. Architectures with limited recurrence and marginal impact were excluded.

Second, for the chapter discussing the fundamental limitations of quadratic and sub-quadratic architectures, we conducted a systematic literature review. This involved querying several academic databases with the search term

("fundamental limitation") AND ("transformer" OR "attention" OR "subquadratic") AND ("natural language processing" OR "NLP" OR "language model")

to identify relevant theoretical and empirical work. The results, i.e., number of hits for each platform, and the search space (full text or abstract only), are stated in the following:

- ACL: 300 (full text)
- Semantic Scholar: 258 (full text)
- Google Scholar: 4430 (full text)*
- IEEE: 4 (abstract)

We then condensed our findings and reported on the very core of limitations that the other findings build upon. Secondary limitations were moved to Appendix A.4. *For Google Scholar, we used additional filtering to address the high number of hits and relatively low overall relevance. The SLR cutoff was 2025-06-18, but we continued to include relevant individual papers until the paper submission.

A.2 Honorable Mentions

In our work, we have encountered various interesting and previously impactful subquadratic architectures, which, however, we were not able to include in the main body of this paper. This was usually due to a combination of limited space and our findings that these architectures were outperformed by others before they became relevant in the long run. For completeness, this section gives a brief overview of these works.

- **Extended Long Short-Term Memory (xLSTM)** xLSTM (Beck et al., 2024) enhances the LSTM architecture by incorporating state expansion, normalization, and stabilization techniques and also using exponential gating and matrix memory. It stacks two specialized LSTM modules: sLSTM, with scalar memory and update mechanisms for efficient state mixing and tracking, and mLSTM, with matrix memory and a covariance-based update rule for improved memorization and parallelism. The mLSTM's matrix memory supports tasks like Multi-Query Associative Recall. xLSTM has linear time and constant memory complexity, but incurs overhead from complex memory operations, only partially offset by hardware-aware optimizations.
- **HGRN** The Hierarchically Gated Recurrent Neural Network (HGRN) (Qin et al., 2023) consists of stacked layers comprising token-mixing (HGRU) and channel-mixing (GLU) modules. Unlike S4 or RWKV-4, HGRN also uses data-dependent, dynamic decay rates via forget gates, allowing lower layers to focus on short-term and higher layers on long-term dependencies. Learnable lower bounds on forget gates prevent vanishing gradients. To address limited recurrent state size, HGRN2 (Qin et al., 2024d) expands the state non-parametrically, improving scaling and outperforming Mamba on Long Range Arena (Tay et al., 2021), though pretrained transformers like LLaMA (Touvron et al., 2023) still perform better on long-context tasks. HGRN2 has been scaled to 3B parameters.

A.3 Benchmark Score Sourcing

The exact methodology through which benchmark scores are reported on in the literature referenced throughout this survey varies strongly. Table 2 represents our best effort at consolidating these results

without introducing yet another evaluation suite. We chose to use the scores from Yang et al. (2025a) for Qwen2.5 and Llama 3.1, and Peng et al. (2025) for Llama 3.2 and RWKV7, due to their consistent evaluation approach. For all other models, we gathered the results from their original technical papers, ensuring consistency to the best of our knowledge. Nevertheless, some inconsistencies, namely in the number and type of tokens used during training, and differences in the number of shots for some task/model combinations, remain.

The model references are as follows: Titans (Behrouz et al., 2024), Griffin (De et al., 2024), GLA (Yang et al., 2024a), HGRN2 (Qin et al., 2024d), Mamba2 (Dao and Gu, 2024), xLSTM (Beck et al., 2024), BMoJo (Zancato et al., 2024), RWKV7 (Peng et al., 2025), Samba (Ren et al., 2025), Jamba (Lenz et al., 2025), Qwen2.5 (Yang et al., 2025a), Llama3.1 (Grattafiori et al., 2024), Mixtral (Jiang et al., 2024).

The references for the benchmarks are MMLU (Hendrycks et al., 2021), Lambada (Paperno et al., 2016), PIQA (Bisk et al., 2020), BBH (Suzgun et al., 2023), ARC-E and ARC-C (Clark et al., 2018), Winogrande (Sakaguchi et al., 2021), HellaSwag (Zellers et al., 2019), GSM8k (Cobbe et al., 2021), and HumanEval (Chen et al., 2021), all used scores are accuracy based.

A.4 Additional Limitations of Attention

In this section, we list interesting additional limitations that were not included in the main body of the paper.

- Hahn (2020) prove that pure attention Transformers cannot handle bracket matching, iterated negation, or non-counter-free regular languages on long inputs, nor emulate stacks or arbitrary finite-state automata (unless layers or heads scale with input length).

- Sanford et al. (2023) show that single-layer, multi-head Transformers require polynomially more heads or dimensions to solve certain triple detection tasks, and likely struggle with higher-order tasks like Match3 (Sanford et al., 2023) without hints or augmentation. However, most real-world sequence problems decompose into pairwise relationships, aligning well with transformer capabilities.

- Huang et al. (2025) propose a theoretical framework to investigate length generalization in causal transformers that use learnable absolute positional encodings. By introducing constraints on how positional information can be utilized, their framework allows them to derive results for multilayer models. They formally prove problems with poor length generalization, such as copying sequences containing repeated strings. Although it remains an open question whether the expressivity of transformers goes beyond the complexity class TC^0, their findings suggest a potential distinction between problems solvable within TC^0 and those for which length generalization is feasible with absolute positional encodings.

- Bavandpour et al. (2025) investigate systematic lower bounds on the number of CoT steps required for various algorithmic problems within a hard-attention setting. Their analysis demonstrates that the required CoT length must necessarily scale with the input length, thereby constraining the ability of self-attention models to solve these tasks efficiently with small inference-time compute.

- Peng et al. (2024) prove that a single transformer layer is not able to do function composition if the domain size of the functions is larger than the dimension parameters of the transformer. Moreover, they show that if we leverage CoT, the model needs to generate a $\Omega(\sqrt{n})$ long prompt to solve iterated function composition, with n being the number of tokens in the prompt. They assume that multi-layer transformers struggle as well.

- Saparov et al. (2025) argue that transformers with standard training will not have robust searching and planning abilities, no matter their number of parameters. For small graphs, a model with effectively limitless and idealized training data can learn to search. Nevertheless, according to them, even if a model can use search in-context (i.e., CoT), it still struggles with search on larger graphs.

- Han et al. (2024) show that linear attention is not injective, often assigning identical attention weights to different queries and causing semantic confusion. They also demonstrate that linear attention struggles with effective local modeling, a strength of softmax attention.

Moreover, the low-rank nature of linear attention's feature map can further hinder modeling of complex spatial or local information (Fan et al., 2025).

A.5 Modern $\mathcal{O}(n^2)$ Attention

The core principle of quadratic attention has not changed much in recent years. However, system-level improvements significantly influence the discussion and use of attention today. Although these methods are not the focus of our work since they do not change the $\mathcal{O}(n^2)$ bottleneck, many attention variants deliver substantial practical speedups with no reduction in quality compared to standard attention. We briefly cover the most common techniques to establish a fair context for the later discussion of alternative architectures.

KV Cache Optimizations During inference, attention's keys and values are often cached to avoid redundant computation, making efficient *key-value* (KV) cache management key for reducing memory requirements. *Multi-Query Attention* (MQA) (Shazeer, 2019) and *Grouped-Query Attention* (GQA) (Ainslie et al., 2023) share key and value matrices across attention heads, reducing cache size by a constant factor at the cost of reduced expressivity. *Multi-Head Latent Attention* (MLA), introduced by DeepSeek (DeepSeek-AI et al., 2024; DeepSeek-AI et al., 2025), shares a latent matrix among heads, which is projected back individually, achieving similar cache savings with better performance than MQA and GQA. Refer to Li et al. (2025a) and Luohe et al. (2024) for a detailed overview of KV cache techniques.

The *Paged Attention* algorithm (Kwon et al., 2023) enables the storing of attention keys and values in non-contiguous paged memory. More specifically, it improves inference memory efficiency by partitioning the KV cache into fixed-size pages and tracking them via a page table, boosting throughput 2–4× and eliminating padding.

Flash Attention FlashAttention (Dao et al., 2022) and its successors exploit GPU memory hierarchies to make attention both faster and more memory-efficient, reducing memory usage to be linear in sequence length and delivering 2–4× runtime speedups over strong baselines. FlashAttention-2 (Dao, 2024) improved thread work partitioning for further speedup (as proven by GPT-style (Brown et al., 2020) LLM training), while FlashAttention-3 (Shah et al., 2024), specialized for Hopper GPUs, adds asynchrony and low-precision operations for an additional 1.5–2× boost.

Forgetting Attention The *Forgetting Transformer* (FoX) (Lin et al., 2025) modifies standard softmax attention by adding a learned forget gate that controls how strongly past tokens remain available. Instead of treating all previous context equally persistently, it applies a decay factor so that older or less relevant information gradually fades. FoX can improve long context language modeling and length extrapolation, but still computes the full attention matrix, so it remains quadratically scaling with context length. Forgetting Attention is compatible with the FlashAttention algorithm.

Why Low-Resource NLP Needs More Than Cross-Lingual Transfer: Lessons Learned from Luxembourgish

Fred Philippy[1], Siwen Guo[2], Jacques Klein[1], Tegawendé F. Bissyandé[1]

[1]SnT, University of Luxembourg, Luxembourg
[2]Luxembourg Institute of Science and Technology, Luxembourg
Correspondence: fred.philippy@uni.lu

Abstract

Cross-lingual transfer has become a central paradigm for extending natural language processing (NLP) technologies to low-resource languages. By leveraging supervision from high-resource languages, multilingual language models can achieve strong task performance with little or no labeled target-language data. However, it remains unclear to what extent cross-lingual transfer can substitute for language-specific efforts. In this paper, we synthesize prior research findings and data collection results on Luxembourgish, which, despite its typological proximity to high-resource languages and its presence in a multilingual context, remains insufficiently represented in modern NLP technologies. Across findings, we observe a fundamental interdependence between cross-lingual transfer and language-specific efforts. Cross-lingual transfer can substantially improve target-language performance, but its success depends critically on the availability of sufficiently high-quality, task-aligned target-language data. At the same time, such resources, particularly in low-resource settings, are typically too limited in scale to drive strong performance on their own. Instead, such resources reach their full potential only when leveraged within a cross-lingual framework. We therefore argue that cross-lingual transfer and language-specific efforts should not be viewed as competing alternatives. Instead, they function as complementary components of a sustainable low-resource NLP pipeline. Based on these insights, we provide practical guidelines for integrating and balancing cross-lingual transfer with language-specific development in sustainable low-resource NLP pipelines.

1 Introduction

Recent advances in multilingual language models have dramatically improved the feasibility of developing NLP systems for low-resource languages. Cross-lingual transfer, where supervision in a high-resource language is leveraged to enable performance in other languages, has emerged as a particularly attractive paradigm. In zero-shot and few-shot settings, multilingual models can often achieve competitive performance on downstream tasks in languages with little or no labeled data, reducing the immediate need for costly language-specific annotation (Lin et al., 2022).

For low-resource communities, these developments are transformative, as cross-lingual methods make it possible to bootstrap systems for classification, inference, or retrieval using predominantly English or other high-resource supervision. This development suggests a compelling narrative: perhaps language-specific resources are no longer strictly necessary if transfer is sufficiently strong.

However, this perspective requires nuance. Cross-lingual transfer comes with several well-known limitations. Its effectiveness can vary substantially across language pairs, and even within the same pair it may behave asymmetrically depending on the transfer direction (Malkin et al., 2022). More broadly, transfer performance is shaped by multiple interacting factors, often resulting in lower-than-expected gains (Philippy et al., 2023).

In this paper, we consolidate lessons learned from a series of studies on Luxembourgish that highlight these issues in a particularly instructive setting. Despite its low-resource status, Luxembourgish occupies a comparatively advantageous position, given its intensely multilingual context and its typological proximity to high-resource neighboring languages, factors that would be expected to support strong cross-lingual transfer. Yet, the performance of existing language models on Luxembourgish remains limited and falls short of what might be expected given these favorable conditions.

At the same time, these structural advantages make Luxembourgish a relative best-case scenario among low-resource languages. Studying where

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 82–93

transfer still breaks down in this context allows us to identify core bottlenecks that cannot simply be attributed to extreme data scarcity, and that are therefore likely to generalize beyond this case. In this sense, Luxembourgish may serve as a practical upper bound on what cross-lingual transfer can realistically achieve for low-resource languages (Figure 1).

Across findings, we observe that cross-lingual transfer consistently shows strong potential for Luxembourgish, but also shows clear limits to its effectiveness. These constraints often arise from insufficient cross-lingual signal, driven by factors such as low target-language data quality, misalignment between downstream objectives, or unreliable target-language evaluation. In short, cross-lingual transfer is powerful, but not self-sufficient. At the same time, purely language-specific efforts pursued in monolingual isolation are equally limited, as they forgo the benefits of cross-lingual signals.

Therefore, in this paper, we argue that cross-lingual and language-specific efforts are best understood as complementary components of a shared development cycle. Drawing on a series of empirical findings, we illustrate this interplay from several angles.

Taken together, these findings suggest a nuanced perspective. Cross-lingual transfer is crucial for bootstrapping low-resource NLP, but sustainable performance requires complementary language-specific efforts that ground models in the target language's linguistic and cultural context.

By grounding this discussion in a concrete low-resource case study, we aim to challenge reductive framings that oppose cross-lingual transfer to language-specific development, and instead promote more sustainable and pluralistic research strategies.

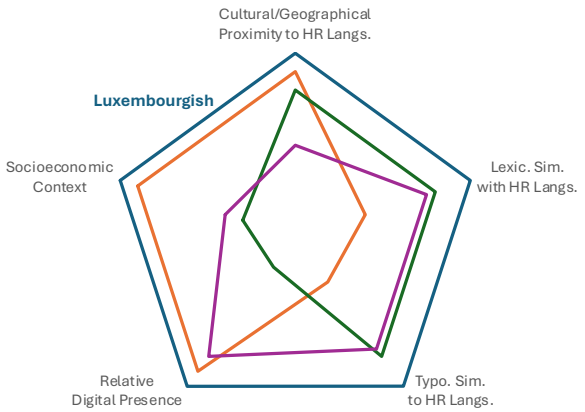


Figure 1: **Radar chart illustrating structural and sociotechnical dimensions associated with NLP inclusion and cross-lingual transfer: cultural/geographical proximity to high-resource languages, lexical similarity, typological similarity, relative digital presence, and socioeconomic context.** Luxembourgish (—) approximates an upper bound among lower-resource languages, combining structural proximity with strong institutional and socioeconomic support. In contrast, many other low-resource languages tend to lack strength along one or more of these axes. Some benefit from favorable socioeconomic conditions and digital capacity but have limited lexical or typological similarity to major high-resource languages (—) (e.g., Irish, Basque, Icelandic, Welsh, Finnish, Maltese). Others are structurally and culturally close to at least one high-resource language but have weaker socioeconomic and digital capacity, often due to limited institutionalization or concentration in economically disadvantaged regions (—) (e.g., Scots, Galician, Lombard). Others show relatively strong digital presence driven by active online communities despite limited institutional or socioeconomic support, while remaining culturally or geographically distant from most high-resource languages even if structurally related to at least one of them (—) (e.g., Haitian Creole, Nigerian Pidgin).

2 The Limits of "Emergent" Cross-Lingual Transfer

Cross-lingual transfer is often presented as an emergent property of multilingual language models (Wang et al., 2024). Empirical results demonstrate strong transfer performance of multilingual models, often without explicit alignment objectives. The implicit logic is straightforward: once sufficiently diverse multilingual pretraining data are mixed at scale, knowledge is assumed to generalize automatically across languages. This assumption is also reflected in contemporary model development practices. For instance, the technical report of Qwen-3 (Yang et al., 2025) describes the models as exhibiting "improved cross-lingual understanding and generation capabilities", yet does not report the use of any explicit cross-lingual alignment objectives or dedicated transfer-enhancing mechanisms beyond large-scale multilingual pretraining. The implicit premise is that cross-lingual competence will emerge naturally from joint multilingual training.

While this assumption is not unfounded, it is incomplete. Although cross-lingual transfer is not an intrinsic objective of most multilingual training regimes, it is a by-product of shared parameterization and distributional overlap. As a consequence,

cross-lingual transfer is neither uniform nor guaranteed, and its success depends on a range of factors, including pretraining data composition and lexical or typological proximity between languages (Philippy et al., 2023). Consequently, languages that are typologically distant, morphologically rich, or underrepresented in web-scale corpora consistently benefit less from the "emergent" transfer abilities.

In such cases, transfer performance can often be substantially improved through targeted interventions. Approaches such as supervised alignment techniques (Hämmerl et al., 2024), adapter-based frameworks (Pfeiffer et al., 2020; Parović et al., 2022), or continued pre-training (Zheng et al., 2024; Fujii et al., 2024) have all been shown to boost cross-lingual generalization. While these approaches are valuable and frequently presented as language-agnostic, many implicitly rely on the existence of high-quality resources in the target language. As we show in this paper, this assumption often fails in low-resource contexts, where the scarcity of reliable supervision fundamentally constrains the effectiveness of otherwise promising techniques.

3 Luxembourgish as a Promising Language for NLP: A Theoretical Perspective

While Luxembourgish has a relatively small speaker base of approximately 400,000, its structural characteristics and sociolinguistic profile position it (theoretically) as an ideal candidate for inclusion in multilingual NLP.

Strong institutional support and standardization. Luxembourgish is the official national language of Luxembourg, which ensures formal recognition and sustained government-backed language planning. Public institutions actively maintain linguistic resources, including standardized orthography guidelines and lexicographic databases (Conseil fir d'Lëtzebuerger Sprooch and Zenter fir d'Lëtzebuerger Sprooch, 2019). In addition, organizations such as the *Centre for the Luxembourgish Language* provide access to high-quality, curated dictionary resources (Zenter fir d'Lëtzebuerger Sprooch, 2025)[1].

Favorable conditions for cross-lingual transfer. Luxembourgish is well positioned for transfer-based approaches due to its close relationship with German and its extensive lexical borrowing from French. This dual proximity to two major high-resource languages makes it theoretically highly amenable to multilingual modeling and cross-lingual adaptation. Its use of the Latin script further reduces technical barriers related to tokenization, font handling, and script-specific modeling challenges. Overall, Luxembourgish's strong ties to Germanic languages through German and to Romance languages through sustained French influence, embed it firmly within the Indo-European language family and align it closely with many high-resource languages in NLP (Figure 2).

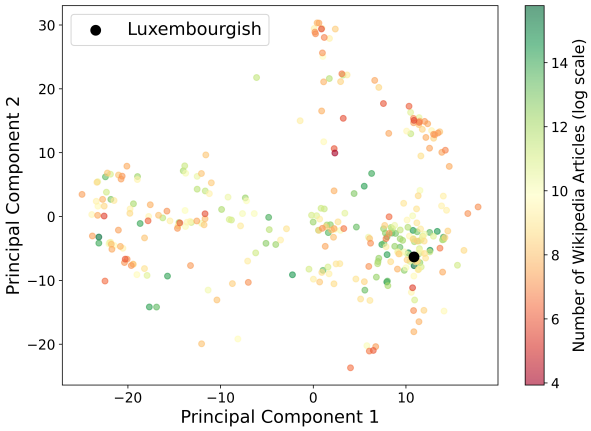


Figure 2: **PCA projection of concatenated syntactic, phonological, inventory, genetic, and geographical representations for each language.** Each point denotes a language; spatial proximity reflects overall linguistic similarity. Colors indicate the logarithm of the number of Wikipedia articles (resource proxy). Luxembourgish is located within a dense cluster of predominantly mid- to high-resource languages. More details are provided in Appendix A.1.

Cultural proximity to high-resource language settings. Luxembourgish is embedded in a broader Western European sociocultural context, with strong ties to both German- and French-speaking regions. This alignment reduces the likelihood of severe distributional mismatches for many common NLP tasks, compared to low-resource languages that are typologically or culturally more distant from the dominant training data of most multilingual language models.

A highly multilingual speaker community. Luxembourgish is embedded in a highly multilingual speech community (Fehlen et al., 2023), in which most native speakers command at least one additional language[2], often at a high or near-

[1] https://lod.lu

[2] typically French, German, or English

native level. This linguistic ecology gives rise to frequent code-switching and cross-lingual interference in both spoken and written communication, providing ecologically valid data for studying mixed-language processing, transfer dynamics, and robustness in multilingual NLP. At the same time, the community's widespread bilingual and trilingual proficiency, at least in theory, broadens the pool of potential cross-lingual annotators (though generally non-expert), thereby facilitating resource creation for tasks such as machine translation.

Disproportionately high digital presence for its size. Luxembourgish benefits from a relatively rich online ecosystem, including news content[3], its own Wikipedia edition[4], and an active multilingual media environment. Despite its small speaker population, Luxembourgish stands out as having unusually high digital coverage relative to its size. Figure 3 illustrates this imbalance: several languages with vastly larger speaker communities are represented by less data[5], while almost none with similar or smaller speaker population sizes are better represented.

Despite these advantages, Luxembourgish remains underrepresented in modern NLP systems. It is frequently absent from multilingual model documentation, inconsistently supported in pretraining data, and often excluded from standard multilingual benchmarks. Large closed-source language models remain noticeably less fluent in Luxembourgish than in neighboring high-resource languages. Language confusion, particularly between Luxembourgish and German, is common, and code-switching phenomena are often mishandled.

Moreover, foundational NLP infrastructure for Luxembourgish remains limited. Clean, widely adopted benchmarks are scarce. Core tools such as part-of-speech taggers, dependency parsers, and spellcheckers exist but often lack the robustness and evaluation depth seen for higher-resource languages. For example, the first treebank for Luxembourgish has only recently been created by Plum et al. (2024), but contains to this date merely 20 annotated sentences. In short, Luxembourgish is structurally well-positioned for inclusion in NLP, yet practically under-integrated.

[3] https://www.rtl.lu

[4] https://lb.wikipedia.org/wiki/HaaptsÃ¤it

[5] Languages such as Oromo, Sindhi, Sundanese, Igbo, and Yoruba each have between 50 and 100 times as many speakers as Luxembourgish, yet they are represented by fewer Wikipedia articles and smaller CommonCrawl corpora.

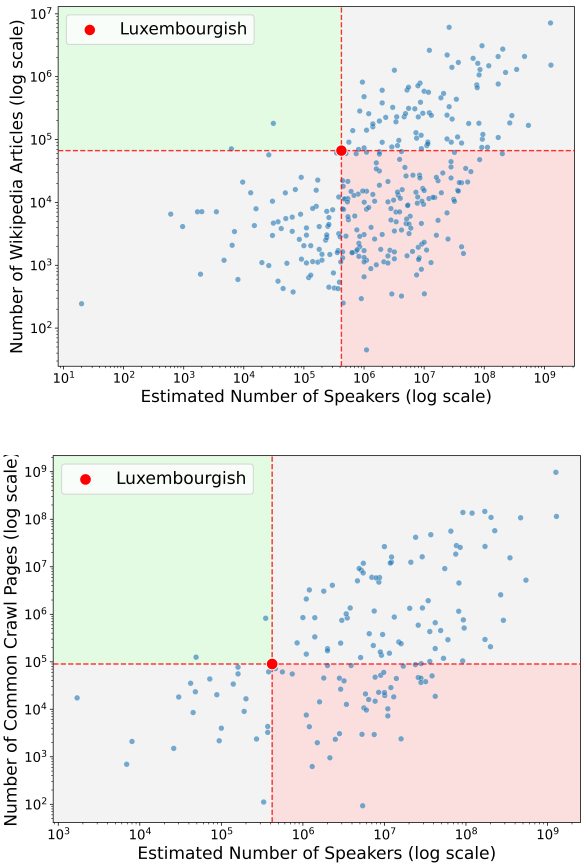


Figure 3: **Estimated number of speakers vs. number of Wikipedia articles (top) / Common Crawl pages (bottom) across languages.** Each point represents a language (both axes shown on a log scale). Shaded quadrants indicate languages with (i) fewer speakers and fewer articles (bottom left), (ii) more speakers but fewer articles (bottom right), (iii) fewer speakers but more articles (top left), and (iv) more speakers and more articles (top right). More details are provided in Appendix A.2.

This contrast makes Luxembourgish a particularly informative case study: if cross-lingual transfer were truly automatic, a language with these characteristics should already exhibit strong and stable performance across multilingual systems.

4 Pushing Transfer in Practice: Lessons from Luxembourgish

Our empirical work on Luxembourgish shows that cross-lingual transfer can be highly effective, but it is not self-sustaining. In practice, meaningful improvements require deliberate, language-aware interventions rather than purely language-agnostic scaling.

Dataset	% Non-LB Text
WikiMatrix	0.77
NLLB	21.39
KDE4	70.05
CCMatrix	99.42

Table 1: Proportion of EN–LB sentence pairs in which the LB segment was identified as non-LB. For CCMatrix and NLLB, estimates are computed from 100k-sample subsets due to corpus size. More details are provided in Appendix A.3.

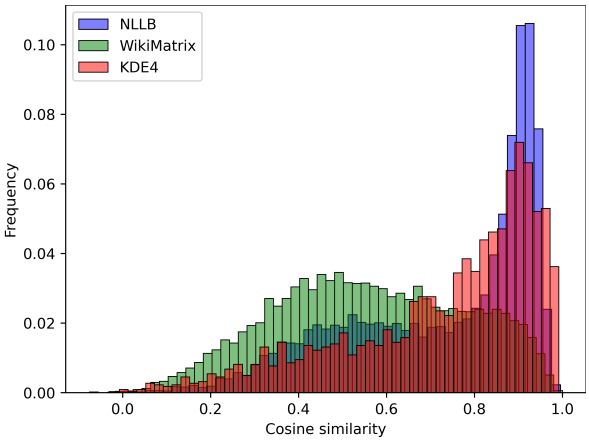


Figure 4: **Distribution of cosine similarities between EN-LB sentence pairs.** More details are provided in Appendix A.4.

Lesson 1: Reliable Resources Require Language-Specific Effort High-quality parallel data collection plays a crucial role for low-resource languages. Carefully curated bitext substantially improves embedding alignment and downstream cross-lingual performance, including cross-lingual retrieval, semantic search, and transfer-based classification. Parallel data acts as an anchoring mechanism: it reinforces semantic correspondence between languages, stabilizes multilingual representations, and reduces drift during training and adaptation.

However, parallel data is also precisely where low-resource settings often fail in practice. While modern bitext mining systems for high-resource language pairs are remarkably strong and can extract parallel sentences reliably at scale, their performance degrades substantially for low-resource languages. This is not only due to limited data volume, but also due to domain mismatch, weaker multilingual encoders for the target language, and higher noise levels in the crawled web (Kreutzer et al., 2022).

Moreover, even when parallel corpora are reported as available, they may be unusable in practice. For Luxembourgish, we find that a non-trivial portion of Luxembourgish segments in widely used parallel datasets are not actually even Luxembourgish (Table 1), and many mined sentence pairs are only weakly related or entirely unrelated (Figure 4). This pattern of data degradation is well-documented across the low-resource landscape. Similar quality concerns have been raised regarding Sinhala and Tamil (Ranathunga et al., 2024), Catalan (de Gibert et al., 2022), and a wide array of other languages analyzed in the comprehensive audit by Kreutzer et al. (2022).

This suggests that for low-resource languages, the bottleneck is often not the absence of parallel data per se, but the absence of *reliable* parallel data.

As a result, building usable parallel corpora for low-resource languages frequently requires human intervention, not necessarily in the form of full manual translation, but through targeted guidance and language-aware constraints that improve mining precision.

A concrete example is presented by Philippy et al. (2025b), who construct high-quality Luxembourgish–French and Luxembourgish–English parallel corpora from multilingual news articles published by a Luxembourgish news provider. Although this multilingual data source is publicly available, existing large-scale automatic pipelines have so far been unable to reliably mine parallel corpora for Luxembourgish from it. The difficulty lies in three practical challenges: (1) the three languages (EN, FR & LB are published under different domains and websites with differing structures, which may break automatic link-based parallel webpage mining pipelines (e.g., Liu et al., 2014); (2) not every article is published in all three languages, creating an asymmetry across language pairs; and (3) even when articles report on the same event, they are not always literal translations and may content-wise differ systematically due to target-audience differences.

Consequently, standard language-agnostic bitext crawlers and same-domain URL pairing fail, and full-document matching is unreliable. Instead, Philippy et al. (2025b) first perform broad semantic article matching, further restricting candidate pairs to articles published within a three-day window to improve robustness. Only after this coarse alignment step, they proceed to sentence-level matching, using additional constraints such as length-based

filtering to reduce noise.

In other words, high-quality parallel data could be created precisely because the mining pipeline incorporated language- and source-specific knowledge, human-provided cues, and structural constraints. This type of targeted intervention is difficult, and in some cases impossible, to replicate with a purely automatic, fully language-agnostic bitext mining pipeline.

Lesson 2: The Complementarity of Language-Specific and Cross-Lingual Signals While language-specific efforts are indispensable for developing NLP resources for Luxembourgish, these efforts should not result in linguistic isolation. As seen in lesson 1, low-resource languages require tailored data collection and preprocessing because large "language-agnostic" pipelines often fail to capture their structural specificities. However, once such resources are built, it becomes equally important to consider how they connect to the broader multilingual landscape.

Köksal et al. (2025) introduced an automatic, largely language-agnostic pipeline for constructing instruction-tuning data across multiple languages, including Luxembourgish. The core idea is to generate instructions based on existing content in the target language, thereby producing synthetic supervision without requiring large-scale manual annotation. Conceptually, such an approach promises to reduce reliance on high-resource languages and strengthen monolingual supervision.

However, subsequent analysis by Philippy et al. (2025a) identified several limitations of this approach in the Luxembourgish setting[6]. Due to the comparatively weak performance of LLMs in Luxembourgish, automatically generated instructions often suffered from reduced fluency, semantic imprecision, and structural inconsistencies. These issues resulted in a dataset of uneven quality, highlighting that language-agnostic generation pipelines are not immune to representational asymmetries between languages.

To address these limitations, Philippy et al. (2025a) adapted the pipeline specifically to Luxembourgish by incorporating additional language resources and applying more rigorous filtering and heuristic cleaning procedures. More fundamentally, the instruction generation process was modified and, instead of generating instructions directly in Luxembourgish, instructions were produced in English, French, and German. In this setup, the model only needed to understand Luxembourgish source content, but did not have to generate complex instructional phrasing in Luxembourgish itself. This shift leveraged the model's stronger generative capabilities in high-resource languages while maintaining Luxembourgish task grounding.

Few-shot prompting experiments showed that demonstration examples containing instructions in English, French, or German, while requiring outputs in Luxembourgish, led to higher performance than demonstrations written entirely in Luxembourgish. In other words, cross-lingual instructions did not merely compensate for weak Luxembourgish generation; they frequently outperformed fully monolingual prompting configurations.

This finding carries an important implication. Language-specific data is essential, but simply localizing all components of the training or prompting pipeline to the target language does not automatically improve results. When underlying multilingual representations are asymmetric, high-resource languages may provide a more stable scaffolding for task reasoning and instruction following. Carefully combining cross-lingual strengths with target-language grounding can therefore yield better performance than purely monolingual approaches. This phenomenon is further supported by the empirical findings of Chen et al. (2024) and Li et al. (2024).

This lesson reinforces the broader thesis of this paper: language-specific efforts are necessary, but they must be integrated with, rather than isolated from, cross-lingual transfer mechanisms.

Lesson 3: It Is Not Only How We Transfer, but What We Transfer A traditional and still widely adopted approach to zero-shot topic classification involves fine-tuning a model on Natural Language Inference (NLI) datasets (Yin et al., 2019). In NLI training, the model learns to determine the relationship between a premise and a hypothesis, typically classifying it as entailment, contradiction, or neutral. At inference time, topic classification is reframed within this paradigm: the input text is treated as the premise, and each candidate topic label is formulated as a hypothesis (e.g., "This text is about politics."). The model assigns an entailment score to each hypothesis, and the label with the highest entailment probability is selected as the predicted topic.

[6]These limitations do most likely also apply to many other languages.

Given the scarcity of NLI datasets for low-resource languages, this framework is often regarded as particularly suitable for cross-lingual transfer. In practice, a model can be fine-tuned on NLI data in English, or another high-resource language, and subsequently applied directly to a target low-resource language in a zero-shot setting.

However, in practice, this assumption does not always hold for low-resource languages. NLI constitutes a cognitively and linguistically demanding task, as it requires fine-grained semantic understanding, logical reasoning, and sensitivity to subtle pragmatic cues in the language under consideration. For language models operating in low-resource settings, such capabilities may be underdeveloped due to limited pretraining exposure. As a result, fine-tuning on NLI data, whether in the target language itself or transferred from high-resource languages, may fail to yield substantial gains, because the task complexity can exceed the model's effective linguistic competence in that language. In this context, it may be more beneficial to rely on a comparatively simpler proxy objective that enables the model to acquire at least partial semantic competence, rather than optimizing it for a demanding inference task from which it cannot fully benefit.

This perspective is empirically supported by the findings of Philippy et al. (2024), who investigated zero-shot topic classification for Luxembourgish. Their experiments demonstrated that the conventional NLI-based paradigm is suboptimal for Luxembourgish, regardless of whether NLI supervision is provided directly in Luxembourgish or transferred from high-resource languages such as German, English, or French. Instead, they introduced an alternative based on a Luxembourgish-specific lexical resource containing synonyms, translations, and example sentences. From this resource, they constructed a training dataset and optimized the model using a contrastive learning objective: given a sentence containing a target word, the model must determine whether a candidate word constitutes a valid synonym or translation. In other words, rather than training the model to perform premise–hypothesis entailment, they directly reinforced Luxembourgish semantic relationships through a more accessible and linguistically grounded signal, which is considerably easier to obtain than large-scale NLI annotations for many under-resourced languages.

More broadly, the results underscore a central principle: cross-lingual transfer is shaped not only by how knowledge is transferred, but by what kind of knowledge is transferred. Effective transfer requires a match between the complexity of the objective and the model's existing linguistic competence in the target language. For under-resourced languages in particular, carefully designed, linguistically grounded proxy tasks may offer a more reliable path toward robust performance than directly transferring high-level reasoning objectives.

5 Towards Balancing Transfer and Language-Specific Effort

Building sustainable NLP systems for low-resource languages requires more than applying cross-lingual transfer at scale. The Luxembourgish case shows that transfer succeeds only when certain linguistic, data, and modeling conditions are met. The following guidelines outline practical recommendations for balancing cross-lingual leverage with targeted language-specific development, treating the two as complementary components of a robust low-resource NLP pipeline.

Treat Transfer as Conditional, Not Automatic Cross-lingual transfer should not be treated as a guaranteed by-product of multilingual pretraining. While multilingual models often show strong zero-shot or few-shot capabilities, their ability to generalize across languages depends on conditions that are not uniformly satisfied across languages. Shared parameterization alone does not ensure stable or robust transfer.

Before relying on transfer-based methods, it is therefore essential to assess whether minimal target-side prerequisites are in place. These prerequisites may include sufficient pretraining exposure, coherent subword representations, basic lexical and semantic competence. If these foundational elements are weak or absent, transfer may appear unstable, inconsistent, or task-dependent.

In practical terms, this implies that researchers and practitioners should conduct lightweight diagnostic checks prior to downstream deployment. Examples include inspecting tokenization fragmentation, testing basic sentence similarity performance, or probing for language confusion in multilingual contexts. Such diagnostics help determine whether the model possesses enough internal grounding in the target language for transfer-based approaches to be effective.

In short, cross-lingual transfer depends on underlying conditions rather than being an uncondi-

tional property of multilingual models. Verifying these conditions can prevent misattributing transfer failures to model architecture or task difficulty when they are in fact rooted in insufficient target-language grounding.

Assess and Curate Resource Quality Before Use The mere availability of large-scale datasets does not guarantee their usefulness in low-resource settings. While multilingual corpora, such as automatically mined parallel datasets, often contain valuable material, they may also include substantial noise, language misidentification, weak semantic alignment, or domain mismatches, all of which tend to affect low-resource languages more severely.

As illustrated in the Luxembourgish case, a non-trivial portion of widely used parallel data may not even be written in the intended target language, and many mined sentence pairs may exhibit only superficial or partial semantic correspondence. Without careful filtering and validation, such noise can undermine representation quality, weaken alignment, and ultimately reduce downstream performance. Therefore, before integrating large-scale resources into a transfer pipeline, it is crucial to evaluate their reliability. A smaller, high-quality dataset may anchor cross-lingual representations more effectively than a large but noisy corpus.

In essence, the question is not whether data exists, but whether it is sufficiently trustworthy to serve as stable supervision.

Leverage High-Resource Languages as Scaffolding High-resource languages can serve as valuable scaffolding when developing resources for low-resource settings. Although fully monolingual resources may appear conceptually cleaner, their practical impact depends on the model's existing competence in the target language. If representation in the underlying LLM is weak, even carefully constructed monolingual datasets may yield limited gains, as the model may struggle to meaningfully interpret or generalize from them.

In such contexts, selectively incorporating high-resource languages can provide stabilizing anchors. Combining low-resource materials with related high-resource content can strengthen semantic alignment and facilitate learning, allowing the model to better contextualize and use the target-language data. Rather than diminishing language-specific efforts, this cross-lingual grounding can enhance their effectiveness within a broader multilingual framework.

Favor Targeted Interventions Over Large-Scale Expansion In low-resource settings, improvements often arise from small, targeted interventions rather than large-scale resource expansion. Addressing specific bottlenecks, such as correcting systematic tokenization issues, introducing small curated evaluation sets, or adding narrowly focused training data, can yield disproportionate gains relative to the effort involved. Because low-resource pipelines are particularly sensitive to representation gaps and data noise, incremental improvements guided by careful evaluation can be more effective than simply scaling up data collection. Prioritizing targeted refinements allows to progressively strengthen weak points in the system while maintaining overall stability.

6 Conclusion

Cross-lingual transfer is central to extending NLP to low-resource languages. However, our findings show that transfer alone rarely yields robust systems. Using Luxembourgish as a case study, we demonstrate that even languages with favorable conditions, including typological proximity to high-resource languages, a shared script, and a relatively strong digital presence, still face substantial limitations when relying solely on cross-lingual transfer.

We observe that cross-lingual transfer is most effective when supported by targeted language-specific efforts. High-quality parallel data and linguistically grounded supervision help stabilize multilingual representations and enable meaningful transfer. Conversely, purely language-specific approaches are rarely sufficient, since low-resource settings typically lack enough data to achieve strong performance without cross-lingual signals. Cross-lingual transfer and language-specific development are therefore best understood as complementary strategies. Language-specific resources ground the modeling of the target language, while cross-lingual transfer allows these resources to benefit from the broader capacity of multilingual models.

More broadly, the Luxembourgish case highlights an important implication for multilingual NLP. Structural proximity to high-resource languages does not guarantee strong cross-lingual performance. If limitations appear even under such favorable conditions, they are likely to be even

more pronounced for languages that are typologically distant, digitally underrepresented, or socioeconomically marginalized.

Ultimately, advancing NLP for the majority of the world's languages requires moving beyond the assumption that cross-lingual transfer alone can close the resource gap. Sustainable progress instead depends on integrating cross-lingual modeling with deliberate language-specific resource development.

Limitations

This work assumes that Luxembourgish approximates an upper bound for cross-lingual transfer among low-resource languages. This assumption is partly motivated by empirically studied factors known to facilitate transfer, such as linguistic similarity and lexical overlap with high-resource languages. At the same time, it also relies on additional contextual properties, such as the institutionalization of the language or its relatively strong digital presence, that are intuitively beneficial but have received less systematic empirical investigation in multilingual NLP.

As a result, the upper-bound framing should be interpreted as a theoretical approximation rather than a strictly validated claim. While these properties plausibly create favorable conditions for language technology development, they are not necessarily fully predictive of model performance. Multilingual models may also be influenced by less visible or difficult-to-measure factors, such as properties of pretraining data or other aspects of training pipelines. Consequently, although Luxembourgish exhibits many characteristics that are advantageous compared to most low-resource languages, we do not exclude the possibility that other languages with similar conditions may exist.

References

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Conseil fir d'Lëtzebuerger Sprooch and Zenter fir d'Lëtzebuerger Sprooch. 2019. *D'Lëtzebuerger Orthografie*, 6 edition. Ministère fir Educatioun, Kanner a Jugend, Lëtzebuerg.

Ona de Gibert, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé, and Maite Melero. 2022. Quality versus quantity: Building Catalan-English MT resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69, Marseille, France. European Language Resources Association.

Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. Openlid-v3: Improving the precision of closely related language identification – an experience report. *Preprint*, arXiv:2602.13139.

Fernand Fehlen, Peter Gilles, Louis Chauvel, Isabelle Pigeron-Piroth, Yann Ferro, and Etienne Le Bihan. 2023. Rp 1st results 2021 n°08 "linguistic diversity on the rise". Posted online on 12/07/2023.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Transactions of the Association for Computational Linguistics*, 13:1032–1055.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 546–566,

Bangkok, Thailand. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Le Liu, Yu Hong, Jun Lu, Jun Lang, Heng Ji, and Jianmin Yao. 2014. An iterative link-based method for parallel web page mining. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1216–1224, Doha, Qatar. Association for Computational Linguistics.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Fred Philippy, Laura Bernardy, Siwen Guo, Jacques Klein, and Tegawendé F. Bissyandé. 2025a. Luxinstruct: A cross-lingual instruction tuning dataset for luxembourgish. *Preprint*, arXiv:2510.07074.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025b. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, use a dictionary: Zero-shot topic classification for low-resource languages with application to Luxembourgish. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 97–104, Torino, Italia. ELRA and ICCL.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The first Universal Dependency treebank for Luxembourgish. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30–39, Hamburg,Germany. Association for Computational Linguistics.

Surangika Ranathunga, Nisansa de Silva, Menan Velayuthan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian's, Malta. Association for Computational Linguistics.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified metadata for thousands of languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Zenter fir d'Lëtzebuerger Sprooch. 2025. De lëtzebuerger dictionnaire. PDF version dated 14 November 2025.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7725–7738, Miami, Florida, USA. Association for Computational Linguistics.

A Details About Figures and Tables

A.1 Figure 2

Figure 2 visualizes a two-dimensional Principal Component Analysis (PCA) projection of language representations constructed by concatenating syntactic, phonological, inventory, genetic (language family), and geographical feature vectors obtained from *lang2vec* (Littell et al., 2017). For syntactic, phonological, and inventory features, we relied on the KNN-based representations to guarantee vectors of consistent dimensionality across languages. The resulting vectors were standardized and reduced to two dimensions using PCA. Language-level resource availability was operationalized as the number of Wikipedia articles per language, extracted from the Wikimedia statistics page[7]. In the visualization, each point corresponds to a language positioned according to its first two principal components, and colored by the logarithm of its Wikipedia article count (log(n+1)). Luxembourgish (lb) is highlighted in black for reference.

A.2 Figure 3

The estimated number of speakers per language was obtained from LinguaMeta (Ritchie et al., 2024).

For Wikipedia, the number of articles per language edition was collected from the Wikimedia statistics page[8].

For Common Crawl, we used language-level page counts extracted from the CC-MAIN-2026-04 crawl [9].

A.3 Table 1

To estimate the proportion of non-Luxembourgish segments, we apply automatic language identification to the Luxembourgish side of each English–Luxembourgish sentence pair. We use OpenLID-v3 (Fedorova et al., 2026) with a threshold of 0.5, which we found to perform reliably for Luxembourgish in preliminary experiments. We evaluate the largest English–Luxembourgish corpora listed in OPUS[10] (Tiedemann, 2012): WikiMatrix (Schwenk et al., 2021a), CCMatrix (Schwenk et al., 2021b), NLLB (Team et al., 2022), and KDE4[11], sampling 100,000 sentence pairs from CCMatrix and NLLB due to their size.

[7] https://meta.wikimedia.org/wiki/List_of_Wikipedias visited on March 3, 2026

[8] https://meta.wikimedia.org/wiki/List_of_Wikipedias visited on March 3, 2026

[9] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

[10] https://opus.nlpl.eu/corpora-search/en&lb

[11] https://huggingface.co/datasets/Helsinki-NLP/kde4

A.4 Figure 4

We first remove all segments predicted as non-Luxembourgish by OpenLID-v3 (Fedorova et al., 2026) from NLLB (Team et al., 2022), WikiMatrix (Schwenk et al., 2021a), and KDE4. From each filtered dataset, we then select up to 10,000 English–Luxembourgish sentence pairs (using all available pairs in cases where fewer remain). For this subset, we compute sentence embeddings with LaBSE[12] (Feng et al., 2022) and calculate the cosine similarity for each aligned pair. The resulting similarity scores are visualized as normalized histograms.

[12] https://huggingface.co/sentence-transformers/LaBSE

Building Arabic NLP from the Ground Up: Twenty Years of Lessons, Failures, and Open Problems

Wajdi Zaghouani
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

This paper reflects on twenty years of building NLP resources and research infrastructure for Arabic, a language spoken by hundreds of millions yet historically underserved relative to languages such as English or Chinese. The first decade focused on foundational linguistic infrastructure; the second shifted toward computational social science, social media analysis, and socially oriented applications. Rather than cataloguing outputs, the paper examines what the experience of building them revealed. Three counterintuitive lessons emerge: building datasets is as much a social process as a technical one; communities formed around shared tasks often matter more than the tasks themselves; and moving from language resources to computational social science exposes challenges that traditional NLP training does not address. We discuss three failures: a depression detection corpus that never reached clinical practice, a period of spreading across too many shared tasks without sufficient depth, and a long-standing assumption that Modern Standard Arabic infrastructure would transfer cleanly to dialectal tasks. These experiences suggest that the hardest problems in developing NLP for underserved communities are not linguistic but social, institutional, and epistemic, and require competencies the field rarely teaches.

1 Why Write This Paper

Most papers in NLP report successes. A new dataset achieves higher coverage than its predecessor. A new model beats the previous state of the art. A shared task attracts more participants than the last edition. This is how the field communicates progress, and it is broadly fine. But it produces a literature with a systematic blind spot: the accumulated wisdom about what does not work, what surprised the people who built things, and what they would do differently, rarely gets written down in a form anyone else can learn from.

When something fails in NLP, the failure is usually absorbed quietly. The dataset sits on a server and is not cited. The shared task runs once and is not repeated. The deployment pilot ends without a follow-up paper. The PhD student who ran the failed experiment writes up a different result instead. None of this is dishonest, exactly, but the cumulative effect is a literature that tends to overrepresent successful results relative to failed experiments, a pattern well documented in metascience more broadly (Fanelli, 2012). For senior researchers reflecting on a body of work, there is also a subtler pressure: the narrative of a research program tends toward coherence and inevitability in retrospect, because that is how people tell stories about themselves. Fighting that pressure is one reason this paper exists.

This paper attempts to fill some of that gap for one particular research program: twenty years of building Arabic NLP, spanning roughly a first decade of foundational linguistic infrastructure and, beginning in the mid-2010s, an expanding focus on social media analysis, computational social science, and socially oriented applications in and around the Arab world. The program is long enough now to look back on honestly. Some things worked better than expected. Some things failed. Some failures were obvious in retrospect. Some successes turned out to matter less than anticipated. And a few lessons emerged that seem genuinely transferable to any researcher trying to build NLP for an underserved language or community.

The paper is organized around three questions the Big Picture workshop explicitly invites: What was the larger vision? What worked and what did not? And what comes next, not just for this program but for the set of problems it represents?

One framing note before proceeding: nothing in this paper is specific to Arabic in a way that makes it irrelevant to other contexts. The challenges of building language infrastructure for a morphologi-

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 94–106

cally complex, dialectally diverse, politically sensitive, and historically under-resourced language generalize. Researchers working on low-resource African languages, South Asian languages, Indigenous languages of the Americas, or regional languages of Europe will recognize most of what follows. That generalizability is part of the point.

The paper situates itself within a body of work interrogating how NLP infrastructure gets built and for whom. Proposals like datasheets for datasets (Gebru et al., 2021) and data statements for NLP (Bender and Friedman, 2018) advocate for transparency in resource documentation. Critiques of benchmark culture (Bowman and Dahl, 2021; Ethayarajh and Jurafsky, 2020) have questioned whether leaderboard-driven evaluation advances understanding or merely advances numbers. Participatory approaches to ML (Birhane et al., 2022) argue that communities whose language is being studied should shape what is built and how. This paper operates in the same register: its claims are grounded in practice, and it is offered as a position piece that others can test against their own programs.

2 The Vision: What We Thought We Were Building

The research program began, like many in NLP, with a resource gap. In the mid-2000s, Arabic was a language spoken by hundreds of millions of people yet had no publicly available syntactically annotated corpus comparable to those for English, very limited large-scale error-corrected text, almost no social media datasets, and few evaluation benchmarks. The initial vision was straightforward: build the infrastructure, and the research will follow.

That vision was correct in the narrow sense. Infrastructure does enable research. But it was incomplete in a way that took years to fully understand: infrastructure does not automatically enable the *right* research, by the *right* people, for the *right* purposes. It enables whatever research the people who have access to it choose to do, which is not always what the people who built it imagined. This gap between what a resource is built for and what it is actually used for is probably universal in dataset work, but it is especially consequential in low-resource language NLP, where each resource represents a significant fraction of the total infrastructure and therefore shapes downstream research directions more heavily.

Roughly the first decade of the program, from 2004 to 2014, was spent primarily on foundational resources: the Arabic Treebank series (Maamouri et al., 2010; Zaghouani, 2010), the Arabic PropBank (Zaghouani et al., 2010; Palmer et al., 2008), named entity recognition systems (Zaghouani, 2012; Zaghouani et al., 2010), error correction corpora and shared tasks (Zaghouani et al., 2014; Mohit et al., 2014; Rozovskaya et al., 2015), morphological resources (Zaghouani et al., 2016), and dialect corpora (Bouamor et al., 2018). These were necessary. They also consumed enormous time, and the vision of what they were *for* evolved considerably during their construction.

Beginning in the mid-2010s, the program expanded toward using those resources to study real social phenomena: who speaks Arabic online and how (Zaghouani and Charfi, 2018; Rangel et al., 2020); how hate speech and misinformation spread in Arabic social media (Charfi et al., 2024; Zaghouani et al., 2024); how mental health signals appear in Arabic text (Zaghouani, 2018; Zaghouani and Biswas, 2025); how political discourse is framed in Arabic news and social media (Shurafa et al., 2020; Shestakov and Zaghouani, 2024); and what digital citizenship means for Arabic-speaking communities (Al Heraki and Zaghouani, 2025; Zaghouani et al., 2026).

The transition from resource building to social analysis felt natural from inside the program. It was not a pivot; it was a maturation. But looking back, it is clear that the two phases required fundamentally different competencies, and assuming the first would automatically prepare you for the second was a mistake that cost time and produced some work that was thinner than it should have been. There is also a vision-level failure worth naming explicitly: as the program moved into socially oriented work in its second decade, we started with the implicit assumption that building better NLP tools was sufficient, and that social scientists and policymakers would eventually discover and use them. This assumption is widespread in the field, and it is largely false. Policymakers rarely engage directly with NLP research publications, and social scientists are not waiting for a better hate speech classifier to appear in an ACL proceedings volume. The gap between producing a technical result and having that result affect anything outside academia is enormous, and bridging it requires forms of engagement that NLP researchers are not trained for and that academic incentive structures do not re-

ward. Recognizing this earlier would have changed how several projects were designed.

3 What Worked: Three Counterintuitive Lessons

3.1 Datasets Are Social Infrastructure, Not Just Technical Artifacts

The first lesson sounds obvious but took years to fully internalize: the value of a dataset is not in the data. It is in the community that forms around it.

The Qatar Arabic Language Bank (QALB) (Zaghouani et al., 2014) was designed as an error correction corpus. Its technical contribution was well defined: a large collection of manually corrected Arabic texts with detailed annotation guidelines (Zaghouani et al., 2015). By any standard metric of dataset quality, it was good. But what made it genuinely impactful was not the corpus itself. It was the shared tasks we organized around it at EMNLP 2014 (Mohit et al., 2014) and ACL 2015 (Rozovskaya et al., 2015), which brought together teams who would not otherwise have had reason to work on Arabic error correction. The task created a reason to engage with the data, a competitive structure that motivated effort, and a venue where results could be compared and discussed. Remove the shared task and QALB is a dataset. Add the shared task and QALB becomes a coordination mechanism for a research community.

The pattern repeated with the AraP-Tweet corpus (Zaghouani and Charfi, 2018) and the author profiling shared task at FIRE 2019 (Rangel et al., 2019), and again with ADHAR (Charfi et al., 2024), the MAHED multimodal shared task (Zaghouani et al., 2025), and ImageEval 2025 (Bashiti et al., 2025). In every case, the dataset alone attracted limited engagement. The shared task built around the dataset created a network. Some of those networks persist long after the task itself has ended: collaborations that formed during a shared task evaluation cycle have continued for years, producing work that has nothing to do with the original task.

This has a practical implication that the NLP literature rarely discusses: if you are building a dataset for an underserved language and you want it to have impact beyond your own group, the data release is not the end of the work. It is the beginning of a community-building project that requires sustained organizational effort, outreach, and maintenance. Researchers who treat dataset release as the finish line often find that their corpora are cited but not actually used in downstream work by people outside their circle. Citation is not impact. Actual use by people who were not involved in building it is closer to impact.

The Arabic Natural Language Processing Workshop (WANLP) series, which we have co-organized since 2014 (Habash et al., 2017; Zitouni et al., 2020; Habash et al., 2021; Bouamor et al., 2022), illustrated this at workshop scale. The technical papers at WANLP are individually unremarkable by top-conference standards. What WANLP built over a decade was a community: a set of people who know each other, share datasets informally, collaborate across institutions, and maintain shared standards about what counts as good work in Arabic NLP. That community is more durable than any individual resource. When a new student arrives wanting to work on Arabic NLP, they do not start from scratch; they start from a community with accumulated norms, shared baselines, and accessible mentorship. Building that takes a decade. It cannot be replicated by a grant or a paper.

3.2 The Shared Task Is a Research Instrument, Not Just an Evaluation Exercise

The second lesson follows from the first and is equally underappreciated: shared tasks, run well, are one of the most efficient mechanisms for accelerating research in a new area, precisely because they force clarity on questions that individual researchers can avoid indefinitely.

The field tends to think of shared tasks as evaluation exercises. You build a benchmark, teams submit systems, you measure who performs best. That is a reasonable description of what happens technically. But what actually happens socially is different. A well-designed shared task forces a community to agree, at least provisionally, on what a problem *is*, how it should be measured, and what counts as a valid solution. These are contested questions in any active research area, and the process of contesting them in public, with real systems and real results, advances understanding faster than individual papers arguing theoretical positions.

The CheckThat! Lab collaborations (Nakov et al., 2018, 2022; Hasanain et al., 2024), which ran over multiple years on fact-checking, misinformation, and subjectivity detection, demonstrated this clearly. The technical problem of claim checkworthiness estimation sounds well defined until you try to annotate it. Annotation disagreements

in early rounds surfaced genuine conceptual ambiguities: is a claim check-worthy because it is important, because it is verifiable, or because it is likely to be believed by a specific audience? These are not annotation errors. They are different theories of what check-worthiness means. Resolving them required cross-team discussion that would never have happened without the shared task as a forcing function. Several papers produced during CheckThat! participation were less about system performance than about reframing what the task should be, which is exactly the kind of output a maturing research area needs.

The lesson for researchers entering new domains is this: if you are building resources for a problem that has not been studied computationally before, organizing a shared task early, before your own models are mature and before the task definition has stabilized, is probably more valuable than publishing another dataset paper. The task will teach you things about your own dataset that your own analysis would not reveal, because other people bring assumptions you did not know you were making.

There is also a less obvious benefit: the shared task as a mentorship infrastructure. For a research group based in a region that is geographically distant from the major NLP conference hubs, shared task co-organization is one of the most effective ways to bring junior researchers into contact with the broader community. Students who would never have had a reason to interact with senior researchers at CMU, NYU, or the University of Edinburgh found those interactions through shared task evaluation discussions. That is not a trivial benefit.

3.3 Moving Into Social Science Requires Unlearning Some NLP Habits

The third lesson is the most uncomfortable to write about, because it touches on the limits of the training most NLP researchers receive and the assumptions that training instills without announcing them.

When the research program shifted from building linguistic infrastructure to studying social phenomena, we carried a set of habits from NLP that turned out to be poorly suited to the new domain. The most consequential of these was treating annotation disagreement as noise to be minimized.

In syntactic annotation, inter-annotator disagreement usually means the guidelines are unclear or the annotator made an error. The fix is better guidelines, more training, and adjudication procedures. This worked reasonably well for the Arabic Treebank and for QALB, both of which deal with phenomena that have defensible correct answers. It worked badly for hate speech, stance detection, and emotion recognition.

When we built the ADHAR hate speech corpus (Charfi et al., 2024) and the multi-label hate speech dataset (Zaghouani et al., 2024), we encountered annotation disagreement rates that were high by NLP standards but turned out to be meaningful rather than erroneous. Two Arabic speakers from different countries, educational backgrounds, and political perspectives often genuinely disagreed about whether a statement was offensive, and about what made it so. The disagreement was not a measurement error. It was the phenomenon. The distribution of opinions about what counts as offensive language in Arabic social media *is* the thing we wanted to understand. Collapsing that distribution into a majority-vote gold label discards the most socially important information in the annotation.

The NLP literature has become more sophisticated about this (Plank, 2022), and work on learning from disagreement, soft labels, and annotator modeling is growing. But the field's dominant evaluation framework, built around single gold labels and metrics that assume ground truth, still treats inter-annotator disagreement as a problem to be solved rather than a signal to be preserved. For any task involving human judgment about social phenomena, this is a systematic limitation that produces benchmarks which look clean but measure something subtly different from what the task was supposed to capture.

Unlearning this habit took longer than it should have. The deeper lesson is that moving from language technology to computational social science is not just a change of application domain. It is a change of epistemic framework. Social science has decades of sophisticated thinking about measurement validity, construct validity, and the relationship between operationalization and theory. NLP researchers entering social science domains typically bring none of that training. We did not bring it either, and the work suffered for it in ways that are visible only in retrospect.

4 What Did Not Work: Three Honest Failures

4.1 The Gap Between Detection and Deployment

In 2018, we published a large-scale social media corpus for Arabic youth depression detection (Zaghouani, 2018). The motivation was clear and genuine: mental health stigma in Arab societies suppresses help-seeking behavior to a degree that has measurable public health consequences. People who would not speak to a doctor do express distress online, and computational tools that could identify at-risk individuals earlier than clinical referral pathways seemed both feasible and valuable.

The corpus was well constructed. The models trained on it achieved reasonable performance on held-out test data. The paper was published and received citations. To our knowledge, no clinical deployment or policy adoption resulted from the work.

Looking back, the failure was not technical. It was a failure to engage, from the beginning, the clinical, ethical, and regulatory infrastructure that would be required for any actual deployment. Mental health detection from social media text raises serious questions that we were not equipped to answer: What is the consent framework for using someone's social media posts to infer their mental health status? What happens when the system produces a false positive? Who is responsible if an at-risk individual is flagged and no clinical response follows? How does a model trained on Arabic text from one country generalize to another where the cultural expression of distress is different? We did not have clinical collaborators, ethical review partnerships, or deployment relationships in place. We had a dataset and a model, which turned out to be the easy part.

This is not an isolated failure. A significant fraction of NLP-for-social-good work follows the same pattern: a technically interesting problem, a corpus, a classifier, a publication, and then silence. The medical NLP literature is full of systems that achieve impressive performance on benchmark datasets and have never been used by a clinician. The gap between benchmark performance and clinical deployment is well documented (Obermeyer and Emanuel, 2016) but rarely discussed candidly in the papers that report the benchmark results. The lesson has since shaped how subsequent mental health NLP work is framed. The MINDSCAPE-QA proposal currently in development begins with clinical collaborators and ethical review structures before the first annotation decision is made. The corpus design is constrained by what a clinical deployment would actually need, not by what is easiest to annotate. This is slower and harder. It is also the only approach that has any chance of producing something that matters outside a benchmark.

Based on this experience, we propose a minimal governance checklist that any mental health NLP project should satisfy before annotation begins: (a) a signed memorandum of understanding with at least one clinical partner who has agreed to participate in deployment planning; (b) IRB or equivalent ethics board approval covering the data collection protocol and annotator wellbeing; (c) a written data minimization policy specifying what is collected, from whom, for how long it is retained, and under what conditions it is deleted; (d) a false-positive response protocol defining what happens when a system flags a user who does not need intervention; and (e) a cultural adaptation review by mental health professionals familiar with the target community, because distress expression varies substantially across Arabic-speaking societies. None of these prerequisites are technically demanding. All of them require institutional relationships that take time to build. Starting that work at the proposal stage, not the publication stage, is the difference between research that could be deployed and research that cannot. The broader point, which is the most important takeaway from this section, is that NLP researchers working on socially oriented problems should integrate domain experts, ethics review boards, and policy partners at the project design stage, not as downstream consultants once the dataset and model already exist. Current practice typically reverses this order, and the cost of that reversal is research that does not translate into use.

4.2 The Breadth-Depth Tradeoff in Shared Task Participation

Between 2023 and 2025, MarsadLab participated in a large number of shared tasks across the ArabicNLP community, including AraGenEval, AraHealthQA, BAREC, NADI, PalmX, MAHED, TAQEEM, and others (Biswas et al., 2025; Bessghaier et al., 2025; Ibrahim et al., 2025; Biswas et al., 2025,?; Zaghouani et al., 2025; Bessghaier et al., 2025). The stated goal was twofold: train students through the discipline of a submission deadline, and establish the group's presence across

multiple research fronts.

The honest assessment is that this strategy produced many papers but uneven scientific depth. Several of the shared task submissions were primarily engineering exercises: fine-tuning pre-trained Arabic models with task-specific augmentation rather than contributions to understanding the underlying problems. We were optimizing for participation breadth rather than for insight, and the submissions showed it. A student who fine-tunes AraBERT for eight different tasks in a year learns something about practical NLP engineering. They do not develop the deep engagement with a single problem that produces real understanding.

Some of the most interesting problems we touched during this period deserved more sustained attention than a shared task submission cycle allows. Arabic readability, for instance, is a genuinely underexplored problem with real educational implications. Multimodal propaganda detection in Arabic memes raises hard questions about the relationship between visual and linguistic meaning that a shared task run cannot resolve. We produced submissions on both topics. We did not produce the work those topics deserved.

The practical lesson for research groups building capacity through shared tasks is that participation is useful training for junior researchers but is not a substitute for sustained engagement with a problem. There is an important distinction between using shared tasks to train students and using them as a primary publication strategy. The former is defensible. The latter tends to produce thin work that has high paper counts and low scientific impact. If a shared task consistently reveals something genuinely surprising about how models handle a problem, that surprise is worth following into original research. If the result is always approximately what you expected, the submission may not be worth the cost to the students who ran it.

4.3 The Assumption That MSA Resources Transfer to Dialectal Tasks

The early years of the program were spent building resources for Modern Standard Arabic (MSA), the formal register used in news, official communication, and education across the Arab world. MSA is important, heavily used in writing and formal speech, and linguistically well defined. The resources built during the treebank and error correction years remain in active use. But those years also embedded an assumption that took a decade to fully dislodge: that MSA infrastructure would transfer, with some adaptation, to the dialectal Arabic that most people actually use when they communicate informally online or in speech.

It does not transfer nearly as well as expected (Abdul-Mageed et al., 2021).

Arabic dialects are not stylistic variations of MSA in the way that formal and informal registers of many languages relate to each other. They have phonological systems that diverge substantially from MSA, distinct vocabulary sets with massive English and French loanword influence in some varieties, morphological patterns that differ systematically, and discourse conventions that MSA text simply does not represent. A model trained on MSA newswire text, even one fine-tuned on dialectal data, consistently makes errors on Gulf Arabic, Moroccan Darija, or Egyptian colloquial that are not just quantitatively worse but qualitatively different: the model fails on constructions that do not exist in MSA and has no framework for handling them.

We recognized this problem relatively early and contributed to the corrective through the MADAR corpus (Bouamor et al., 2018) and dialect orthography guidelines (Habash et al., 2018). But the field as a whole, including this program, continued for too long to treat dialect adaptation as a minor engineering challenge. The framing was: we have MSA resources, we add some dialectal data, we fine-tune, problem partially solved. The more accurate framing is: dialectal Arabic is a set of genuinely distinct language varieties that each require dedicated resource development from the ground up, and the fact that they share a name and a script with MSA is a source of confusion as much as a foundation to build on. This failure generalizes directly to other language communities. Any language with significant internal variation, including Hindi and Urdu, Mandarin and Cantonese, Brazilian and European Portuguese, and written and spoken varieties of Norwegian, is at risk of producing NLP infrastructure that serves its formal prestige variety while leaving the varieties that most speakers actually use in most of their daily lives severely underserved. The field tends to treat the prestige variety as the language and the other varieties as variants, which is a sociolinguistic assumption embedded in resource design decisions, not a neutral technical choice.

5 Wider Reflections: What This Means Beyond Arabic

This section steps back from Arabic specifically and asks what the experience of building this research program suggests about NLP more broadly, including for high-resource language research.

Compressed infrastructure timelines create compounding debt. English NLP had decades to build its infrastructure before the era of deep learning. Arabic NLP has tried to compress that timeline dramatically, and the compression has costs that accumulate in ways that are hard to see in the moment. When you build a treebank, then immediately build a social media corpus using the same annotation team, then immediately build a hate speech dataset using the same guidelines philosophy, you inherit the assumptions and limitations of each previous step into the next. MSA-centric assumptions persisted into social media annotation longer than they should have, because the same people were doing both and had internalized those assumptions as defaults. Low-resource language NLP programs should build in deliberate pauses for retrospective audit: structured moments to ask whether the assumptions that made sense at step one still make sense at step four. This is almost never funded and almost never done. But the audit is the work that prevents a decade of compounding debt. In practice, such an audit need not be elaborate. A structured two-day workshop involving the annotation team and at least one external reviewer, convened every two to three years, with an explicit mandate to question whether annotation guidelines, dialect coverage, and task framing still reflect the phenomena the program is trying to study, would be sufficient. The cost is low. The expected benefit, in catching assumptions that have quietly stopped being true, is high.

The hardest problems are not linguistic, they are social. The problems that most slowed this research program were not morphological complexity or dialectal diversity, though those were genuine and time consuming. They were social: getting clinical partners to engage with mental health NLP on terms that would support actual deployment; getting platform partners to provide data access that did not evaporate when a terms-of-service update changed API policies overnight; getting annotation workers to complete sensitive tasks about hate speech and trauma without suffering psychological harm; getting policymakers to treat computational findings as evidence worth engaging with rather than as academic noise.

None of these problems appear in the NLP curriculum. None of them appear as explicit criteria in grant review panels, which evaluate technical merit and broader impact but rarely require evidence that a team has the social infrastructure to bridge the gap between those two things. And yet all of them determine, more than the technical quality of the models, whether a body of work has any real-world effect.

The lesson is not that NLP researchers should become clinicians or policymakers. It is that teams working on socially oriented NLP should include people with those competencies from the beginning of a project, not as downstream consultants once the technical work is done. This is an argument for interdisciplinary team composition that the field talks about frequently and practices rarely.

Community infrastructure outlasts any individual resource. The Arabic Treebank from 2010 will eventually be superseded by larger, more diverse, neural-era corpora. The specific shared tasks from 2014 and 2015 are no longer running. Individual datasets become obsolete as the phenomena they capture shift or as the models that use them are replaced. What does not become obsolete is a research community that knows how to build things together: shared norms about annotation quality, shared practices about data release, shared venues where disagreements can be aired and partially resolved, and personal relationships that make cross-institutional collaboration something other than a bureaucratic exercise.

The most durable investments in this research program were the community investments: WANLP and its surrounding network, the shared task infrastructure, the training and mentorship relationships with students who are now doing independent research. Individual papers report results; communities build fields. For a language community that is trying to build NLP infrastructure from a deficit position, this means that community-building is not a secondary activity to be done after the real research. It is the primary activity, and the research is what sustains it. This is one place where the conventional wisdom of the field is most misleading: career incentives in NLP reward individual technical contributions, while the work that most reliably matures a research area is the unglam-

orous coordination work that no single paper can capture.

Annotator perspective is data, not noise. This deserves emphasis because the dominant evaluation culture in NLP actively works against it. When annotators disagree about whether a tweet is hate speech, or whether a social media post expresses depression, the standard procedure is majority-vote adjudication. The resulting gold label is clean, model-friendly, and epistemically misleading. If 40% of Arabic speakers consider a statement offensive and 60% do not, that split is a fact about the social reality of hate speech in Arabic-speaking communities. A gold label of "not hate speech" erases that fact. Systems trained on adjudicated labels also learn to replicate the adjudicator's demographic perspective, regardless of what the guidelines said about neutrality, a problem documented for English (Sap et al., 2019; Davidson et al., 2019) but understudied for Arabic. The pressure to produce clean gold labels moreover creates perverse design incentives: guidelines minimize disagreement rather than capture genuine conceptual boundaries, and tasks are scoped to problems annotators will agree on, which tends to mean problems that are less socially significant. The field is slowly learning this (Plank, 2022), but evaluation frameworks built on single gold labels remain the norm. High inter-annotator agreement is a quality signal only conditionally. For tasks involving social judgment, it may mean the task was scoped too narrowly to matter.

A concrete corrective: release per-annotator labels alongside aggregated labels, include annotator demographic summaries in datasheet documentation (Gebru et al., 2021), and report model performance under majority-vote, soft-label, and per-annotator aggregation schemes so that sensitivity to label aggregation is visible rather than assumed away. This is a documentation norm, not a research burden.

6 What Comes Next

From detecting harm to understanding flourishing. The bulk of this program's social media work has been oriented toward detection: hate speech, misinformation, mental health signals, polarization. This is necessary work. But designing systems primarily around harm detection produces a distorted picture of online discourse. Recent dataset work on hope speech (Sharqawi and Zaghouani, 2026; Zaghouani and Biswas, 2025), women's empowerment discourse (Zaghouani et al., 2026), and social cohesion (Ali Al-Athba and Zaghouani, 2026) reflects a deliberate shift toward also modeling constructive discourse. You cannot design interventions to support something you have not characterized, and a research literature on harm without a corresponding literature on flourishing is an incomplete basis for policy.

Arabic LLM evaluation as a first-class research priority. Large language models are being deployed in Arabic-speaking contexts at scale, yet systematic evaluation infrastructure does not yet exist. PalmX (Biswas et al., 2025) and AraGenEval (Biswas et al., 2025) are early steps. Arabic LLM evaluation must at minimum span four axes: dialectal coverage across Gulf, Levantine, Egyptian, Maghrebi, and Sudanese varieties; cultural fidelity; safety auditing for harms Arabic-speaking communities would recognize but English reviewers would not; and factual accuracy on Arab history and current affairs. Two decades of dialectally diverse, socially grounded Arabic corpora are exactly the substrate this agenda needs.

Taking the translation problem seriously. The digital citizenship project underway attempts to translate findings into something that affects the media literacy of young Arabic speakers in Qatar. Whether it succeeds will depend on pedagogical and institutional work far more than on NLP. For most of this program's history, a successful project produced publications and trained students. Those outcomes remain entirely inside academia. Building something that reaches a 14-year-old in a Doha school navigating the phenomena this program has spent a decade studying requires a different theory of impact and a willingness to be evaluated by criteria unrelated to citation counts. That shift is overdue.

7 Key Takeaways for the NLP Community

A dataset without a community is an archive. Impact comes from the organizational and community-building work surrounding a release, not from the data alone.

Run the shared task before your models are ready. Tasks designed before the field has converged on a problem force honest engagement with ambiguity that post-hoc evaluation cannot surface.

Prestige-variety infrastructure does not transfer downward automatically. The formal standard of a language is not the language. Building NLP for the prestige variety first and adapting later consistently underserves the communities who most need the technology.

The hardest competencies for socially oriented NLP are the ones NLP programs do not teach. Clinical partnership, ethical governance, and policy translation determine whether technically sound work produces any effect outside a benchmark. Integrating these competencies from the start of a project, not after the technical work is done, is the practical change that follows from twenty years of this experience.

8 Conclusion

What went wrong: assuming infrastructure automatically produces impact; treating annotation disagreement as noise; skipping clinical governance in mental health work; chasing shared task breadth over depth; and treating MSA-to-dialect transfer as a minor engineering task. What was surprising: how much community building mattered relative to any individual resource, and how consistently the hardest obstacles were social rather than technical. What comes next: modeling flourishing alongside harm, rigorous Arabic LLM evaluation, and translating findings into tools that reach communities outside academia.

None of these conclusions are uniquely Arabic. Any researcher building NLP for an underserved language community will recognize the compressed timelines, the annotation philosophy that travels from prestige registers to vernacular ones without sufficient scrutiny, the gap between detection and deployment, and the community-building work that turns individual resources into durable fields. The Arabic experience does not solve those problems. It has accumulated enough honest failure, and enough honest success, to be worth learning from.

Limitations

This paper reflects on a twenty-year research program through the perspective of one researcher and therefore necessarily represents a partial account. Large collaborative programs evolve through the contributions of many students, collaborators, and institutional partners. Those collaborators would likely emphasize different episodes, successes, or failures, and may interpret some of the lessons differently. The narrative presented here should therefore be read as a reflective synthesis rather than a definitive historical record of the program.

A second limitation concerns the evidentiary basis of the claims. The arguments in this paper are grounded primarily in accumulated experience rather than systematic empirical analysis across multiple programs or language communities. While many of the lessons described here likely generalize to other low-resource language contexts, this generalization has not been formally tested. Researchers working on African languages, South Asian languages, or Indigenous language communities may encounter similar structural challenges but also face distinct institutional and sociolinguistic conditions that shape infrastructure development differently.

Third, the paper focuses heavily on research infrastructure, dataset development, and community organization within the academic NLP ecosystem. It does not provide a systematic evaluation of downstream real-world impact. As discussed in Section 4.1, the gap between research outputs and deployment remains large, and the extent to which any particular dataset or model influences policy, education, or public discourse is difficult to measure. This paper therefore evaluates impact primarily through indicators internal to the research community (shared tasks, collaborations, student training, and dataset reuse) rather than through external social outcomes.

A fourth limitation concerns retrospective interpretation. Reflective papers inevitably introduce narrative coherence into events that were experienced as messy and contingent when they occurred. Some failures appear obvious in hindsight but were not obvious at the time, given the constraints under which projects were conducted. Similarly, some successes described here may partly reflect broader shifts in the field rather than decisions made within this specific research program.

Finally, the analysis emphasizes structural and institutional lessons rather than detailed technical analysis. Readers seeking technical evaluation of specific models, datasets, or annotation frameworks should consult the original publications referenced throughout the paper. The purpose of this article is not to document individual technical contributions but to extract higher-level insights about the process of building NLP infrastructure and research communities over an extended period.

These limitations should be understood not as weaknesses but as boundaries on the type of claim the paper is making. The goal is to contribute a practice-based perspective that complements more formal analyses of dataset design, evaluation methodology, and NLP system development.

Ethical Considerations

Several strands of work discussed in this paper involve sensitive forms of data collection and analysis, particularly research on mental health, hate speech, online harassment, and political discourse. These areas raise ethical questions related to privacy, consent, annotator wellbeing, and potential misuse of computational tools.

8.1 Privacy and Data Governance

Much of the research described here relies on publicly available social media data. Although such data are technically accessible, their use still raises important ethical questions regarding user expectations of privacy and the potential for unintended harm. Individuals posting online may not anticipate that their content will be aggregated into research datasets or analyzed by automated systems.

To mitigate these concerns, responsible data practices should include data minimization, removal of personally identifiable information when possible, and careful documentation of collection protocols. When datasets are released publicly, documentation such as datasheets or data statements should clearly describe the data sources, collection procedures, and known limitations. These practices help ensure transparency and allow downstream users to understand the context in which the data were created.

8.2 Mental Health Research and Risk of Harm

Research on depression detection and other mental health signals from social media text raises particularly serious ethical concerns. Predictive systems that attempt to infer psychological states from language can generate false positives or false negatives with significant consequences if deployed in real-world contexts.

As discussed in Section 4.1, a key lesson from early work in this area is that technical model development alone is insufficient. Responsible research in this domain requires collaboration with clinical experts, ethical review structures, and clear protocols for how predictions would be used in practice. Without these safeguards, there is a risk that models could be misinterpreted as diagnostic tools or used in ways that harm the individuals they aim to help.

8.3 Annotator Wellbeing

Annotation work involving hate speech, harassment, or traumatic content can impose psychological burdens on annotators. Exposure to harmful language or disturbing material over extended periods can negatively affect mental health. Ethical dataset construction therefore requires attention not only to annotation quality but also to the wellbeing of the people performing the annotation.

Practical safeguards may include limiting daily exposure to harmful content, providing clear opt-out mechanisms for annotators, offering mental health resources, and designing annotation workflows that distribute difficult tasks across teams rather than concentrating them on a few individuals.

8.4 Bias, Representation, and Cultural Context

Arabic is a linguistically and culturally diverse language family encompassing dozens of dialects across multiple regions. Datasets that overrepresent certain dialects, social groups, or geopolitical contexts risk producing models that perform unevenly across communities. Similarly, annotation decisions about hate speech, offensiveness, or political framing are shaped by cultural and demographic perspectives.

Ethical NLP research should therefore treat annotation disagreement and demographic variation as important signals rather than as noise to be eliminated. Where possible, dataset documentation should describe annotator backgrounds and dataset composition so that users can interpret model behavior appropriately.

8.5 Dual-Use Risks

Finally, tools developed to detect harmful content or analyze public discourse may also be used for surveillance or censorship. Systems designed to identify political framing, online dissent, or controversial speech could potentially be repurposed in ways that restrict legitimate expression.

Researchers cannot fully control how computational tools are used after publication, but awareness of dual-use risks should inform dataset design,

documentation, and release decisions. Clear documentation of intended research uses and limitations can help reduce the likelihood that models are interpreted as authoritative decision-making systems.

Taken together, these considerations highlight a broader point emphasized throughout this paper: the ethical challenges of socially oriented NLP research are not peripheral concerns but central design constraints. Addressing them requires interdisciplinary collaboration among NLP researchers, social scientists, clinicians, and policy experts from the earliest stages of a project.

Acknowledgments

Some of the projects and research activities reported in this paper received partial support from the Qatar National Research Fund (QNRF) and QRDI under grants MCSC 02-0217-250013, NPRP 14C-0916-210015, NPRP 13S-0206-200281, CWSP 18-W-0206-20044, NPRP 11S-1112-170006, and NPRP 9-175-1-033.

References

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2021). ARBERT and MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 7088–7105.

Ali Al-Athba, A. and Zaghouani, W. (2026). Cohesion-6K: An Arabic Dataset for Analyzing Social Cohesion and Conflict in Online Discourse. In *Proceedings of LREC 2026*.

Al Heraki, H. and Zaghouani, W. (2025). Analyzing Digital Polarization on Hijab: A Dataset of Annotated YouTube Comments. In *Proceedings of ICWSM 2025*, pages 2350–2360.

Bashiti, A., Aljabari, A., Hamoud, H. K., Biswas, M. R., Shalash, B. M., Jarrar, M., Zaraket, F., Mikros, G., Asgari, E., and Zaghouani, W. (2025). ImageEval 2025: The First Arabic Image Captioning Shared Task. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 376–389.

Bender, E. M. and Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bessghaier, M., Ibrahim, S., Biswas, M. R., and Zaghouani, W. (2025). MarsadLab at AraHealthQA: Hybrid Contextual-Lexical Fusion with AraBERT. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 233–238.

Bessghaier, M., Biswas, M. R., Dhouib, A., and Zaghouani, W. (2025). MarsadLab at TAQEEM 2025. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 998–1002.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., and Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2022)*.

Biswas, M. R., Bessghaier, M., Alam, F., and Zaghouani, W. (2025). MarsadLab at AraGenEval Shared Task. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 88–93.

Biswas, M. R., Attia, K., Ibrahim, S., Bessghaier, M., and Zaghouani, W. (2025). MarsadLab at NADI Shared Task. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 752–756.

Biswas, M. R., Ibrahim, S., Attia, K., Alam, F., and Zaghouani, W. (2025). MarsadLab at PalmX Shared Task. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 818–824.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of LREC 2018*.

Bouamor, H., Al Khalifa, H., Bougares, F., Darwish, K., Rambow, O., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W. (Eds.) (2022). *Proceedings of the Seventh Arabic Natural Language Processing Workshop*. Association for Computational Linguistics.

Bowman, S. R. and Dahl, G. E. (2021). What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of NAACL 2021*, pages 4843–4855.

Charfi, A., Bessghaier, M., Atalla, A., Akasheh, R., and Zaghouani, W. (2024). Hate Speech Detection with ADHAR: A Multi-Dialectal Hate Speech Corpus in Arabic. *Frontiers in Artificial Intelligence*, 7, Article 1391472.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of EMNLP 2020*, pages 4846–4853.

Fanelli, D. (2012). Negative Results Are Disappearing from Most Disciplines and Countries. *Scientometrics*, 90(3):891–904.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92.

Habash, N., Diab, M., Darwish, K., El-Hajj, W., Al-Khalifa, H., Bouamor, H., Tomeh, N., El-Haj, M., and Zaghouani, W. (Eds.) (2017). *Proceedings of the Third Arabic Natural Language Processing Workshop*. Association for Computational Linguistics.

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Al-Shargi, F., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of LREC 2018*, pages 3628–3637.

Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S. (Eds.) (2021). *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics.

Hasanain, M., Suwaileh, R., Weering, S., Li, C., Caselli, T., Zaghouani, W., Barrón-Cedeño, A., Nakov, P., and Alam, F. (2024). Overview of the CLEF-2024 CheckThat! Lab Task 1. In *Working Notes of CLEF 2024*, pages 276–286.

Ibrahim, S., Biswas, M. R., Bessghaier, M., and Zaghouani, W. (2025). MarsadLab at BAREC Shared Task 2025. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 274–279.

Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D., and Ciul, M. (2010). From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In *Proceedings of LREC 2010*.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The First QALB Shared Task on Automatic Text Correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47.

Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., and Da San Martino, G. (2018). Overview of the CLEF-2018 CheckThat! Lab. In *Lecture Notes in Computer Science*.

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J. M., Mandl, T., Miguez, R., Caselli, T., Kutlu, M., and Zaghouani, W. (2022). The CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In *Proceedings of ECIR 2022*, pages 416–428.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the Future: Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13):1216–1219.

Palmer, M., Babko-Malaya, O., Bies, A., Diab, M., Maamouri, M., Mansouri, A., and Zaghouani, W. (2008). A Pilot Arabic PropBank. In *Proceedings of LREC 2008*.

Plank, B. (2022). The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of EMNLP 2022*.

Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., and Sánchez-Junquera, J. (2019). Overview of the Track on Author Profiling and Deception Detection in Arabic. In *Working Notes of FIRE 2019*, pages 70–83. CEUR Workshop Proceedings.

Rangel, F., Rosso, P., Zaghouani, W., and Charfi, A. (2020). Fine-grained Analysis of Language Varieties and Demographics. *Natural Language Engineering*, 26(6):641–661.

Rozovskaya, A., Bouamor, H., Habash, N., Zaghouani, W., Obeid, O., and Mohit, B. (2015). The Second QALB Shared Task on Automatic Text Correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing, ACL 2015*.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of ACL 2019*, pages 1668–1678.

Sharqawi, E. A. and Zaghouani, W. (2026). AraHopeCorpus: Annotation Guidelines and Dataset for Hope Speech in Arabic Social Media Crisis Discourse. In *Proceedings of LREC 2026*.

Shestakov, A. and Zaghouani, W. (2024). Analyzing Conflict Through Data: A Dataset on the Digital Framing of Sheikh Jarrah Evictions. In *Proceedings of the Second Workshop on NLP for Political Sciences, LREC-COLING 2024*, pages 55–67.

Shurafa, C., Darwish, K., and Zaghouani, W. (2020). Political Framing: US COVID-19 Blame Game. In *Social Informatics (SocInfo 2020)*, LNCS 12467. Springer.

Zaghouani, W. (2010). Arabic Treebank Part 1 Version 4.1. LDC Catalog No. LDC2010T13. Linguistic Data Consortium.

Zaghouani, W., Pouliquen, B., Ebrahim, M., and Steinberger, R. (2010). Adapting a Resource-Light Highly Multilingual Named Entity Recognition System to Arabic. In *Proceedings of LREC 2010*.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The Revised Arabic PropBank. In *Proceedings of the 4th Linguistic Annotation Workshop, ACL 2010*.

Zaghouani, W. (2012). RENAR: A Rule-Based Arabic Named Entity Recognition System. *ACM Transactions on Asian Language Information Processing*, 11(1), Article 2.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of LREC 2014*, pages 2362–2369.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus. In *Proceedings of LAW IX, co-located with NAACL 2015*.

Zaghouani, W., Hawwari, A., Diab, M., O'Gorman, T., and Badran, A. (2016). AMPN: A Semantic Resource for Arabic Morphological Patterns. *International Journal of Speech Technology*, 19(2):281–288.

Zaghouani, W. and Charfi, A. (2018). AraP-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of LREC 2018*.

Zaghouani, W. (2018). A Large-Scale Social Media Corpus for the Detection of Youth Depression. *Procedia Computer Science*, 142:347–351.

Zaghouani, W., Mubarak, H., and Biswas, M. R. (2024). So Hateful! Building a Multi-Label Hate Speech Annotated Arabic Dataset. In *Proceedings of LREC-COLING 2024*, pages 15044–15055.

Zaghouani, W. and Biswas, M. R. (2025). EmoHope-Speech: An Annotated Dataset of Emotions and Hope Speech in English and Arabic. In *Proceedings of RANLP 2025*, pages 1406–1412.

Zaghouani, W., Biswas, M. R., Bessghaier, M., Ibrahim, S., Mikros, G., Hasnat, A., and Alam, F. (2025). MAHED Shared Task: Multimodal Detection of Hope and Hate Emotions in Arabic Content. In *Proceedings of ArabicNLP 2025 Shared Tasks*, pages 560–574.

Zaghouani, W., Bessghaier, M., Biswas, M. R., and Ibrahim, S. A. (2026). Audience Engagement with Arabic Women's Social Empowerment and Wellbeing: A Decadal Corpus. In *Proceedings of LREC 2026*.

Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W. (Eds.) (2020). *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics.

Speaking of Language: Reflections on Metalanguage Research in NLP

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Abstract

This work aims to shine a spotlight on the topic of metalanguage. We first define metalanguage, link it to NLP and LLMs, and then discuss our two labs' metalanguage-centered efforts. Finally, we discuss four dimensions of metalanguage and metalinguistic tasks, offering a list of understudied future research directions.

1 Introduction

Language is so powerful that it can be reflected back on itself. All of the following English sentences expressly concern linguistic inventories, structures, and behaviors:

(1) People often confuse the "George" universities.

(2) The expression "kick the bucket" is an idiom meaning "die".

(3) Hebrew has a zero copula in the present tense.

(4) Now read it in an Irish accent.

Sentences such as these may concern a particular instance of language use, or properties of a language or speaker in general; either way, they are **metalinguistic** in making linguistic phenomena (rather than the external world) the subject matter of a linguistic utterance.

Any kind of formal notation that elucidates linguistic properties can also be considered metalanguage, e.g.:

(5) One morning I shot [$_{\text{NP}}$ an elephant in my pajamas].

(6) an$_{\text{DT}}$ elephant$_{\text{NN}}$ in$_{\text{IN}}$ my$_{\text{PRP\$}}$ pajamas$_{\text{NNS}}$

(7) $shot(I, elephant) \wedge in(elephant, pajamas)$

Both kinds of metalanguage enable humans to reflect on linguistic form, meaning, and use, which is why metalanguage is central to fields such as linguistics, language pedagogy, rhetoric, and even law and policy.

The recent advent of fluent multipurpose chatbot tools powered by large language models (LLMs) puts a new focus on metalanguage—or at least we argue that it should. Despite some research on metalanguage and on the metalinguistic abilities of LLMs, the topic remains an understudied one. Yet many real-world tasks in spheres such as language learning and law rely on metalinguistic reasoning rather than simple content understanding. Evaluating whether LLMs can process, generate, and learn from metalanguage therefore provides a crucial test of adaptability and robustness. Zeroing in on the processing of metalanguage, we believe, will ultimately serve applications in important domains where meaning and interpretation are central.

2 Defining "Metalanguage"

The human ability to use language to communicate rests on knowledge that is mostly *implicit*. Linguistics and other fields **communicate explicit conceptualizations of language phenomena** via metalanguage (Berry, 2005). For example, the statement "5-year-old children can productively form regular plurals of nouns" will be incomprehensible to most adults, not to mention the 5-year-olds in question. Not only do the plural-producing 5-year-olds not know the *word* "noun", they presumably have not learned any grammar to the point of comprehending the *concept* of "noun". The metalanguage of a discipline such as linguistics thus reflects many of the concepts at the core of disciplinary expertise.

We provide brief terminological definitions below to setup the stage for the rest of the paper. **Natural metalanguage** is text in a natural language that is interpreted to be about language, grounded in particular utterances or general behavior by a speaker or language community. **Symbolic metalanguage** is formal notation that encodes aspects

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 107–118

	symbolic	natural
instance-level	tagging, parsing	answering a question about the grammaticality of a sentence
system-level	grammar rule induction	generating a dictionary definition for a word

Table 1: Examples of metalinguistic description tasks (where the metalanguage is language-system–oriented in nature). The metalinguistic inputs and/or outputs in question may be symbolic or natural, and can be formulated at the level of individual instances (tokens) or at the level of an entire system of language (generalizations).

of language (or instances of language use) in a way that explicitly surfaces relationships and patterns within the language system.[1] In general, *a* symbolic metalanguage is a formal system or controlled vocabulary for describing a phenomenon. Finally, we note that quotations (a reference to another speech act or text) are a specific instance of metalanguage that serves (usually) a more narrow/specific communicative purpose.

Broadly, a **metalinguistic task** is one that necessarily processes, leverages, produces, or is defined with metalanguage. Below we discuss kinds of metalinguistic tasks in the context of LLMs (§3.1).

3 Why Study Metalanguage (in NLP)?

Metalinguistic inquiry of one sort or another is common in a wide range of fields and applications. Linguistics, of course, is entirely about the study of language. Consider also: language teaching and learning (e.g., second language learners asking for advice about how to use a word or construction); lexicography; literary studies; and law (the interpretation of legal rules). In these domains, amateur or professional language analysts *consume* instances of language use (in some cases using highly customized corpus search tools), and/or *produce* large quantities of textual metalanguage in textbooks, dictionaries, online discussion forums, legal opinions, and scholarly publications. One impetus, then, for NLP study of metalanguage is to develop tools for metalinguistic inquiry.

Another motivation comes from the inherent goals of modeling language and linguistic meaning. For humans engaged in scientific work, natural as well as formal languages are indispensable when developing theories and making predictions. Within NLP, the tradition of analyzing linguistic grammar and meaning with symbolic structures is one incarnation of metalinguistic NLP (Opitz et al., 2025). (Thus, the study of syntactic parsing, for example, is inherently metalinguistic.)

3.1 Metalanguage and LLMs

In light of the current fascination with large language models (LLMs), it is worth breaking down where metalanguage may come into play in this paradigm, and what studies have or might shed light on its role.

The phenomenon of metalanguage confronts different forms of interaction with an LLM system. We outline these metalinguistic modes below.

Metalinguistic instructions. If a chatbot user explicitly requests a piece of writing—whether it is a summary, translation, homemade pizza recipe, or humorous limerick about cheese—that user is speaking metalinguistically.[2] Thus the practices of instruction tuning and prompting are tied up with metalanguage to some extent. This implies that the presence (or not) and the extent of metalinguistic intent should perhaps inform the evaluation of LLMs in general: do they perform better or worse when the instructions are metalinguistic or not?

Metalinguistic description tasks. By this we mean tasks that center *systematic* aspects of language (or a language). Requesting a definition, grammatical analysis, or explanation of meaning all presuppose a set of conventions constituting a linguistic system, and seek a description that somehow unpacks the conventions or how they apply to a particular instance. Table 1 illustrates tasks that involve *symbolic* or *natural* metalanguage describing linguistic *instances* or *generalizations*. Not all language-manipulation tasks qualify here: machine translation of a sentence, for example, does not in itself reference the organization of either language system (though a translation may be part of a larger

[1] We focus here on metalanguage that is expressed in human-understandable formats, so we will not discuss "style vectors" or "task embeddings" as potential metalanguage.

[2] We consider an instruction metalinguistic if it makes any reference to communication or the linguistic nature of the input or output. Thus "Tell me a joke." and "limerick about cheese" are metalinguistic; "What is the capital of France?" is not.

explanation of linguistic patterns constituting a metalinguistic description).[3]

Metalinguistic interpretation and explanation. To better understand how "black box" models of language operate, one route is to look for correlations between representations or behaviors in the model and their counterparts as described metalinguistically for human language. For example, attempts have been made to localize grammatical knowledge amongst a tangled web of neural network components (e.g., Liu et al., 2019; Tenney et al., 2019; Aoyama and Schneider, 2022; Wang et al., 2022). Other work has investigated model behavior vis-à-vis its generations or probability distributions (e.g., Warstadt et al., 2020; Hu and Levy, 2023). We return to metalinguistic interpretability in §6.

4 A Tale of Two Labs

Research programs within the authors' research groups have prioritized metalinguistic NLP.[4] We give an overview of several such efforts:

- in Anastasopoulos's lab at George Mason University, studies featuring documentary linguistics and low-resource NLP (§4.1 and §4.2);
- in Schneider's lab at Georgetown University, studies motivated by second language learning and legal interpretation (§4.3 and §4.4).

The concluding sections will discuss emerging themes and dichotomies.

4.1 Learning from Reference Grammars

The idea of learning using already-defined grammars is not new; it is in fact one of the first ideas tried out in the early era of symbolic NLP. However, using the text of a reference grammar *as is* to facilitate the creation of language technologies for a given language has only recently come within reach, due to LLMs' capabilities.

Calls to "mobilize the archive", in particular for data-scarce languages, aim to encourage research that leverages linguistic documentation efforts (Bird, 2022). In seminal work, Tanzer et al. (2024) did exactly that, incorporating dictionaries, sentences, and grammar books to perform machine translation using LLMs in a zero-shot setting, i.e., in a language without *any* other data available ("Machine Translation from One Book"). This is perhaps akin to how a documentary linguist or any second-language learner could potentially learn a new language (at least if they did not have access to a teacher or said language's speakers).

In follow-up work, Hus and Anastasopoulos (2024) explored this grammar-based paradigm on 16 languages. Our initial findings were particularly encouraging. For translating extremely low-resource languages like Chuvash, Dogri, and Kalamang into English, providing a combination of dictionary entries and the full grammar book yields almost usable translations (with chrF++ scores between 25–55). Other ongoing work attempts to integrate such approaches into a documentary linguist's workflow—in this case, working on Nepal's Kulung languages (Taguchi et al., 2025).

However, concurrent and followup work has called into question whether the current generation of LLMs can trully understand and leverage metalinguistic content in the form of a reference grammar (Aycock et al., 2025; Marmonier et al., 2025). Regardless, we believe that the potential for reducing data requirements for under-resourced languages of already extremely under-served communities makes this a worthy research direction.

4.2 Inducing and Describing Patterns in Data

Another line of work aims at *generating* metalanguage (symbolic or natural). In particular, we aim at simulating the work of a linguist or a language teacher, producing output that describes a language system, based on raw text samples.

The notion of describing a language "in its own terms" based solely on raw data has an established tradition in descriptive linguistics (Harris, 1951). Early work included discovering morphosyntactic agreement (Chaudhary et al., 2020) or lexical selection preferences (Chaudhary et al., 2021), tying it also to educational applications, by presenting these rules along with selected examples to be used by L2 teachers (Chaudhary et al., 2023) – see an illustration in Figure 1. Earlier work by Howell et al. (2017) aimed to predict the case systems of endangered languages and Zamaraeva (2016) inferred morphotactics from IGT using k-means clustering.

[3] There is probably no bright line that demarcates this category. Between "Proofread this paragraph" and "Indicate the grammatical errors in this paragraph", the latter more overtly invokes a goal of linguistic system–based description, but in practice these serve very similar user needs. An alternative definition going beyond our focus could take into account user intent so that e.g., producing a free translation for an interlinear gloss of an example in a reference grammar could qualify as such a metalinguistic description task.

[4] Supported in part by NSF awards "CAREER: Metalinguistic Natural Language Understanding" (Schneider) and "CAREER: Leveraging Grammar Books to Develop Language Technologies for Data-Scarce Languages" (Anastasopoulos).

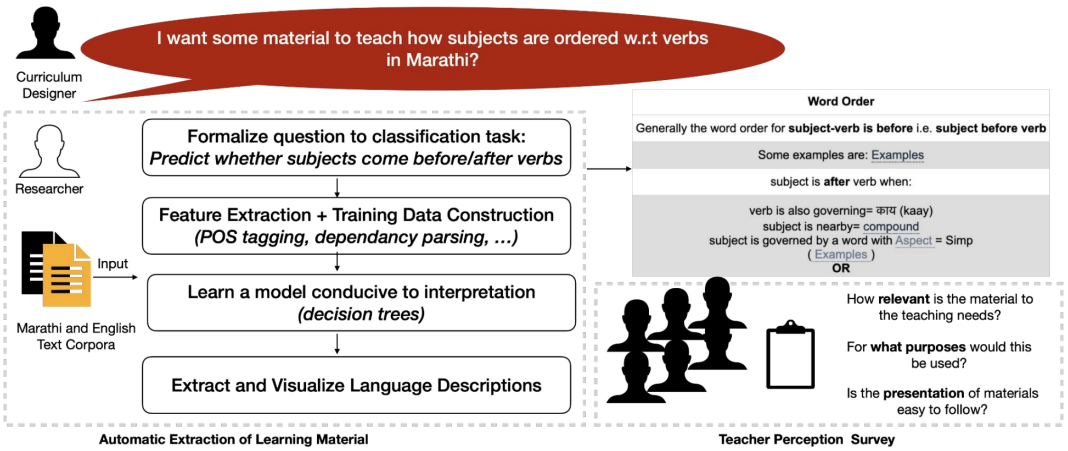


Figure 1: Workflow for the collaboration of NLP researchers and language-learning curriculum designers, to create pedagogical materials (Chaudhary et al., 2023). The input and intermediate and final outputs include metalanguage.

The above line of work is somewhat orthogonal to the works that have typologists as their target audience. Linguists and researchers have long undertaken initiatives to collect linguistic properties in machine-readable formats. *WALS* (Dryer and Haspelmath, 2013) is one such example which can tell us, for instance, that English objects occur after verbs, or that Turkish pronouns have symmetrical case. *Grambank* (Skirgård et al., 2023) is the latest typological database: it covers 2,467 language varieties, capturing a wide range of grammatical phenomena in 195 features, from word order to verbal tense and many other well-studied comparative linguistic variables. Most works on the NLP side aim to fill in the missing entries in such databases, producing a structured typological description of a language based on raw text samples (Daumé and Campbell, 2007; Bjerva et al., 2020, *inter alia*). See Baylor et al. (2023) for additional discussion on the usefulness of such work for NLP in general. More recently, Arčon et al. (2026) converted each entry of the WALS database into a question-answer pair, in order to test LLMs' typological knowledge. Hus and Anastasopoulos (2026) pose similar typological questions but they also assume access to reference grammars for the languages in question.

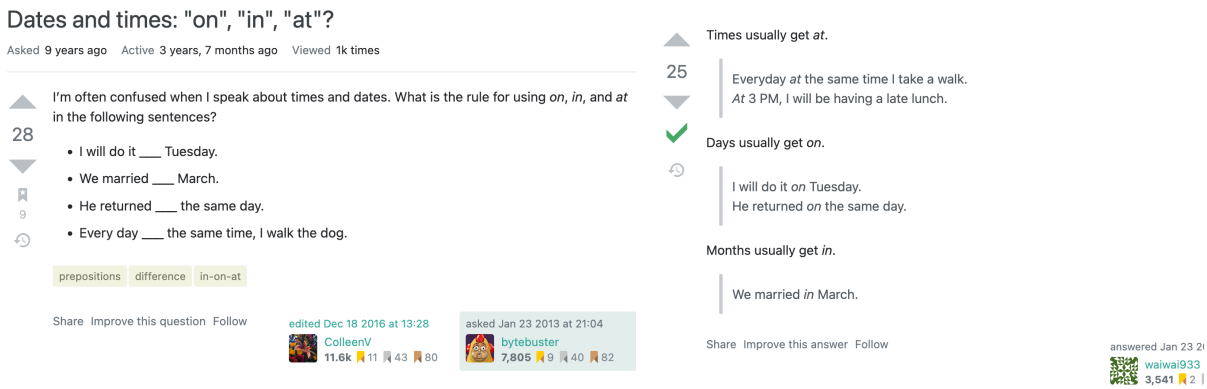
Notably, to our knowledge, there is no substantial progress in integrating LLMs with field linguists' or typologists' workflow. Most works listed above are from the pre-LLM era. One exception is the rather exciting exploration of the metalinguistic reasoning capabilities of LLMs focused on small, artificial problems inspired (or directly taken) from linguistic olympiads, such as the PuzzLing Machines (Şahin et al., 2020) and LingOly (Bean et al., 2024) benchmarks (see also §6). But all such problems with their guarantees of a single correct solution and carefully curated data to reach it barely mimic real-world settings, where incomplete and ambiguous data render the task significantly harder. The very recent work of Yang et al. (2025) that tests whether LLMs can be used to gloss unknown lexical items is one step in the above direction.[5]

4.3 Language Learning Domain

An important domain for metalanguage is the realm of language teaching and learning, whether in a classroom or in a less formal context such as an online forum or chatbot interaction. Complementing the extraction of symbolic rules for educational purposes described above (§4.2), it is valuable to investigate **natural** metalanguage scenarios in the language learning domain.

Here we highlight two studies of LLM processing of natural metalanguage (NML). Behzad et al. (2023) collected data from two online English discussion forums (one designed specifically for L2 English learners) in order to construct a benchmark for English Language Question Answering (ELQA). The metalinguistic questions in this dataset span a range of topics including vocabulary, grammar, and meaning; some, for example, inquire about sentence grammaticality, while others request help with expression or passage interpretation, or asked general questions about linguistic conventions. An example of a question and corresponding answer, both user-generated, appear in

[5] We note, although, that it still operates in a less noisy environment and with more available information as input than one might expect in real-world settings.



(a) Question **(b)** Answer

Figure 2: Screenshots of a page on the English Language Learner Stack Exchange site, which is included in the ELQA dataset (from Behzad et al., 2023). The source page is `https://ell.stackexchange.com/questions/12/dates-and-times-on-in-at`.

Figure 2: the question/answer pair mix a system-level overarching issue (about the rules for prepositions in times and dates) with specific exemplar instances.

Sampling questions and answers from ELQA, Behzad et al. (2023) conducted a human evaluation pitting user responses against responses from LLMs (including GPT-3 with few-shot learning or finetuning). Note that this kind of question answering is an NML-to-NML task. Overall, the best GPT-3 setup was highly fluent across the board, and gave accurate answers for many of the questions, but in some cases underperformed the highest-rated user answer for accuracy. This suggests that LLMs may be helpful interlocutors in answering learners' metalinguistic questions, but at times will make mistakes (with the caveat that today's state-of-the-art models have not yet been evaluated).

In a followup study, Behzad et al. (2024) addressed the *crosslinguistic* dimension of learner QA, speculating that L2 learners using an LLM may frame questions in their native language. This raises the question of whether the pairing of prompt-language and target language matters. This study used a controlled paradigm of minimal pair grammaticality judgment[6] (given an original and a corrected sentence from a grammatical error correction dataset) so that it would be possible to independently manipulate the language of the prompt template and the language of the target sentence. The languages tested were English, German, Russian, Ukrainian, and Korean. The tested models displayed a great deal of sensitivity to the choice of prompt-language, suggesting that the apparent multilinguality of an LLM does not guarantee stable metalinguistic behavior across languages.

[6]Other work has studied metalinguistic prompting for grammaticality judgments (e.g., Hu and Levy, 2023).

4.4 Legal Interpretation

First, a bit of background. **Legal interpretation** is the enterprise of determining the meaning of a rule expressed in natural language (Brannon, 2023). This frequently arises in judicial cases, where a judge must determine the extent of a category in order to decide whether it applies to the facts of the case. (If the rule is "no vehicles are allowed in the park", and "vehicle" is not specifically defined, should it be read to encompass skateboards? Wheelchairs? Ambulances? Are there contextual clues that shed light on the scope of the category?) The vehicle rule is a hypothetical example, but real cases similarly discuss the semantic interpretation of a term, such as the "landscaping" controversy summarized in Figure 3. Other cases implicate grammatical ambiguities in the wording of a statute or contract. Many U.S. judges subscribe to the philosophy that the starting point for interpreting legal language is the general language—they seek to determine the "ordinary meaning" of the text per contemporaneous usage (as members of the public might interpret it today, or when the law was enacted). This "textualist" perspective has engendered scholarly examinations of the nature of meaning in language, and the principles judges have articulated in an attempt to make language analysis rigorous and objective. But the practice of textualism has come under fire from scholars who contest the supposed neutrality of these principles—suggesting that they are poorly formulated, rooted

John is a contractor with insurance that covers property loss, damage, or personal injury claims that arise due to his 'landscaping' work.

John is employed by a family, the Smiths, to install an in-ground trampoline in the family's backyard. A few years after John completes the project, the Smiths successfully sue John for injuries that their daughter sustained while playing on the trampoline. John files a claim with his insurance company to recover losses incurred from the lawsuit.

Considering just how "landscaping" would be understood by ordinary speakers of English, is John covered by the insurance—yes or no?

Figure 3: A legal interpretation scenario represented as a QA task with binary questions. The example is based on the case *Snell v. United Specialty Insurance Co.* and constructed in the style of one of the prompting formats studied by Purushothama et al. (2026a).

in misconceptions about linguistic meaning, or so malleable that they can be manipulated to support any outcome (Eskridge et al., 2023).

Through collaborations with law professor Dr. Kevin Tobia—who has advocated for empirical approaches like survey research to ascertain ordinary meaning (e.g., Tobia, 2020, 2022; Waldon et al., 2025a)—Dr. Schneider and his lab are investigating computational linguistic and NLP tools for textualist inquiry. There are several threads of investigation.

Are LLMs trustworthy as tools for answering interpretive questions? Already judges have begun to entertain the possibility that difficult interpretive questions might be outsourced to LLM chatbots, on the rationale that huge amounts of ordinary usage in training data would entail accurate (perhaps superhuman) metalinguistic conclusions about meaning.[7] Waldon et al. (2025b) push back against this assumption, attributing it to myths about how LLMs work. Further experimentation by Purushothama et al. (2026a) and Petersen et al. (2026) examines prompt sensitivity and alignment with human consensus. (One of the tested prompt formats appears in Figure 3.) The results so far indicate that state-of-the-art platform models are less sensitive to prompt framing than smaller-scale models, and achieve some level of correlation with human judgments, but are not immune to giving an implausible answer when asked for a binary judgment. They are therefore *not* a silver bullet for resolving difficult questions. If they have any utility for interpretive reasoning, it is probably as a brainstorming aid: the system can be asked to generate arguments for and against a position, provided the human judge critically evaluates all claims (Waldon et al., 2025b).

How prevalent are different facets of metalanguage in judicial opinions? Kranzlein et al. (2024); Kranzlein (2024, ch. 5) examined this as a computational social science question by taxonomizing and tagging different facets of metalanguage in a corpus of Supreme Court opinions. An example of a metalinguistic sentence from one of these opinions (Breyer, 2023):

(8) First, the Act defines "pollutant" broadly, including in its definition, for example, any solid waste, incinerator residue, "heat," "discarded equipment,"' or sand (among many other things). §502(6), 86 Stat. 886.

Notably, sentence (8) includes several parts: **metalinguistic cue** words that denote linguistic units or processes ("defines", "definition"); a **focal term** being defined ("pollutant"); portions of a **definition** of the focal term; and a citation to a **legal source**. Kranzlein et al. (2024) envision metalanguage category tagging as an information extraction task, and annotate these categories in the CuRIAM corpus. (The full list of categories appears in Table 2.)

Kranzlein (2024, ch. 5) then trains a tagger for these categories: a **metalanguage identification** task. With automatic tagging, he conducts a content analysis of three decades of Supreme Court opinions. His analysis of metalanguage use over time points to an increase of some of the categories from 1986 to 2018, consistent with the growing popularity of textualism as an interpretive philosophy.

Do judicial canons of construction reflect accurate generalizations about linguistic usage? Over time, textualist judges have developed a suite of heuristics known as **canons of construction**. These assert preferences for resolving certain ambi-

[7] One judge writes: "models train on a mind-bogglingly enormous amount of raw data [across many genres]. Because they cast their nets so widely, LLMs can provide useful statistical predictions about how, in the main, ordinary people ordinarily use words and phrases in ordinary life" (Newsom, 2024). This fails to appreciate the distinction between learning implicit usage patterns, and being able to articulate those patterns metalinguistically in response to a prompt.

Category	Definition
Focal Term (FT)	Word or phrase used metalinguistically and/or whose meaning is under discussion.
Definition (D)	Succinct, reasonably self-contained description of what a word or phrase means. Need not be exhaustive. May also be negative—defining a word by what it's not.
Metalinguistic Cue (MC)	Word or short phrase cueing nearby metalanguage.
Direct Quote (DQ)	Span of text inside quotation marks.
Legal Source (LeS)	Citation or mention appealing to a legal document or authority.
Language Source (LaS)	Citation or mention appealing to an authority on language.
Named Interpretive Rule (NIR)	Mention of a well-established interpretive rule or test used to support an argument about the meaning of a word or phrase.
Example Use (ES)	Intuitive, quoted, or hypothetical examples that demonstrate a word/term can or cannot be used in a certain way.
Appeal to Meaning (ATM)	An explicit argument, implicit value judgment, or other statement indicating how one should go about interpreting meaning (e.g., by appealing to common sense, ordinary meaning, or the language of another statute).

Table 2: Categories of metalanguage annotated in the CuRIAM corpus (from Kranzlein et al., 2024).

guities in the text (Scalia and Garner, 2012; Brannon, 2025).[8] In response to calls for basing canons on stronger empirical foundations (Tobia et al., 2022), we have sought to critically examine the canons via computational linguistic techniques: namely corpus analysis (e.g., compiling judicial opinions referencing a particular canon in order to establish how it tends to be applied), treebanking (e.g., to establish the most frequent resolution of syntactic ambiguity in statutory text; Waldon et al., 2025c), and semantic annotation (Wells et al., 2025). These investigations are ongoing. Our hope is that they will lead to a clearer articulation of the canons (with precise terminology from linguistics), as well as empirical data about the reliability of each canon in practice.

5 Dichotomies in Metalanguage Research

Organizing metalanguage research, even when not taking a very broad view of metalanguage, requires considering multiple axes of analysis. We observe the following notable dichotomies (two of which are highlighted in Table 1):

- *system-level vs. instance-level metalanguage:* system-level metalanguage targets general properties of a linguistic system (e.g., grammatical rules, constructions, or typological facts), at the level of the entire language or subpopulation of the language community; instance-level metalanguage concerns specific linguistic tokens or contexts (e.g., explaining why a given sentence is ambiguous).
- *monolingual vs. multilingual:* the necessary LLM capabilities as well as system requirements and design would likely need to differ for metalinguistic inquiries targeting a single language, contrasting a pair of languages, or if the focus is specifically on second-language settings.
- *symbolic vs. natural metalanguage:* employing formalized notations such as parse trees, symbolic metalanguage enables precision but is less accessible to lay users than using ordinary (natural) language to describe linguistic phenomena.
- *processing vs. generation of metalanguage:* while processing metalanguage as input might largely evaluate models' comprehension, generation reveals whether models externalize linguistic reasoning in useful ways. Of course, many applications such as legal analysis and educational assistance require both capabilities.

By necessity, any metalanguage-related work will occupy a position along many of these axes, as do many of our works discussed in §4.

6 Research Directions

Research connecting metalanguage and LLMs asks how well LLMs fare on different kinds of metalinguistic tasks; why; and to what end. We list a few specific directions below. Some of these have already been subjects of inquiry, for which we give illustrative citations from our labs and others. Many

[8] The Nearest-Reasonable-Referent Canon, for example, makes recommendations about how to disambiguate the syntactic attachment of a modifier (Scalia and Garner, 2012).

directions though are, to the best of our knowledge, heretofore unexplored:[9]

Intrinsic evaluation questions Here we focus on research directions that aim to evaluate the metalinguistic capabilities of LLMs:

1. Can an LLM solve linguistic structure NLP tasks (e.g., Ettinger et al., 2023; Tian et al., 2024), analysis problems in theoretical linguistics (Beguš et al., 2025), or language puzzles (Rozner et al., 2021; Şahin et al., 2020; Bean et al., 2024; Chi et al., 2024; Sánchez et al., 2025; Choudhary et al., 2025)?
2. How well can models understand self-referential language ("This sentence is short.")? (Thrush et al., 2024)
3. How well can models distinguish mentions vs. uses?[10] (Kranzlein, 2024) (discussed in §4.4)
4. How does the choice of the (natural or formal) language in which metalanguage is formulated affect model behavior in relation to the described language? (Behzad et al., 2024) (discussed in §4.3)
5. Are LLMs sensitive to pragmatic phenomena like so-called *metalinguistic negation* (Horn, 1985),[11] where the speaker uses negation to signal disagreement with a choice of words, and quotation, where the speaker is not necessarily committing to the same perspective as the source they are quoting? (Gligorić et al., 2024, focusing on detecting whether hate speech and misinformation reflects the speaker's perspective)

Interpretability questions Next we outline inquiries that are paramount in order to understand *why and how* metalinguistic abilities arise in LLMs:

6. How well-calibrated is *explicit* metalinguistic output with respect to the system's *implicit* linguistic generalizations? (Hu and Levy, 2023; Song et al., 2025)
7. Can metalinguistic distinctions such as use vs. mention be traced to internal model representations?
8. How much of model behavior on metalinguistic tasks can be attributed to metalinguistic text in pretraining data or data provided at inference time (or not, as discussed, e.g., in Aycock et al. (2025) and Marmonier et al. (2025))?
9. To the extent that metalinguistic meaning is grounded in linguistic usage, is distributional learning from form alone (discussed, e.g., in Bender and Koller, 2020; Pavlick, 2023) fundamentally different from learning of non-metalinguistic meaning?

Extrinsic uses Finally, we delineate some underexplored uses of metalanguage for further downstream applications:

10. Can metalinguistic data such as syntax trees or grammar rules/descriptions be leveraged for inductive biases in pipelined systems (Wein and Schneider, 2024), integrated within LM architectures (Prange et al., 2022; Gessler and Schneider, 2023), or via in-context learning (Court and Elsner, 2024; Ginn and Palmer, 2025; Pei et al., 2025; Nakashole, 2026; Purushothama et al., 2026b)?
11. How well can a system perform metalinguistic question answering? (Behzad et al., 2023) (discussed in §4.3)
12. How can NLP shed light on how people use metalanguage? (Kranzlein et al., 2024) (discussed in §4.4)
13. Can we build LLM-powered assistants that deploy metalanguage effectively for language documentation, education, and scholarship? (also discussed in §4.1 and §4.2)

7 Conclusion

Metalanguage is inherently multifaceted. As we outlined in the possible dimensions in §5 above, it spans multiple levels of abstraction, heterogeneous representational formats, and diverse forms of linguistic reasoning. This diversity should be taken into consideration as we devote greater attention to metalinguistic tasks and applications.

We are particularly excited about the interpretability questions around metalanguage. Metalinguistic tasks may be particularly challenging where they require a model to *both* articulate and apply linguistic reasoning.

Metalanguage research also offers a promising pathway to studying learning and generalization. Humans frequently learn from explicit metalinguistic instructions and explanations and are able to

[9] Here we include work with transformer models like BERT and GPT-2, though our main focus in this paper is on contemporary prompt-based LLMs.

[10] *Mentioned language* when text refers explicitly to a linguistic entity like a word or sentence. A thorough definition is given by Wilson (2011, ch. 2), and a study of statistical classifiers is presented by Wilson (2013).

[11] Also known as *frame-rejecting negation*; an example is "John isn't being thrifty, he's just downright stingy" (Fillmore, 1985, p. 243).

apply this new knowledge to new examples. One should expect models to be able to do the same. Advancing research on how models learn from and operationalize metalanguage should dramatically improve the frontier of LLM abilities in general.

Limitations

We do not aim for this work to be a complete survey of metalanguage research. We by design draw heavily from our own work and our own perspective on the field.

Acknowledgments

We thank our collaborators on our metalinguistic journeys, including members of the NERT lab and the George Mason NLP lab, Dr. Amir Zeldes, and Dr. Kevin Tobia. We benefited from the Metalinguistic NLP Bibliography (https://github.com/nert-nlp/metalinguistic-nlp-bib) spearheaded by Abhishek Purushothama. This work was supported in part by NSF awards IIS-2144881 (Schneider) and IIS-2439202 (Anastasopoulos).

References

Tatsuya Aoyama and Nathan Schneider. 2022. Probeless probing of BERT's layer-wise linguistic knowledge with masked word prediction. In *Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 195–201, Hybrid: Seattle, Washington + Online.

Tjaša Arčon, Matej Klemen, Marko Robnik-Šikonja, and Kaja Dobrovoljc. 2026. Evaluating metalinguistic knowledge in large language models across the world's languages. arXiv:2602.02182 [cs].

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Simaan. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? In *International Conference on Learning Representations*, volume 2025, pages 12334–12357.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169, Singapore. Association for Computational Linguistics.

Andrew Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A., Ryan Chi, Scott A. Hale, and Hannah R. Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2025. Large linguistic models: investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, 6(12):3453–3467.

Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. ELQA: A corpus of metalinguistic questions and answers about English. In *Proc. of ACL*.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2024. To ask LLMs about English grammaticality, prompt them in a different language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15622–15634, Miami, Florida, USA. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proc. of ACL*, pages 5185–5198, Online.

Roger Berry. 2005. Making the most of metalanguage. *Language Awareness*, 14(1):3–20.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics.

Valerie C. Brannon. 2023. Statutory interpretation: theories, tools, and trends. Report R45153, Congressional Research Service.

Valerie C. Brannon. 2025. Canons of construction: a brief overview. In Focus IF12992, Congressional Research Service.

Stephen Breyer. 2023. *County of Maui v. Hawaii Wildlife Fund*. 140 S. Ct. 1462.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proc. of EMNLP*.

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2023. Teacher perception of automatically extracted grammar concepts for L2 language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3776–3793, Singapore. Association for Computational Linguistics.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is *wall* a *pared* and when a *muro*?: Extracting rules governing lexical selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6911–6929, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: a novel dataset for testing linguistic reasoning in language models. In *Proc. of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Mukund Choudhary, KV Aditya Srivatsa, Gaurja Aeron, Antara Raaghavi Bhattacharya, Dang Khoa Dang Dinh, Ikhlasul Akmal Hanif, Daria Kotova, Ekaterina Kochmar, and Monojit Choudhury. 2025. UNVEILING: What makes linguistics olympiad puzzles tricky for LLMs? In *Proc. of COLM*.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Hal Daumé, III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proc. of ACL*, pages 65–72, Prague, Czech Republic.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

William N. Eskridge, Brian G. Slocum, and Kevin Tobia. 2023. Textualism's defining moment. *Columbia Law Review*, 123(6):1611–1698.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "You are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Luke Gessler and Nathan Schneider. 2023. Syntactic inductive bias in transformer language models: especially helpful for low-resource languages? In *Proc. of CoNLL*, pages 238–253, Singapore.

Michael Ginn and Alexis Palmer. 2025. LLM dependency parsing with in-context rules. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 186–196, Vienna, Austria. Association for Computational Linguistics.

Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proc. of NAACL-HLT*, pages 5942–5959, Mexico City, Mexico.

Zellig S. Harris. 1951. *Methods in Structural Linguistics*. University of Chicago Press.

Laurence R. Horn. 1985. Metalinguistic negation and pragmatic ambiguity. *Language*, 61(1):121–174.

Kristen Howell, Emily M Bender, Michel Lockwood, Fei Xia, and Olga Zamaraeva. 2017. Inferring case systems from igt: Enriching the enrichment. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 67–75.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proc. of EMNLP*.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Hus and Antonios Anastasopoulos. 2026. A rag approach for typological database completion. In *Proceedings of the Eighth Workshop on Computational Research in Linguistic Typology*, Rabbat, Morocco. Association for Computational Linguistics.

Michael Kranzlein. 2024. *Unpacking Meaning with Natural Language Processing: Legal Metalanguage Analysis and Long-Tail Calibration*. Ph.D. dissertation, Georgetown University.

Michael Kranzlein, Nathan Schneider, and Kevin Tobia. 2024. CuRIAM: Corpus Re Interpretation and Metalanguage in U.S. Supreme Court Opinions. In *Proc. of LREC-COLING*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proc. of NAACL-HLT*, pages 1073–1094, Minneapolis, Minnesota.

Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit learning and the LLM in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31372–31422, Suzhou, China. Association for Computational Linguistics.

Ndapa Nakashole. 2026. Grammar as control: Modular language generation for the long tail. In *Proc. of ACL*, San Diego, California.

Kevin Newsom. 2024. Concurring opinion in *Snell v. United Specialty Insurance Co.* United States Court of Appeals For the Eleventh Circuit, 22-12581.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing RELIES on linguistics. *Computational Linguistics*, 51(3):1009–1032.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Dawson Petersen, Abhishek Purushothama, and Nathan Schneider. 2026. Sense and sensitivity: "reasoning" models are more robust, but can diverge from human consensus in a legal interpretation task. In *Proc. of CoNLL*, San Diego, California.

Jakob Prange, Nathan Schneider, and Lingpeng Kong. 2022. Linguistic frameworks go toe-to-toe at neuro-symbolic language modeling. In *Proc. of NAACL-HLT*, pages 4375–4391, Seattle, United States.

Abhishek Purushothama, Junghyun Min, Brandon Waldon, and Nathan Schneider. 2026a. Prompting from the bench: Large-scale pretraining is not sufficient to prepare LLMs for ordinary meaning analysis. In *Proc. of the Ninth Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, Montréal, Canada. arXiv preprint: 2510.25356 [cs].

Abhishek Purushothama, Emma Thronson, Alexia Guo, and Amir Zeldes. 2026b. Syntax as a Rosetta Stone: Universal Dependencies for in-context Coptic translation. In *Findings of ACL*. arXiv preprint: 2604.18758 [cs].

Josh Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: semantically complex wordplay puzzles as a target for NLP. In *Advances in Neural Information Processing Systems*, volume 34, pages 11409–11421.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: a challenge on learning from small data. In *Proc. of ACL*.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Arina Turkatenko, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2025. Linguini: A benchmark for language-agnostic linguistic reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Antonin Scalia and Bryan A. Garner. 2012. *Reading law: the interpretation of legal texts*. Thomson/West, St. Paul, MN.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. In *Proc. of COLM*.

Chihiro Taguchi, J Elizabeth Liebl, Antonios Anastasopoulos, David Chiang, and Géraldine Walther. 2025. Digital documentation for diasporic data: challenges, opportunities, and solutions for working with diaspora communities. In *9th International Conference on Language Documentation & Conservation (ICLDC)*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Learning Representations*, volume 2024, pages 18955–18985.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proc. of ACL*, pages 4593–4601, Florence, Italy.

Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, and Douwe Kiela. 2024. I am a strange dataset: metalinguistic tests for language models. In *Proc. of ACL*.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Large language models are no longer shallow parsers. In *Proc. of ACL*, pages 7131–7142, Bangkok, Thailand.

Kevin Tobia. 2022. Experimental jurisprudence. *The University of Chicago Law Review*, 89(3):735–802.

Kevin Tobia, Brian G. Slocum, and Victoria Nourse. 2022. Statutory Interpretation from the outside. *Columbia Law Review*, 122(1):213–330.

Kevin P. Tobia. 2020. Testing ordinary meaning. *Harvard Law Review*, 134(2):726–806.

Brandon Waldon, Cleo Condoravdi, James Pustejovsky, Nathan Schneider, and Kevin Tobia. 2025a. Reading law with linguistics: the statutory interpretation of artifact nouns. *Harvard Journal on Legislation*, 62(2):415–467.

Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes, and Kevin Tobia. 2025b. Large language models for legal interpretation? Don't take their word for it. *Georgetown Law Journal*, 114(1):115–183.

Brandon Waldon, Micaela Wells, Devika Tiwari, Meru Gopalan, and Nathan Schneider. 2025c. Legal-CGEL: Analyzing legal text in the CGELBank framework. In *Proc. of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 148–153, Ljubljana, Slovenia.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv:2211.00593 [cs].

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? Reducing translation effect using Abstract Meaning Representation. In *Proc. of EACL*, pages 753–765, St. Julian's, Malta.

Micaela Wells, Brandon Waldon, and Nathan Schneider. 2025. Scope ambiguity resolution of negated connectives in English corpora. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 47, page 6608.

Shomir Wilson. 2011. *A computational theory of the use-mention distinction in natural language*. Ph.D. dissertation, University of Maryland, College Park, Maryland.

Shomir Wilson. 2013. Toward automatic processing of English metalanguage. In *Proc. of IJCNLP*.

Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025. LingGym: How far are LLMs from thinking like field linguists? In *Proc. of EMNLP*, pages 1314–1340, Suzhou, China.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In *Proc. of the SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Harnessing the Latent Space: From Steering Vectors to Model Calibrators for Control and Trust

Nishant Subramani CMU
CMU: Carnegie Mellon University, Language Technologies Institute
nishant2@cs.cmu.edu

Abstract

Language models have changed from unreliable text generators to highly-capable large models with trillions of parameters. Capability increases come hand-in-hand with increases in scale, making understanding the internal representations of models more challenging. Since millions of users increasing rely on language models to interact with external tools or make decisions in medium or high-stakes scenarios, we need to establish control over model behavior and know when to trust model outputs. In this paper, we discuss our contributions on harnessing the latent spaces by proposing steering vectors for *control* and developing latent space-based model calibrators for *trust*. Together, our contributions help demystify the latent spaces of language models and offer new insights into how to harness model internals to build more trustworthy language technology.

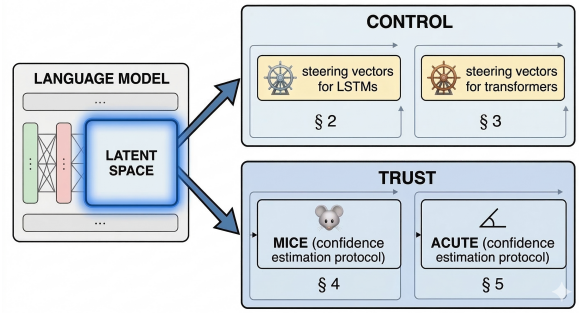


Figure 1: Our contributions on harnessing the latent spaces of language models: §2 and §3 focus on control, proposing steering vectors for the first time for LSTMs and transformer-based models. §4 and §5 focus on trust, building model-internal confidence estimators to assess confidence of language model output generations.

1 Introduction

Neural network language models (LMs) have evolved from small, unreliable text generators to very large models capable of solving complex reasoning tasks (Peters et al., 2018; Radford et al., 2019; Groeneveld et al., 2024; Yang et al., 2025; Team et al., 2025, *inter alia*). Despite the vast capability increases, analyzing the internal representations of trillion-parameter models is challenging. Due to this, the NLP community has increasingly treated models as black boxes, neglecting understanding the inner-workings of models. Even though large language models (LLMs) are scaled to millions of users, increasingly interact with external tools (Qu et al., 2024), and make decisions in medium and high-stakes scenarios (Thirunavukarasu et al., 2023), we rely by-and-large on simple behavioral observation (Hendrycks et al., 2021; Srivastava et al., 2023; Liang et al., 2023, *inter alia*). As a community, we must build fundamental understanding of the inner-workings of models and operationalize the internal representations of LLMs. We need to establish **control** over model behavior to ensure safety and alignment and establish confidence estimation mechanisms which can accurately adjudicate **trust**.

We present four threads of research aimed to demystify and harness the latent spaces of language models. To achieve model control, we show that LSTM-based language models can be minimally steered for exact generation (§2). We then adapt to transformer-based models in §3, showing both fine-grained and coarse-grained control via exact and concept-based steering. Shifting to trustworthiness, we build model-internal confidence estimators (MICE) to calibrate LLM generations in tool-calling scenarios (§4). Lastly, we broaden the framework to new model families and tasks by proposing activation-based confidence, utility, and trust estimators (ACUTE; §5). Together, our contributions offer actionable recipes to harness model internals to build more controllable, well-calibrated, and trustworthy language technologies.

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 119–130

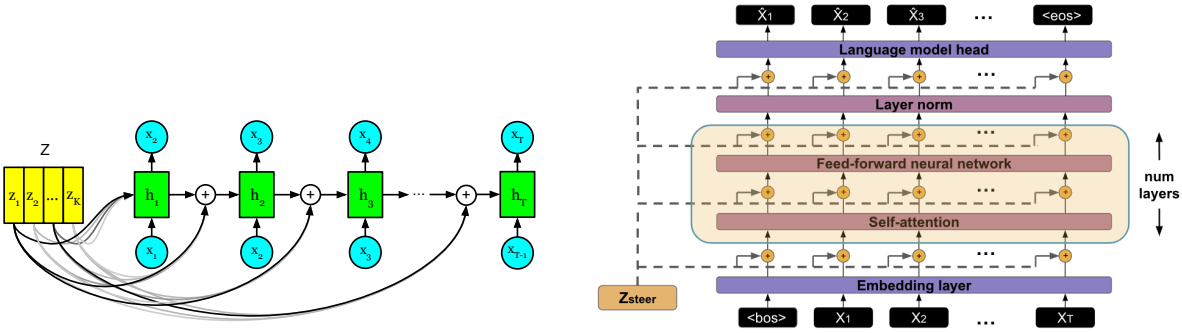


Figure 2: Here we show how the steering vector z_{steer} can be injected into an LSTM-based language model (left) and a transformer-based one (right). On the left, Z is shown to have a larger dimension than the model dimension. If dim(Z) equals the model dimension, $K = 1$, and thus there is just one vector z_1.

2 Control: Steering Vectors for LSTMs (Subramani et al., 2019)

We focus on control, specifically trying to answer one key question:

Key Question 1

Can LSTM-based language models be steered to generate a desired sequence exactly without updating a single parameter?

2.1 Prior Work

In 2018, the transformer architecture proposed in Vaswani et al. (2017) had yet to fully permeate the language model landscape and long short-term memory models (LSTMs; Hochreiter and Schmidhuber (1997)) were still the predominant architecture for language modeling. These language models were unreliable text generators. However, they have the potential to learn useful representations, so LMs started to be seen as general-purpose encoders (Dai and Le, 2015; Peters et al., 2018; Devlin et al., 2019, *inter alia*). Precise control of language model output, on the other hand, remained far out of reach, primarily due to the low quality of the underlying language models of the time.

2.2 Our Contributions

We explore whether language models could be used as *general-purpose decoders*, something that we now take for granted, but at the time was an unknown. For a pretrained language model to be used a general-purpose decoder, we need (1) to find a continuous-valued sentence representation (a steering vector) that can be fed into the frozen language model, (2) an encoder, likely task-specific, that can convert task inputs into steering vectors, and (3) for those steering vectors to *causally* generate the desired output. At the time, no work had shown this was possible, but now we take this for granted with advances in prompting and decoder-only LLMs. In our work, we explore the possibilities of this in LSTM-based models, before prompting became popular. Specifically, we ask whether LSTM-based models can be steered to generate a desired sequence exactly while keeping the underlying language model frozen.

Background To ask this, we first define the *sentence space* of a recurrent language model. Since the recurrent transition function $f_\theta = \mathbb{R}^d \times V \to \mathbb{R}^d$ defines a dynamical system based on the observations of tokens in a sequence. As a result, the language model embeds a sequence of length T as a $T + 1$ step trajectory in a d-dimensional space, where d is the dimension of the hidden state of the recurrent LM. Next, we parametrize the sentence space into a flat-vector space $\mathcal{Z} \in \mathbb{R}^{d'}$ to better understand the sentence space of the LM.

To map the trajectory of hidden states to a flat vector in $\mathcal{Z}$, we add a bias term $z_{steer} \in \mathcal{Z}$ to the previous hidden and cell state at each time step in the model and optimize z_{steer} to maximize the log-probability of a given sequence. Since we're adding z_{steer} to every hidden and cell state as well as at every timestep, information contained in z_{steer} will not degrade as quickly as if we just intervened at one location at one timestep. Using this formulation, we can go back and forth, from vectors to sequences and vice-versa, and thus design experiments to test whether a frozen model can be steered to generate any sequence of interest.

To map from sequences to steering vectors (forward estimation), we modify the recurrent transi-

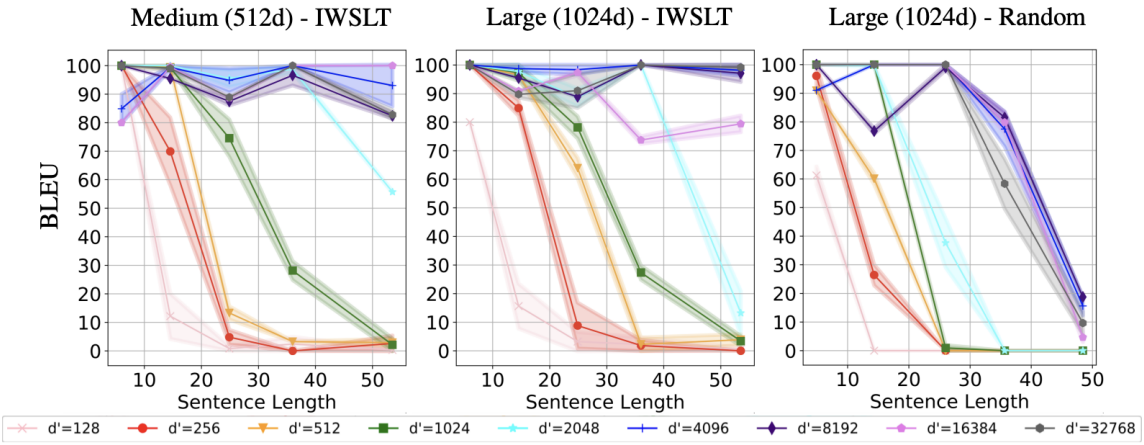


Figure 3: Recovery on IWSLT16 for `medium` (left) and `large` (center) and on random data for `large` (right).

tion function, see Figure 2 for details:

$$h_t = f_\theta(h_{t-1} + z_{steer}, x_t) \quad (1)$$

Here, we assume the dimension of z_{steer} is equal to the model dimension d^*. If this is not the case, we can up or down project z_{steer} without the addition of any parameters. See Subramani et al. (2019) for details. We then optimize z_{steer} to maximize the log-probability of the sequence as mentioned before via any off-the-shelf gradient-based optimization algorithm (*e.g.*, gradient descent, nonlinear conjugate descent, etc.).

To map steering vectors back to sequences (backward estimation), we intervene on a language model by injecting z_{steer} and using beam search to decode starting with a <bos> token. We stop when an <eos> token or 100 total tokens is reached. We measure how well the original sequence matches the generated sequence via three string overlap metrics: token-level exact match, BLEU score (Papineni et al., 2002), and longest prefix match.

Experimental Setup First, we train our own language models on 50M sentences from the English Gigaword corpus (Graff and Cieri, 2003), with a 879k sentence development set and a 878k sentence test set stratified by article publishing date. We use byte-pair encoding with 20,000 merges for a vocabulary of 20,234 subword tokens (Gage, 1994; Sennrich et al., 2016). Our model is a 2-layer language model with LSTM units of three sizes, `small` ($d = 256$), `medium` ($d = 512$), and `large` ($d = 1024$) with shared input and output embeddings (Press and Wolf, 2017), and dropout (Srivastava et al., 2014).

To train the model, we use stochastic gradient descent with Adam with a learning rate of 1e-4 and a batch size of 100 (Kingma and Ba, 2015). To learn steering vectors, we sample 100 randomly selected sentences from the development set as well as 50 sentences from the IWSLT16 En-De translation dataset to measure out-of-distribution generalization (Cettolo et al., 2016). We use nonlinear conjugate gradient (Wright and Nocedal, 1999) for optimization due to the highly non-convex nature of the objective function and use beam search with a width of 5 for backward estimation (Graves, 2012). See Subramani et al. (2019) for more details.

2.3 Takeaways

We can find steering vectors for every sequence that achieve near perfect recoverability (token-level exact match ≥ 0.99) on `large`, offering an avenue for direct causal control. Additionally, we find that larger, better trained models have higher recoverability and longer sequences are harder to recover. One key question is whether our forward estimation procedure operates like a naive compressor without any structure. In other words, does the forward estimation procedure have enough capacity to just encode the entire sequence in the vector without leveraging the language model's internal representations of language? To test this, we create a *random* dataset where we sample from the vocabulary with replacement at random where every token has equal probability. We learn steering vectors for these sequences as well as the out-of-domain IWSLT16 data and measure recoverability. In Figure 3, we show that sequences that are lower entropy under the language model (IWSLT data) are much easier to recover at similar sequence lengths as compared to sequences from the *random* data.

However, even for the very high entropy sequences (ones from *random*), z_{steer} has the capacity to encode sequences up to a token length of 28 nearly perfectly.

Limitations: Finding steering vectors was challenging. Optimization via conjugate gradient methods was very slow and could not be easily GPU-accelerated at the time, reducing adoption. Given how good recovery was for random sequences, a natural follow-up would be to understand what steering vectors encode and whether they could be useful beyond single sequence interventions. There is no guarantee that the steering vector space learned here offers any additional utility. This is precisely what we expand upon and explore in the next section.

2.4 Bigger Picture

At the time, being able to intervene on a language model with a single vector and causally force the model to generate *any* sequence of interest without updating a single parameter was highly surprising. This meant that language models had tremendous potential as universal decoders and steering could open up avenues to move away from task-specific finetuning and replace with inference-time steering. Our work could serve as justification to attempt natural language prompting on better trained, stronger models, which occurred in the years that followed. BERT had just recently come out (Devlin et al., 2019), and while this paper was under review, we learned that BERT rediscovered the classical NLP pipeline (Tenney et al., 2019), hinting that internal structure likely exists in transformer-based language models.

3 Control: Steering Vectors for Transformers (Subramani et al., 2022)

We expand upon control, generalizing steering vectors to transformer-based language models for both *exact steering* and *concept-based steering*. We answer the following questions:

Key Question 2

Can transformer-based language models be steered to generate a desired sequence exactly without any parameter updates?

Injection location	Timestep	BLEU-4
Embedding	all timesteps	33.99
Layer 6 (self attn)	**all timesteps**	**100.0**
Layer 6 (self attn)	**first timestep**	**99.80**
Layer 7 (feed fwd)	**all timesteps**	**100.0**
Layer 7 (feed fwd)	**first timestep**	**99.25**
All layers (self attn + feed fwd)	**all timesteps**	**100.0**
All layers (self attn + feed fwd)	first timestep	91.72
LM head	all timesteps	6.72

Table 1: Sentence recovery for steering vectors when injected into different layers of the transformer model and at multiple timesteps.

Key Question 3

Can extracted steering vectors act as useful representations with which we perform concept-based steering at inference-time?

3.1 Prior Work

Transformer language models started becoming popular, outperforming and largely replacing recurrent models (Devlin et al., 2019; Radford et al., 2019; Raffel et al., 2020). In §2, we showed that LSTM-based LMs could be precisely controlled for short sequences with steering vectors, opening up the potential for them to be used as universal or general-purpose decoders. Here, we explore whether higher-quality transformer-based language models could be more easily and efficiently steered, and thus make better universal decoder candidates. This work began prior to the release of GPT3 (Brown et al., 2020), hence the focus on small transformer-based models rather than LLMs.

3.2 Our Contributions

We coin the term *steering vector*. A vector $z_{steer} \in \mathbb{R}^d$ is a *steering vector* for a sequence x under a model M only if M exactly generates x via greedy decoding when z_{steer} is injected into M.[1]

Background We define a flat-vector space $\mathcal{Z} \in R^{d'}$ for a transformer language model, similar to the recurrent language model from §2. To map the trajectory of hidden states for a sequence

[1]Note that steering vectors need not correspond to an exact sequence. They are commonly now used to steer towards a desired concept or attribute.

$x_1, \ldots, x_T$, we add a bias term $z_{steer} \in \mathcal{Z}$ to the first-timestep at a single layer in the transformer stack after the feed-forward layer, see Figure 2 for details.[2] We also optimize z_{steer} to maximize the log-probability of a given sequence, giving us the ability to map sequences to steering vectors (and vice-versa) and measure recoverability, exactly like in LSTM-based models. This process is more efficient in transformers as compared to LSTMs because z_{steer} is only added at one layer and one timestep. We can up or down project z_{steer} if we want to control the capacity of the steering vector.

Experimental Setup We take the GPT2-117M model and learn steering vectors by sampling sequences from four different genres (movies, books, news, and wikipedia) and stratify them based on length for a total of 256 sequences. Recoverability is measured via BLEU score. We vary where (injection location) and when (injection timestep) to intervene with z_{steer}. For optimization during forward estimation, we use Adam with a learning rate of 1.0 and use greedy decoding to recover sequences. We measure the extent to which steering vectors can be used as representations and compare them with mean-pooled hidden states.

Lastly, we explore whether concept-based steering is possible. We first extract steering vectors for sequences for positive and negative sentiment respectively via the Yelp sentiment dataset (Shen et al., 2017). We propose difference-of-means (DiffMean) steering, which works as follows. First, we use vector arithmetic to take the mean of the positive sentiment steering vectors z_{pos} and the negative sentiment ones z_{neg}. This is then operationalized at inference-time. For example, to steer towards positive sentiment, you add a steering vector $z_{steer} = \lambda(z_{pos} - z_{neg})$ at the timestep and location that those steering vectors were extracted from.[3]

3.3 Takeaways

Our experiments show that fine-grained control via the exact steering of transformer-based language models is much easier and more efficient than the exact steering of LSTMs. Table 1 shows that nearly all sequences are perfectly recovered, even when adding z_{steer} at just the first timestep. As long as z_{steer} is injected in the transformer stack (after the embedding and before the final layer), recoverability remains nearly perfect. Linearly interpolating between steering vectors gives us a glimpse into what the steering vector space looks like. Decoding from these intermediate points reveals structure: the space seems relatively smooth with large clusters corresponding to each of the sequences being interpolated between and a smooth transition in both syntax and semantics when moving from one sequence to another. Cosine distances between steering vectors at middle layers reflect semantic similarity better than mean-pooled hidden states when measured on the semantic textual similarity benchmark (Cer et al., 2017), indicating that steering vectors may be better representations than the ones learned by the underlying language models.

Steering vectors provide coarse-grained control, too. Our experiments on unsupervised sentiment transfer via DiffMean steering on the Yelp sentiment dataset show that a single direction in latent space learned via these steering vectors can flip sentiment reliably. We show two examples in Table 2. For the first time, we show that concept steering at inference-time is possible.

Positive Input	the taste is excellent!
$+0.5 * (z_{neg} - z_{pos})$	the taste is excellent!
$+1.0 * (z_{neg} - z_{pos})$	the taste is excellent!
$+1.5 * (z_{neg} - z_{pos})$	the taste is bitter and bitter taste is bitter taste is bitter
$+2.0 * (z_{neg} - z_{pos})$	the taste is unpleasant.
Negative Input	**the desserts were very bland.**
$+0.5 * (z_{pos} - z_{neg})$	the desserts were very bland.
$+1.0 * (z_{pos} - z_{neg})$	the desserts were very bland.
$+1.5 * (z_{pos} - z_{neg})$	the desserts were very tasty.
$+2.0 * (z_{pos} - z_{neg})$	the desserts were very tasty.

Table 2: Concept steering for sentiment for a positive input sentence (top) and negative input sentence (bottom).

3.4 Bigger Picture

As language model quality started improving, control became an achievable goal. Two months after starting this project, GPT3 came out showing that large pretrained language models had the ability to, at inference-time, be few-shot prompted to solve different tasks. Our work could serve as further justification for more ambitious inference-time based control such as in-context learning, alignment, and persona-based steering. Over the past 5 years, language models became more performant with higher quality representations and concept-based steering took off, operating on the activation

[2] z_{steer} could be added anywhere in the transformer stack and repeated across layers or timesteps, but we found that adding it once at a single layer and first timestep was sufficient.

[3] $\lambda \in \mathbb{R}$ is a constant known as the steering strength.

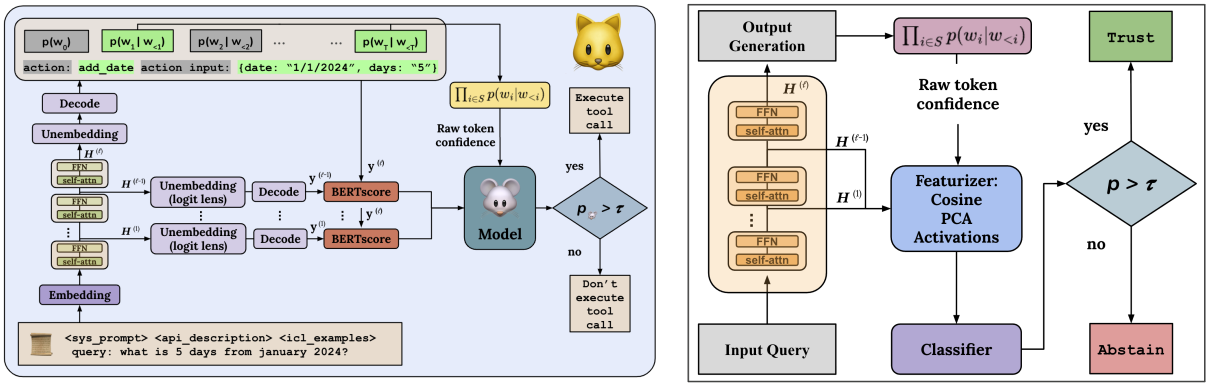


Figure 4: Here we show both the MICE (left) and ACUTE (right) systems to calibrate language model generations. The only major differences between the systems are the featurizers and classifiers used.

space rather than a reparametrized vector space like our steering vector space for inference-time control (Turner et al., 2023; Li et al., 2023; Dunefsky and Cohan, 2025; Arad et al., 2025; Bigelow et al., 2026; Morgulis and Hewitt, 2026; Wurgaft et al., 2026, *inter alia*).

4 Trust: MICE (Subramani et al., 2025a)

We tackle trust by leveraging model internals to try to answer a key question:

Key Question 4

Can we harness the latent spaces of language models to build better confidence estimators for tool-calling agents?

4.1 Prior Work

A confidence estimator is a model that estimates the probability that a different model's output is correct. Since language models have an internal confidence for its output already (*i.e.*, the joint probability of the generated sequence), auxiliary confidence estimators are rarely used. However, raw confidences of language models are known to be poorly calibrated (Desai and Durrett, 2020; Jiang et al., 2021; Zhong et al., 2023). To be well-calibrated, a confidence estimator must be correct approximately as often as it thinks it is (Dawid, 1982).[4]

Confidence estimation in NLP has been studied in tasks such as machine translation (Niculescu-Mizil and Caruana, 2005), semantic parsing (Stengel-Eskin and Van Durme, 2023), and long-form text generation (Band et al., 2024). Calibrating binary classifiers with a single input feature, the raw confidence, is common practice in machine learning with Platt scaling (Platt, 1999), isotonic regression (Barlow, 1972), beta calibration (Kull et al., 2017), and histogram regression estimators with adaptive binning (Nobel, 1996).

Our angle, harnessing the latent spaces of models to build confidence estimation mechanisms, is unique. In fact, there are very few studies that even combine mechanistic interpretability with confidence estimation or trust. Beigi et al. (2024) improves trustworthiness by using contrastive learning on activations and Liu et al. (2025) predicts correctness in question-answering tasks using probes learned on activations.

4.2 Our Contributions

Motivation: We explore whether we can leverage model internals to build a class of model internal confidence estimators (MICE) to better calibrate tool-calling agents. Taking inspiration from the early-exiting and intermediate layer decoding literature (*i.e.*, the observation that a language model can often be decoded from early layers; Geva et al. (2022); Schuster et al. (2022); Belrose et al. (2023); Elhoushi et al. (2024); Yom Din et al. (2024); Merullo et al. (2024)), we theorize that a prediction that is slowly refined through the layers ought to be more trustworthy than one that suddenly appears in the final layer.

MICE: Figure 4 shows MICE on the left. For a given input query q, we pass it through the language model and at each layer i, we use *logit lens* to decode from that intermediate layer (nostalgebraist, 2020), resulting in a candidate generation $y^{(i)}$. Then, we compare how close that intermediate generation is to the final output generation via

[4] Calibration is commonly measured using expected calibration error (ECE; Murphy and Epstein (1967); Naeini et al. (2015)) and Brier Score (Glenn, 1950).

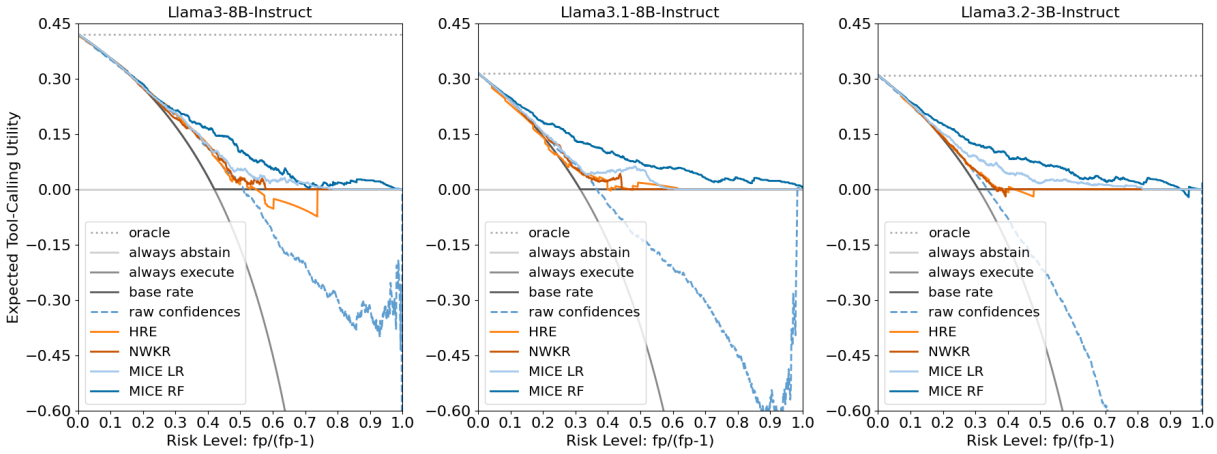


Figure 5: MICE systems outperform baselines on ETCU on the STE test set, especially at high-risk levels.

BERTscore (BERTscore($y^{(i)}, y^{(final)}$); Zhang et al. (2020)) using DeBERTa-xlarge-mnli (He et al., 2021), using those as features to train a correctness classifier. The probability that the classifier assigns to `<correct>` *is* the new calibrated confidence. In practice, we use both BERTscore features and the raw confidence (*i.e.*, the joint probability that the language model assigns to the output sequence) as features to train the classifier.[5]

Measuring Trust: We care about evaluating how *good* a confidence estimator is. This is commonly measured using ECE. However, ECE suffers from two major drawbacks:

1. Cannot distinguish between an oracle and a base rate estimator (*i.e.*, one that just predicts the base rate regardless of correctness)
2. Invariant to the risk level of the task

These drawbacks hinder decision-making utility and thus calibration is necessary but not sufficient for our purpose. We develop our own metric, expected tool-calling utility (ETCU) to solve these drawbacks, offering a better metric with which to evaluate the quality of a confidence estimator.[6] Additionally, we measure calibration error using a recently improved ECE variant called smooth ECE (smECE; Błasiok and Nakkiran (2024)).

Experimental Details: Our experiments use the simulated trial-and-error (STE) dataset, a synthetically generated tool-calling dataset consisting of English-language queries, which call 50 distinct APIs (Wang et al., 2024). We split the data into demonstration (used to construct few-shot examples), training, validation, and test sets consisting of 4250, 1500, 750, and 750 examples each. We 8-shot prompt three LLMs, Llama-3-8B-Instruct, Llama3.1-8B-Instruct, and Llama3.2-3B-Instruct and decode using greedy decoding to generate candidates (Grattafiori et al., 2024). Our MICE classifiers are trained using the training set, hyperparameters are tuned using the validation set, and performance is evaluated on the test set.[7]

4.3 Takeaways

Our experiments confirm that language models are poorly calibrated. Traditional post-hoc recalibration techniques such as histogram regression (HRE) and Nadaraya-Watson kernel regression (NWKR) tend to be highly conservative (Nadaraya, 1964; Watson, 1964), collapsing to base rate estimators with low calibration error, but low decision-making utility. MICE estimators, on the other hand, perform better, maintaining low calibration error, but higher decision-making utility due to a larger spread of probability estimates across examples. In Figure 5, we observe that for tasks with medium or high-risk, both HRE and NWKR remain conservative, always abstaining for all inputs. MICE performs better, correctly increasing its abstention rate as risk-levels increase, while trusting tool-calls some of the time. To quantify how well a confidence estimator does across risk-level settings, we develop an area-under-the-curve (AUC) metric called AUC-ETCU, following Marcum (1960). [8]

To test out-of-domain generalization for tool-

[5] We experiment with two classifiers: a random forest and a logistic regressor.

[6] See Subramani et al. (2025a) for details.

[7] A candidate generation is deemed to be correct if and only if it exactly matches the ground-truth answer.

[8] AUC style metrics are used in many areas of science (Wagner and Ayres, 1977; Geifman et al., 2019; Subramani et al., 2025b, *inter alia*).

	MMLU					APIGen					SCITLDR				
	(↓)	AUC-EURO (↑)				(↓)	AUC-EURO (↑)				(↓)	AUC-EURO (↑)			
estimator	**smECE**	low	med	high	all	**smECE**	low	med	high	all	**smECE**	low	med	high	all
Raw Conf	0.17	0.90	0.71	0.54	0.72	0.22	0.28	0.53	0.80	0.53	0.15	0.36	0.71	**0.92**	0.66
HRE	0.11	0.87	0.72	0.71	0.77	**0.02**	0.82	0.70	0.82	0.78	**0.08**	0.67	0.71	**0.92**	0.77
NWKR	**0.07**	0.90	0.73	0.74	0.79	**0.02**	0.82	0.69	0.82	0.78	**0.08**	0.67	0.71	**0.92**	0.77
ACUTE early act	**0.07**	**0.91**	0.73	0.74	0.79	0.05	0.90	0.82	0.87	0.86	**0.08**	0.69	**0.72**	**0.92**	**0.78**
ACUTE mid act	**0.07**	**0.91**	0.76	0.78	0.82	0.06	0.90	**0.84**	**0.88**	0.87	**0.08**	**0.70**	**0.72**	**0.92**	**0.78**
ACUTE late act	**0.07**	**0.91**	**0.77**	**0.80**	**0.83**	0.07	0.90	**0.84**	**0.88**	0.87	**0.08**	0.69	**0.72**	**0.92**	0.77
ACUTE cosine	0.09	0.90	0.75	0.78	0.81	0.03	0.87	0.77	0.84	0.83	**0.08**	0.68	0.71	**0.92**	0.77
ACUTE pca10	0.08	**0.91**	0.76	0.78	0.82	0.04	0.90	0.82	0.87	0.87	0.09	**0.70**	**0.72**	**0.92**	**0.78**
ACUTE pca20	0.08	**0.91**	0.76	0.77	0.81	0.06	**0.91**	**0.84**	**0.88**	**0.88**	0.09	**0.70**	**0.72**	**0.92**	**0.78**

Table 3: Results on the MMLU test set averaged across all 57 subtasks (left), on the APIGen test subset (middle), and on the SCITLDR dev set (right). All results for all tasks are averaged across the 6 LLMs we test. Lower smECE is better, while higher AUC-EURO is better. **Bold**, underline indicate the best and second best result respectively.

calling, we create a scenario in which new APIs are tested on. We hold out each of the 50 distinct APIs sequentially, resembling 50-fold cross validation and combine predictions across the entire test set. Despite being zero-shot, MICE performs comparably to post-hoc calibration baselines across both smECE and ETCU, while having a larger spread of probability estimates across samples.

Limitations: MICE relies on an auxiliary model to calculate BERTscore features, which can be very expensive. Additionally, we experiment with a single model family, Llama, on a single task, tool-calling. Lastly, ETCU makes one strong simplifying assumption, that a confidence estimator should get a reward of 0 for abstaining regardless of whether the candidate generation was correct or not. We address all of these limitations in §5.

5 Trust: ACUTE (Subramani et al., 2026)

We improve trustworthiness by addressing some limitations of MICE by asking:

Key Question 5

Can we harness the latent spaces of language models to efficiently build better confidence estimators for model generations across a variety of tasks including multiple-choice question answering, tool-calling, and scientific document summarization?

5.1 Our Contributions

ACUTE: We introduce an activation-based confidence, utility, and trust estimation protocol (ACUTE) to appropriately assess the confidence of a language model output. In Figure 4, we show how ACUTE works. First, we take the activations and pass them through a featurizer. Those features are fed into a classifier to predict whether the output generation is correct or not, exactly like MICE. Finally, the probability that the classifier assigns to correct is the new recalibrated confidence. ACUTE removes the reliance on the auxiliary BERTscore model for input featurization.

EURO: We improve upon the ETCU metric from §4 by removing the assumption that abstaining should always get 0 reward by introducing a new general and easily interpretable metric called expected utility renormalized by the oracle (EURO) that balances decision-making utility with calibration. EURO does not make simplifying assumptions, uses a single degree-of-freedom (the normalized net utility of correctly abstaining u_{ca}), and is bounded between an oracle estimator (EURO=1) and anti-oracle estimator (EURO=0). Calculating EURO at different u_{ca} or risk values traces a curve. Measuring the area under that curve provides a score with which to compare confidence estimators called AUC-EURO.[9]

Experimental Setup We apply ACUTE to six new LLMs on three new tasks: multiple-choice question answering (MMLU; Hendrycks et al. (2021)), tool-calling (APIGen; Liu et al. (2024)), and scientific document summarization (SCITLDR; Cachola et al. (2020)).[10] Performance is measured via smECE and AUC-EURO.

[9] See Subramani et al. (2026) for further details, including a detailed derivation in the Appendix of that paper.

[10] Results are averaged across LLMs.

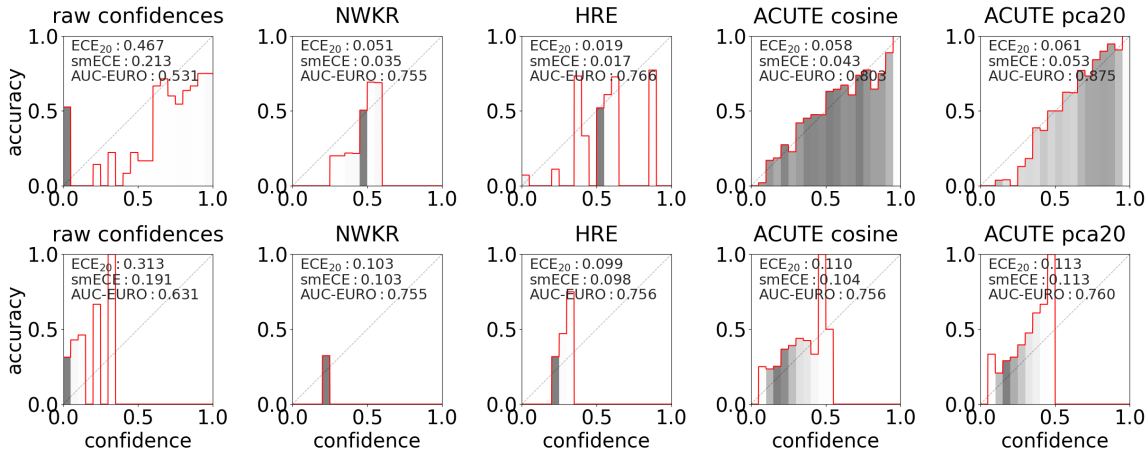


Figure 6: Reliability diagrams for the gemma-3-12b-it model for APIGen (top row) and SCITLDR (bottom row) across 3 baseline estimators and two ACUTE confidence estimators. Darker shading corresponds to higher density of examples in that confidence bin.

5.2 Takeaways

Table 3 and Figure 6 reveals that raw confidences remain poorly calibrated and post-hoc recalibration baselines collapse to base rate estimators as in §4. ACUTE performs well across all three tasks, outperforming baselines on AUC-EURO, while maintaining low smECE. Using activations from later layers (ACUTE late act) and PCA-reduction of all layer activations to 20 components (ACUTE pca20) performed best across all tasks.

5.3 Bigger Picture

Without any post-hoc calibration, language models provide terrible confidence estimates. Most post-hoc calibration methods collapse the often large spread of probability estimates into a much narrower range, reducing calibration error without increasing decision-making utility. Despite these drawbacks, probability estimates with or without post-hoc calibration remain the *de-facto* standard. Our work challenges this by proposing decision-making utility centric confidence estimators that can better adjudicate trust. Further study on this front can help us develop even better confidence estimators and increase the reliability and safety of LLMs. Overall, our work is a small step towards harnessing the latent spaces to appropriately assign trust to LLM outputs.

6 Conclusion

We present four research contributions that demonstrate how we can operationalize the latent spaces of language models for better control and trust. Our work introduces steering vectors for exact and concept-based control on both LSTM- and transformer-based models. Steering vectors provide nearly perfect fine-grained control at inference-time, suggesting that language models could be universal decoders. We propose two methods which harness the latent spaces of models to learn confidence estimators for language model generations to improve trust. Our methods recalibrate model outputs across architectures and tasks effectively, suggesting that the latent spaces contain information to build more reliable and trustworthy language technology, especially in high-stakes scenarios. We hope that our work encourages others to open up the black box and study the latent spaces of language models.

Acknowledgments

We thank the numerous collaborators and authors on each of the individual papers discussed here and both Mona Diab and Nivedita Suresh for feedback on an early version of this work.

References

Dana Arad, Aaron Mueller, and Yonatan Belinkov. 2025. SAEs are good for steering – if you select the right features. In *EMNLP*.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. In *ICML*.

Richard E Barlow. 1972. Statistical inference under order restrictions: The theory and application of isotonic regression. *(No Title)*.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. 2024. InternalInspector i^2: Robust confidence estimation in LLMs through internal states. In *EMNLP Findings*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor V. Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv*.

Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. 2026. Belief dynamics reveal the dual nature of in-context learning and activation steering. *Preprint*, arXiv:2511.00617.

Jarosław Błasiok and Preetum Nakkiran. 2024. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *ICLR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *EMNLP Findings*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Rolando Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign. In *Proceedings of the 13th International Conference on Spoken Language Translation*. International Workshop on Spoken Language Translation.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. *NeurIPS*.

A Philip Dawid. 1982. The well-calibrated bayesian. *Journal of the American statistical Association*, (379).

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Jacob Dunefsky and Arman Cohan. 2025. One-shot optimized steering vectors mediate safety-relevant behaviors in llms. In *COLM*.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. 2024. LayerSkip: Enabling early exit inference and self-speculative decoding. In *ACL*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. Bias-reduced uncertainty estimation for deep neural classifiers. In *ICLR*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *EMNLP*.

W Brier Glenn. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, (1).

David Graff and Christopher Cieri. 2003. English gigaword corpus. *Linguistic Data Consortium*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 22 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 22 others. 2024. OLMo: Accelerating the science of language models. In *ACL*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, (8).

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *TACL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, and 22 others. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*.

Jiarui Liu, Jivitesh Jain, Mona Diab, and Nishant Subramani. 2025. Llm microscope: What model internals reveal about answer correctness and context utilization. *arXiv*.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh RN, and 1 others. 2024. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *NeurIPS*.

J.I. Marcum. 1960. A statistical theory of target detection by pulsed radar. *IRE Transactions on Information Theory*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Language models implement simple Word2Vec-style vector arithmetic. In *NAACL*.

George Morgulis and John Hewitt. 2026. Subliminal steering: Stronger encoding of hidden signals. *arXiv*.

Allan H Murphy and Edward S Epstein. 1967. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, (5).

Elizbar A Nadaraya. 1964. On estimating regression. *Theory of Probability & Its Applications*, (1).

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 1.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *ICML*.

Andrew Nobel. 1996. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, (3).

nostalgebraist. 2020. Interpreting GPT: The logit lens. Blogpost.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, (3).

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Jirong Wen. 2024. Tool learning with large language models: a survey. *Frontiers of Computer Science*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *NeurIPS*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and T. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 22 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*. Featured Certification.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, (56).

Elias Stengel-Eskin and Benjamin Van Durme. 2023. Calibrated interpretation: Confidence estimation in semantic parsing. *TACL*.

Nishant Subramani, Samuel Bowman, and Kyunghyun Cho. 2019. Can unconditional language models recover arbitrary sentences? *NeurIPS*.

Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su, and Sam Thomson. 2025a. MICE for CATs: Model-internal confidence estimation for calibrating agents with tools. In *NAACL*.

Nishant Subramani, Alfredo Gomez, and Mona T. Diab. 2025b. SimBA: Simplifying benchmark analysis using performance matrices alone. In *EMNLP Findings*.

Nishant Subramani, Palash Goyal, Yiwen Song, Mani Malek, Yuan Xue, Tomas Pfister, and Hamid Palangi. 2026. The acute protocol: Operationalizing language model activations for better calibration, utility, and trust. In *ICML*.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *ACL Findings*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *ACL*.

Arun James Thirunavukarasu, Darren S. J. Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

John G. Wagner and James W. Ayres. 1977. Bioavailability assessment: Methods to estimate total area (auc 0-∞) and total amount excreted (ae∞) and importance of blood and urine sampling scheme with application to digoxin. *Journal of Pharmacokinetics and Biopharmaceutics*.

Boshi Wang, Hao Fang, Jason Eisner, Benjamin Van Durme, and Yu Su. 2024. LLMs in the imaginarium: Tool learning through simulated trial and error. In *ACL*.

Geoffrey S Watson. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*.

Stephen Wright and Jorge Nocedal. 1999. Numerical optimization. *Springer Science*, (67-68).

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. 2026. Manifold steering reveals the shared geometry of neural network representation and behavior. *Preprint*, arXiv:2605.05115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv*.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. Jump to conclusions: Short-cutting transformers with linear transformations. In *COLING*. ELRA and ICCL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jason Eisner. 2023. Non-programmers can label programs indirectly via active examples: A case study with text-to-SQL. In *EMNLP*.

Language Models as Measurement Apparatus for Culture

Kent K. Chang
School of Information
University of California, Berkeley
kentkchang@berkeley.edu

Abstract

Language models are increasingly used to quantify cultural phenomena, but what makes such measurement distinctively *cultural*? This paper argues that NLP work on culture is a *material-discursive practice*: the apparatus—model, data, annotation, evaluation—participates in constituting the cultural reality it measures, rather than passively recording it. Drawing on Karen Barad's concept of the *agential cut*—the contingent boundary between phenomenon and instrument—I show that the apparatus's substantive design choices draw such boundaries, and that the boundary is entangled from the start because language models have already internalized much of the cultural material they measure. I illustrate this through three case studies on television and film dialogue and two examinations of the apparatus itself: erasure of character names as cultural markers, and attunement to historically distant Restoration drama. This big picture analysis proposes a research program that is theory-driven, empirically rigorous, and culturally contingent, treating each agential cut as a conscious commitment.

1 Introduction

A growing body of natural language processing (NLP) research engages with cultural objects, often under the rubric of cultural analytics: literary texts, social media, and other artifacts whose significance is irreducible to information content (Piper, 2016, 2017) and constitutes a symbolic form (Cassirer, 2014). For instance, word embeddings have traced historical shifts in cultural concepts (Garg et al., 2018; Hamilton et al., 2016), which has been shown to be robust for humanistic inquiries (Zhou et al., 2025a): their contextualized variants have mapped the geometry of social meaning (Kozlowski et al., 2019; Lucy et al., 2022); connotation frames have measured implicit power and agency in film dialogue (Sap et al., 2017); computational sociolinguistics has modeled stylistic coordination in dialogue (Danescu-Niculescu-Mizil and Lee, 2011); and large language models have been probed for cultural knowledge (Chiu et al., 2025). In this light, this work leverages the affordance of NLP methods to address cultural questions, attending to both empirical rigor and interpretive depth such methods enable. What remains incomplete in this big picture is an explicit account of what it means to *measure* culture—as opposed to measuring sentiment, or syntax, or factual accuracy.

Recent work has begun to address this gap: Zhou et al. (2025b) deftly draws on sociocultural linguistics to argue that cultural NLP needs a coherent theory of culture grounded in indexicality, positionality, and emergence. Building on this, this paper offers a big picture analysis to expound on what it means to *measure* culture with language models, asking: What happens when a language model is used as an instrument of cultural measurement? I argue that NLP work on cultural objects constitutes a *material-discursive practice* (Barad, 2007; Brown and Duguid, 2000): the material configuration of the apparatus (model architecture, training data) and the discursive framework of the researcher (annotation categories, evaluation criteria, interpretive commitments) are entangled in the measurement and inseparable from it.

To make this concrete, I develop the concept of the *agential cut* (Barad, 2007) for NLP: the contingent boundary that an apparatus enacts between what counts as phenomenon and what counts as instrument. In using computational methods to study culture, every design choice—model architectures, taxonomies for classification, adjudication of annotation—draws such a boundary, and the boundary could always have been drawn differently. At the same time, what makes language models distinctive when applied to cultural artifacts is that they often have already encountered snippets (Chang et al., 2023b) or summaries of the cultural material they measure during pre-training:

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 131–143

the boundary between instrument and object is entangled from the start. As large language models (LLMs) are increasingly used for social and cultural measurements (Bamman et al., 2024; Halterman and Keith, 2025), this entanglement raises a problem that is prior to, and distinct from, the representational gap between data and cultural reality (Bode, 2020): while data remains a partial construction of the world, the LLM, the measuring instrument, has already internalized the very material it is asked to measure. The case studies in this paper return to this entanglement repeatedly.

I develop the argument through three case studies from my dissertation on dialogic interactions found in film and television, a site where social identities are constructed and contested. Each case study represents a type of cultural measurement: *structure* (conversation disentanglement), *interaction* (role attribution and gender), and *deviation* (stereotypic relation extraction). The measurements produced—gendered patterns in conversational agency, disparities in role attribution, the formalization of subversion—are constitutively contingent. In treating language models as measurement apparatus, I hope those case studies demonstrate how rigor and contingency are not in tension but mutually constitutive—and acknowledging this entanglement is the precondition for productive research at the intersection of NLP and culture.

2 Measurements of Culture

2.1 Operationalization and measurement

The dominant framework for computational work on culture derives from the social sciences, where *operationalization*—translating a theoretical concept into a measurable variable—sits at its core. Franco Moretti influentially imported this into digital humanities (Moretti, 2013), and subsequent work has refined the call for the computational study of culture to be explicit about its methodological commitments (Piper, 2017; Underwood, 2019). Two recent positions extend this concern to AI. Wallach et al. (2025) argue that evaluating generative AI systems is fundamentally a measurement challenge in which constructs such as helpfulness, fairness, or harm cannot be treated as natural labels but must be defined, instrumented, and validated.

Measurements of culture, in the sense developed below, take up the same problem fromyet a third angle. The question is not merely one of how we measure AI systems, but also how AI systems measure cultural artifacts. Plays, novels, films, and television scenes are not raw observations of social variable; they are organized by genre convention, historical context, audience reception, and critical interpretation. Computational methods can make such artifacts measurable at scale, but only by making them legible in some particular way: as character lines, reply graphs, role labels, relationship types, or deviations from a learned norm.

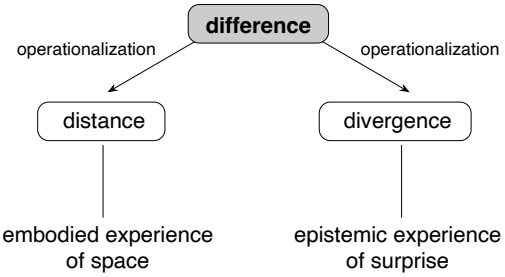


Figure 1: Two paths to operationalizing the concept of difference in computational work on culture (Chang and DeDeo, 2020): distance (a spatial metaphor, subject to metric axioms including symmetry) and divergence (a cognitive metaphor, capturing asymmetric relationships invisible to distance). The choice between them is itself an agential cut that determines which cultural relationships become measurable.

As Piper (2017) argues in his framework for literary modeling, operationalization is itself a reduction—and the reduction is where the measurement happens. Consider a simple example of operationalization—measuring how much two texts differ. Distance metrics like cosine afford spatial proximity arguments and inherit the symmetry of a metric space; divergences like Kullback–Leibler afford asymmetric arguments about encoding cost and surprise (Fig. 1; Chang and DeDeo, 2020). The two operationalize the same word—*difference*—into different cultural facts. The choice between them determines which cultural relationships are made measurable.

Operationalization is productive, but it carries an assumption: that the concept being operationalized exists independently of the measurement procedure. Recent work in cultural analytics questions this: McNulty and Chapot (2025) reconsider the relationship between computation and form in the wake of generative AI, treating models themselves—their architectures and outputs—as cultural-technical "forms" that both enable and require new modes of analysis. Dobson (2025) argues that architecture is not a neutral container but a substantive interpretive choice—a site where meaning is made and

Type	Apparatus configuration	Cultural question	Source of contingency
Structure	Multi-party dialogue → directed reply-to graph (Chang et al., 2023a)	How does conversational agency emerge from the **topology of who-responds-to-whom**?	architecture and input representation define what counts as a thread
Interaction	Audio-visual signal → Goffmanian role labels (Chang et al., 2026)	How are speaking and listening roles **distributed across participants**?	category taxonomy and modality selection determine which roles are measurable
Deviation	Dyadic dialogue → stereotypic relationship type (Chang et al., 2024)	Where does performed interaction **depart from normative expectation**?	norm is learned from training; deviation exists only relative to a trained baseline

Table 1: Three modes of cultural measurement, organized by the type of agential cut the apparatus enacts.

historicity registered. These observations suggest a framework that treats the entire configuration—model, data, annotation, evaluation—as a larger, coherent whole: indeed an *apparatus* that participates in producing its object, not one treating data, task, and algorithm as separable components of a linear research narrative.

2.2 Entanglement and cuts

The linear research narrative—task, data, model, metric—works for many NLP problems by treating these components as separable stages. For cultural measurement, that separation collapses from both sides. Decisions about what to model—how to represent a fictional character, what counts as an interaction, which categories to annotate, how much context to expose—are not preliminary to the measurement but constitutive of it; they determine which cultural realities can emerge and which are foreclosed.

Crucially, the instrument itself is culturally formed: what an LLM knows about culture is what circulates, and what circulates is structured by prestige and the cultural industries long before any researcher uses the model to measure. The problem is not that the model has this cultural past—for measurements of culture, some past is often necessary—but whether that past is acknowledged, tested, and interpreted as part of the apparatus. This is related to, but distinct from, the ontological gap between data and cultural reality that Bode (2020) identifies.

In machine learning terms, decisions about what to model are usually treated as task design, and the cultural formation of the instrument as contamination or memorization (Mallen et al., 2023) that puts pressure on model reliability. Those terms are useful, but too narrow: they treat the apparatus and its object as separable when, for cultural measurement, they are not. From Barad's (2007) reading of Niels Bohr, I take the concept of the *agential cut*: the boundary an apparatus enacts between what counts as phenomenon and what counts as instrument. I use the term here in a constrained methodological sense—not every implementation detail is an agential cut. Random seeds, batch sizes, choice of GPU vendor, logging verbosity: these do not, in any normal range, change what cultural phenomenon can appear in the measurement. A design choice becomes a cut when it does: the label taxonomy, the context window, the modality, the anonymization procedure, the training data, or the norm against which deviation is measured.

The case studies that follow each enact a different cut (summarized in Table 1). Conversation disentanglement cuts continuous dialogue into a directed reply-to graph (Chang et al., 2023a). Conversational role attribution cuts an audiovisual scene into speaker, addressee, and side-participant roles (Chang et al., 2026). Stereotypic relation extraction cuts a trained expectation into a norm against which performance can depart (Chang et al., 2024). Each measurement is empirical, but none is independent of the apparatus that produces it. Framed like this, cultural analytics seeks to hold together the positivist work (of building models that shed light on culture) with the critical work (of insisting on the contingency of every cultural question those models help us ask).

3 Measuring Structure

The first type of cultural measurement involves imposing a formal structure on multi-party dialogue. The agential cut here is *structural*: any conversational exchange can be formalized in multiple ways—as a sequence of turns, a tree of reply-to links, a network of topic threads—and each formalization draws a different boundary between what counts as "structure" and what is relegated to noise.

Conversation analysis has long studied the systematics of turn-taking (Sacks et al., 1974) and the collaborative work of speakers and hearers (Goodwin, 1981); choosing among formalizations is an interpretive commitment about which of those systematics the apparatus will be allowed to see. Choosing reply-to graphs makes conversational floor, address, and initiation visible at scale; it forecloses lexical cohesion, topical drift, and the slow buildup of mutual understanding that fluent dialogue depends on. The gender measurements that follow are visible only inside this cut—they would not survive a re-formalization that, for example, weighted topic continuity over reply structure.

3.1 Conversation disentanglement

The structural formalization I adopt here is *conversation disentanglement*: recovering the thread structure of interleaved multi-party dialogue, a task studied extensively in NLP on IRC chat logs (Elsner and Charniak, 2008; Kummerfeld et al., 2019; Jiang et al., 2018; Zhu et al., 2021). In Chang et al. (2023a), we extend this to scripted multi-party dialogue, developing a BERT-based model (Devlin et al., 2019) that predicts which prior utterance each line responds to, thereby recovering a latent thread structure. Formally, given a sequence of utterances $\{u_1, \ldots, u_n\}$, the model encodes each utterance contextually and scores candidate reply-to links:

$$P(\text{parent}(u_i) = u_j) \propto \exp\big(g(\boldsymbol{h}_i, \boldsymbol{h}_j)\big), \quad (1)$$

where $\boldsymbol{h}_i$ is the contextual representation of utterance u_i and g is a learned scoring function. The result is a directed graph—a thread structure—extracted from continuous dialogue.

In my framework, this is an apparatus-dependent measurement: the threads do not pre-exist in the script but are produced by the apparatus. Here, the start of a thread is grounded in observations in television studies: McKee (2016) argues that speech acts are driven by character need: "all talk responds to a need, engages a purpose, and performs an action." This is central to our annotation scheme, which itself is part of the apparatus: what counts as a "reply," whether breaks a thread or continues it—these are not neutral transcription decisions but interpretive commitments that shape the thread graph the model produces.

At the same time, the original work includes exhaustive experiments across architectures and inpuft representations—different encoders, different context windows, different representations of speaker identity—each constituting a different apparatus configuration applied to the same dialogue. These are design choices that, in addition to enabling comparison between methods, crucially reflect what we think constitutes a speaker that shapes the ensuing analysis. Each combination of annotation scheme, encoder, and input representation enacts a different agential cut on the same underlying dialogue—and the cultural patterns that emerge are inseparable from the apparatus that produced them.

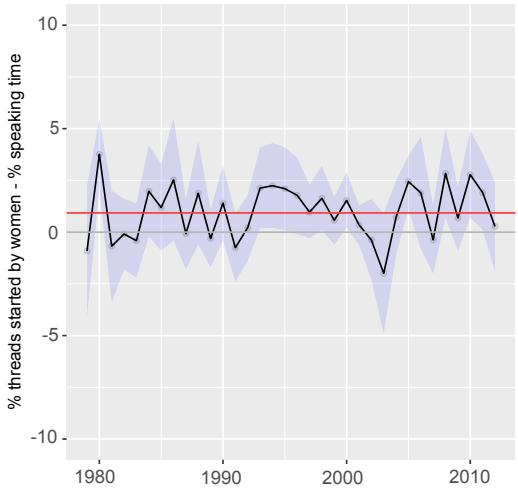


Figure 2: Percentage of conversational threads started by female characters minus their share of speaking time, by year, with 95% confidence intervals (shaded) (Chang et al., 2023a). The red line marks the overall average of $+1.0$ percentage points ($p < 0.05$): despite their under-representation, women initiate threads at a rate that slightly exceeds their speaking time.

3.2 Measurements of agency

Applied at scale to TV and movie transcripts, the resulting thread structures enable measurements of conversational agency: who initiates threads, who sustains them, and how these patterns distribute by gender. In our data from 808 movies, 30.4% of threads are started by female characters—consistent with the well-documented disparity in women's presence and voice in media (Rosen, 1973; hooks, bell, 1992). However, when normalized by each character's share of speaking time (Fig. 2), women initiate threads at a rate that slightly exceeds their speaking time, with an average absolute difference of $+1.0$ percentage points ($p < 0.05$). This finding is surprising: despite their under-representation, female characters are written to claim the conversational floor more often than

their male counterparts would predict.

These findings demonstrate what structural measurement—an affordance of NLP applied to cultural objects—makes possible: by committing to a particular formalization of conversational structure, the apparatus renders visible patterns of agency that would be invisible in the unstructured transcript. A viewer watching a scene experiences dialogue as a continuous flow; the apparatus transforms it into a graph whose structural properties—who initiates, who sustains, who is peripheral—can be counted, compared, and statistically tested.

4 Measuring Interaction

The study of how language use varies with interactional context has deep roots: Ervin-Tripp (1964) showed that speakers shift register depending on topic, setting, and listener; Ng and Bradac (1993) documented how verbal behavior both reflects and constitutes power asymmetries; and studies of scripted dialogue have long recognized that television writing encodes—and sometimes contests—these dynamics (Richardson, 2010; Bednarek, 2023).

To address such cultural questions, the second type of cultural measurement involves classifying participants into sociolinguistic categories—speaker, addressee, side-participant—that formalize who participates and how. The agential cut here is *categorical and modal*: the apparatus partitions a continuous audio-visual stream into a finite set of discrete role labels, and the choice of which roles to recognize—and which modalities to admit as evidence—determines what aspects of interactional positioning can register at all.

4.1 Conversational role attribution

Multimodal video understanding has produced large-scale datasets for television—most notably TVQA (Lei et al., 2018), which benchmarks compositional question answering over TV clips. But existing datasets treat dialogue as a source of answers rather than as an interactional system with its own structure. To address this gap, we operationalize the sociolinguistic organization of conversation itself to devise an annotation scheme, culminated in TV-MMPC (Chang et al., 2026).

The annotation scheme is itself an agential cut: Goffman's theorization of conversation participants, along with Clark and Carlson's (1982) taxonomy of speakers and hearers, is mapped onto discrete labels assignable to individual utterances in scripted television, so the model is tasked, for each utterance, to predict its speaker, intended addressee, and side-participants. This cut turns the continuous, multimodal flow of conversation into a discrete assignment of roles, and the choice of role categories is itself a discursive commitment: it reflects a interpretive commitment about which aspects of interactional positioning matter enough to measure.

In this particular case, the models evaluate a range of models: a text-only model discards everything non-verbal; a multimodal model takes in the full audio-visual signal and maps it into the same label space. These are different apparatuses measuring the same phenomenon, and they produce different results—not only in model performance, but in what counts as a relevant signal: which modalities the researcher selects, which frames the vision–language model samples, whether audio and video are processed jointly or separately, and what computing resources make feasible. For Barad, those are all agential cuts enacted by human and nonhuman actors shaping the apparatus, and consequently, the measurements it produces.

4.2 Measurements of gendered roles

Applied to 350,842 utterances across four TV series in TVQA, the apparatus produces measurements of how interactional roles distribute by gender. To quantify this, we fit a multinomial logistic regression that estimates the log-odds of occupying each role, controlling for show-level effects. With speaker as the reference category and $j \in \{\text{addressee, side-participant}\}$:

$$\log \frac{P(\text{role} = j)}{P(\text{role} = \text{speaker})} = \beta_{j,0} + \beta_{j,\text{female}} \cdot \mathbb{I}(\text{female}) + \sum_{s=1}^{S-1} \gamma_{j,s} \cdot \mathbb{I}(\text{show}_s), \quad (2)$$

This is itself a measurement: from the high-dimensional space of individual role attributions to a two-dimensional summary (one odds ratio per non-speaker role) that isolates the effect of gender. The resulting odds ratios—$\exp(\beta_{j,\text{f}})$—are 1.19 for addressee and 1.20 for side-participant (both $p < 0.001$), indicating that women are approximately 1.2 times as likely as men to appear in a listening role rather than as speaker, after controlling for show.

These measurements provide evidence of how social roles and power dynamics are constructed—and reinforced—in cultural artifacts. They are visible only because the apparatus includes the categories of addressee and side-participant: the theoretical commitment to distinguishing listening roles is what makes the gendered pattern measurable.

5 Measuring Deviation

The third type of cultural measurement involves a normative baseline and quantifying departures from it. Where structural measurement asks "what is the form?" and interaction measurement asks "who participates how?", deviation measurement asks "where does practice depart from expectation?" The agential cut here is *normative*: the apparatus learns a baseline of stereotypic patterns from training data, and what counts as "subversion" exists only relative to that learned norm. A different training corpus would yield a different baseline; a different label inventory—one that included `queer-platonic` or `enmeshed sibling`—would yield different deviations. What this cut makes measurable is stereotypicality and its breach; what it forecloses is meaning that neither stabilizes as a stereotype nor disrupts one—the in-between cases that read as simply ordinary.

Deviation, then, is a relational property: between texts, and between text and apparatus. This case study takes inspiration from cognitive stylistics, what Culpeper (2001) calls "stereotyping": the textual cues by which dramatic figures are constructed as social beings provide the signal, and the model's trained expectations provide the norm against which deviation becomes measurable.

5.1 Stereotypic relation extraction

In Chang et al. (2024), we train a Longformer (Beltagy et al., 2020) encoder on 787 digitized pilot teleplays to predict the social relationship enacted in a dyad's dialogue. Given a scene $\mathcal{S}$ with utterances from a head character c_h and tail character c_t, the model builds a joint representation. A Longformer encoder extracts the CLS token for each speaker's concatenated utterances. To incorporate scene context beyond the target speakers' words, an attentive pooling mechanism weights the hidden states of the full scene, guided by a token-level mask M that zeros out the target speakers' tokens:

$$\alpha = \mathrm{softmax}\big(\boldsymbol{w}_A{}^\top \boldsymbol{h}_\mathcal{S} \odot M\big), \tag{3}$$

$$\boldsymbol{h} = \big[\boldsymbol{e}^{c_h}_{\texttt{<s>}};\ \boldsymbol{e}^{c_t}_{\texttt{<s>}};\ \boldsymbol{h}_\mathcal{S}{}^\top \alpha\big], \tag{4}$$

where $M[j] = 0$ if token j is spoken by either target speaker and 1 otherwise, $\boldsymbol{h}_\mathcal{S}$ is the encoded scene, and $\boldsymbol{w}_A$ is a learned attention vector. The concatenated representation $\boldsymbol{h}$ is projected through a linear classification head:

$$p(r|\boldsymbol{h}) = \mathrm{softmax}(f(\boldsymbol{h})), \tag{5}$$

predicting a relationship type r from a fixed set (e.g., `colleague_of`, `sibling_of`, `spouse_of`).

The architecture enacts a specific agential cut. The mask M directs the model to attend to what *other* characters say and do in the scene—the ambient social context—rather than relying solely on the target dyad's words. This is a deliberate choice about what the apparatus should treat as signal: the broader conversational ecology, not just the dyad in isolation. Character names are anonymized to force the apparatus to operate on dialogic cues—*how* characters talk—rather than on memorized character-relationship associations (see §6 for the empirical consequences of this choice).

5.2 Measurements of subversion

In traditional NLP, model quality is measured by accuracy: how well predictions match ground-truth labels. This evaluation paradigm treats correct labels as the goal and errors as failures to be minimized. But for cultural measurement, this logic inverts—a lesson that working at the intersection of NLP and cultural analysis has made unavoidable. Metrics like accuracy are measurements of *performance*; what cultural analysis requires are measurements of *interpretive significance*.

The model in Chang et al. (2024) is deliberately trained as a "stereotyping reader": it learns what sibling dialogue, or spouse dialogue, *typically* sounds like across hundreds of teleplays. Rather than evaluating accuracy as an end in itself, the work measures *subversion* as the discrepancy between the model's predicted distribution over relationship types and the ground-truth labels. When a model trained on stereotypical patterns of sibling dialogue predicts that two brothers sound like a married couple, the "error" is the finding. The gap between prediction and reality is the very signal to be analyzed—a measurement of how interaction departs from the norm the apparatus has learned.

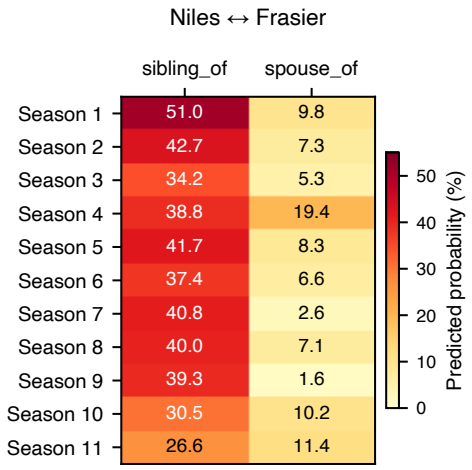


Figure 3: Percentage of predictions for `sibling_of` (ground truth) and `spouse_of` between Niles and Frasier Crane across *Frasier* (Chang et al., 2024).

Applied to *Frasier*, the apparatus identifies the Crane brothers' dialogue as consistently deviating from stereotypical sibling interaction. The model frequently predicts `spouse_of` for Frasier and Niles's exchanges—capturing, and formalizing, what queer theorists would recognize as a form of intimacy that exceeds its nominal category (Sedgwick, 2003; Halperin, 2002): the "crypto-gay" (Clum, 1999) quality that cultural critics have noted in their bickering, codependent, linguistically ornate relationship (Fig. 3). The apparatus produces a formal trace of the gap between performed relationship and normative type that can then become an object of cultural analysis. Butler's (1990) notion of the"subversion of identity" here becomes computationally tractable: notably, it is not meant to be an ontological claim about the characters' sexuality, but as a measurable discrepancy between dialogic performance and the model's trained expectations.

6 Between Measuring Instruments and Cultural Artifacts

The three case studies above focus on what NLP can measure when the cultural material may have been heavily represented in the model's training distribution (well-known TV characters and dialogues from popular series). In this section, we consider two complementary studies to further characterize what the apparatus is actually doing in those measurements: *erasure*, removing the apparatus's purchase on its memory of the cultural artifact in question to reveal what was being measured (§6.1); and *attunement*, deliberately training the apparatus on a corpus it has likely not absorbed, to see what measurement looks like when the apparatus likely has no memory of its object to draw on (§6.2). Together they describe the apparatus from both sides—what it brings to a culturally familiar text and what it has to build for a culturally distant one—and show where the boundary between the measuring instrument and the cultural artifact it measures is actually being drawn.

At issue in both cases is the apparatus's stake in the cultural material it is tasked to measure: what it has absorbed of those things, and what it has not. Brown and Duguid (2000) argue that knowledge cannot be separated from the social practices that produce it: information becomes knowledge only when embedded in a practice that gives it meaning. The apparatus's prior knowledge of, say, *Frasier* is not just stored in its weights but lives in the practice that mobilizes those weights to make a measurement about Frasier and Niles. Here I want to foreground the material side of the apparatus: In Barad's sense, the physical-computational infrastructure—parameters, training data, what the model has absorbed and what it has not. The discursive practice (which question is being asked, what counts as a finding, why these annotation categories) is equally constitutive of how any material result becomes a cultural measurement.

6.1 Erasure

The clearest demonstration of memorization shaping measurement comes from perturbing the apparatus. In the multimodal conversation structure study (§4), replacing character names with anonymous identifiers (e.g., replacing `Sheldon Cooper` to `Character C`) collapses speaker recognition from 78.6 to 13.7 and addressee recognition from 68.1 to 15.7 (Table 2, a). The same model, on the same data, appears to rely on its parametric memory for role attribution. What appeared to be a boundary between a neutral model and a conversational structure was a boundary between two layers of the model's own cultural knowledge: its general language competence and its specific memory of these characters, invoked by their names. Anonymization forces a different cut, one that separates dialogic structure from character identity, and the performance collapse reveals which cut was operative all along.

The same pattern recurs in stereotypic relation extraction (§5; Table 2, b). The Longformer model

Task	Orig.	Anon.	Δ
(a) Multimodal conversation structure			
Gemini 2.0 Flash			
Speaker (Acc)	78.6	13.7	−64.9
Addressee (F_1)	68.1	15.7	−52.5
(b) Stereotypic relation extraction			
Longformer			
Role (Acc)	34.8	24.8	−10.0
+ anon. training	36.7	33.8	−2.9
LLaMA 3-70b			
Role (Acc)	24.3	19.7	−4.6

Table 2: Effects of anonymization across two studies on different test sets: original ("Orig.") and anonymized ("Anon."). Panel (a): for multimodal conversation structure (Chang et al., 2026), speaker and addressee recognition collapse under anonymization; panel (b): for stereotypic relation extraction (Chang et al., 2024): the scene attentive pooling model without anonymized training shows a 10-point accuracy gap; training on anonymized data recovers most of the performance.

trained only on unanonymized data drops 10 points when evaluated on anonymized test data (from 34.8 to 24.8), indicating that the model exploits character names to infer memorized relationships rather than parsing the dialogue. Training on anonymized data recovers most of this gap, which forces the model to attend to how characters talk rather than who they are. LLaMA 3-70b (Grattafiori et al., 2024) zero-shot drops from 24.3 to 19.7 under anonymization, a smaller gap than Gemini's but the same pattern nonetheless. Prompt-based LLMs lack the Longformer's most direct mitigation (adaptation with anonymized data), though input-side anonymization can offer partial alternatives.

These anonymization experiments contextualize the cultural findings from the case studies: the measurements of gendered role attribution and relational subversion are shaped by the model's prior cultural formation. When Gemini attributes a speaking role or LLaMA predicts a character relation, the output reflects the cultural material internalized during training as much as the signal in the input. The material—what the model was trained on, what cultural knowledge its parameters encode—and the discursive—which categories the researcher defines, which tasks the model performs—are entangled in every measurement. This entanglement is at the heart of using language models as measurement apparatus for culture: the instrument might have already internalized some of the cultural material it measures, and the research narrative must account for it, rather than treat data, model, and evaluation as clearly separable stages of a pipeline.

6.2 Attunement

The anonymization analysis showed what the apparatus does when the cultural material may be available in training data. The inverse case is more revealing about what the apparatus *is* when there is no shortcut through prior knowledge, which I explore here in the context of Restoration comedy (1660–1700). Restoration drama is interesting here because it is well-studied in history and criticism but computationally under-explored, in part because the professionally curated and digitized editions sit behind proprietary access. The language is historically distant, and an off-the-shelf LLM, especially a smaller one, has at best a thin grasp of its conventions.

Restoration comedy of manners is driven by characters and archetypes (such as *rake* and *fop*). That said, we do not know *a priori* how many lines we need to have read for a character's archetype to become recognizable, which resembles a sufficient-context problem (López-Monroy et al., 2018) in which the reader does not recognize an archetype all at once but accumulates evidence as the play unfolds, committing when predictions have stabilized enough to act. This problem lets us specify two dimensions of the apparatus directly: what its parameters have been tuned to, and how it reads.

The corpus for this toy experiment is a random sample of 109 plays (1,283 character episodes) from Chadwyck-Healey English Drama collection;[1] each episode $e = (x_{1:T}, y)$ is an ordered sequence of one character's lines. Formally, a backbone f encodes prefix $x_{1:t}$ into a representation $\boldsymbol{h}_t = f_\theta(x_{1:t})$, from which a classifier predicts an archetype label c:

$$p_\theta(c \mid x_{1:t}) = \mathrm{softmax}\big(\boldsymbol{W}\boldsymbol{h}_t\big)_c, \quad c \in \mathcal{C}. \quad (6)$$

Drawing on taxonomies defined in Hirst (1979); Mast (1975), we include the following in $\mathcal{C}$: RAKE, FOP, NATURAL, OBSTACLE, to be predicted from a character's dialogue alone, absent any metadata. The source data has no archetype labels; we developed them iteratively—reading Restoration criticism, hand-annotating, refining prompts—and then scaled with Gemini 2.5 Flash (Gemini Team et al.,

[1] See Appendix A for more details.

2023), achieving Cohen's $\kappa = 0.71$ against the author's annotation of five plays.

To test this apparatus, we compare two reader-models on the same 1B-parameter Gemma backbone (Gemma Team et al., 2024)—a fixed-window reader (first N lines) and the entropy-thresholding reader together with a majority-class baseline. Entropy thresholding models the epistemic experience of the reader: the apparatus reads sequentially, stops when predictive entropy falls below δ, and predicts the maximum-probability class:

$$H_t = -\sum_{c \in \mathcal{Y}} p_\theta(c \mid x_{1:t}) \log p_\theta(c \mid x_{1:t}), \quad (7)$$

$$\tau_\delta(e) = \min\{t : H_t \leq \delta\}, \quad (8)$$

$$\hat{y}(e) = \arg\max_c p_\theta(c \mid x_{1:\tau(e)}). \quad (9)$$

For evaluation, we track macro-F_1 alongside an efficiency-aware objective $\mathcal{J} = \text{macro-F}_1 - \lambda \cdot \overline{\rho}$ ($\lambda = 0.25$ by default, $\rho(e) = \tau(e)/T$ denotes the fraction of lines consumed), which prices each unit of reading against the prediction it enables.

Attunement comes through one epoch of continued pre-training on roughly 174,000 lines of in-domain dialogue, in the spirit of historically attuned LMs (Manjavacas Arevalo and Fonteyn, 2021). The effect is consistent across reader-models: $\mathcal{J}$ rises from 0.17 to 0.29 for fixed-window and from 0.25 to 0.37 for entropy thresholding. The attuned entropy-thresholding reader reaches macro-F_1 0.44 reading only 26% of lines, against the fixed-window reader's 0.38 at 36% and a majority baseline of 0.14. Attunement gives the apparatus a thinner, deliberately-built version of the cultural past; the apparatus itself is layered: a human-calibrated annotator-LLM defines the norm against which an attuned reader-LLM is measured.

Taken together, §6.1 and §6.2 examine the same instrument from two angles. In the first, the apparatus's prior cultural formation is already there and can be perturbed; in the second, it has to be built up, and even built carefully it cannot reach beyond the cuts that defined it. In both cases, the measuring instrument and the cultural artifact are entangled in every result; neither produces it alone.

7 Conclusion

In this paper, I have argued that NLP work on cultural objects is a material-discursive practice in which the apparatus participates in producing the phenomena it measures. Three case studies (§3–§5) developed this along three dimensions of cultural measurement—*structure*, *interaction*, and *deviation*—and §6 examined the apparatus directly through *erasure* and *attunement*. A pattern emerges across both: each study is *theory-driven*. Conversation disentanglement draws on pragmatics and conversation analysis (Sacks et al., 1974; Goffman, 1963); interaction measurement on Goffman's (1981) participation framework; stereotypic relation extraction on cognitive stylistics (Culpeper, 2001), as well as queer theory's attention to the subversion of normative identity categories (Butler, 1990; Sedgwick, 2003); and attunement on dramatic criticism of comedy of manners (Hirst, 1979; Mast, 1975). Each measurement is also *culturally contingent*: the findings depend on the specific apparatus through which they were produced.

What sets the measurement of culture apart from other forms of task and evaluation is that the instrument itself is culturally situated: it carries the training distribution's biases, its era's textual archive, its architecture's affordances. The concepts of agential cut and material-discursive practice provide a framework for taking this reflexive entanglement seriously. This framework has implications for how cultural measurement should be practiced: Accuracy and model performance remain necessary, but they are insufficient. The anonymization experiments show that what a model gets *wrong*—evidence of memorization, deviation of stereotypical expectation—can be more productive than what it gets right. More fundamentally, optimizing for a task and interrogating the contingency of the measurement are complementary: the former establishes what the apparatus can do, the latter reveals what the apparatus is made of—and neither can be separated from the situated practice that produces it. This requires that agential cuts be conscious interpretive and ethical commitments.

In this light, cultural analytics is best understood as an interdisciplinary experiment that treats computational work as a material-discursive practice: it ceaselessly reflects on—and attempts to redefine—its positionality between positivist tradition and the negative movement of theory, while interrogating the reciprocal relations among humans, data, and information that undergird it, in order to shed new light on the worlds we inhabit. To measure culture with a language model is to turn the instrument on itself—and learning to do so deliberately, rigorously, and productively completes the big picture.

Acknowledgments

I thank my thesis advisor, Prof. David Bamman, for years of guidance on the work synthesized here, and my collaborators on the projects underlying this paper, named in the original publications, for the partnership that made each contribution possible. I am also grateful to the Stanford Literary Lab for sharing the Chadwyck-Healey English Drama source data used in §6.2, and to Prof. Mark Algee-Hewitt and Emil Wang for helpful discussions. I thank the anonymous reviewer for their feedback.

Most of all: the students of my Cultural Analytics class at UC Berkeley in Fall 2025, subtitled "Machine Learning and Measurements of Culture."[2] The ideas represented here emerged, died, transformed over the semester that we spent together. Teaching and learning with you was among the most generative experiences of my time at Berkeley. You complete the big picture.

References

David Bamman, Kent K Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. In *Proceedings of the Computational Humanities Research Conference 2024*, Aarhus, Denmark.

Karen Barad. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, Durham, NC.

Monika Bednarek. 2023. *Language and Characterisation in Television Series: A Corpus-informed Approach to the Construction of Social Identity in the Media*. John Benjamins Publishing Company.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv [cs.CL]*.

Katherine Bode. 2020. Why You Can't Model Away Bias. *Modern Language Quarterly*, 80(3).

John Seely Brown and Paul Duguid. 2000. *The Social Life of Information*. Harvard Business Review Press, Boston, MA.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, NY.

Ernst Cassirer. 2014. The concept of symbolic form in the construction of the human sciences. In *The Warburg Years (1919–1933): Essays on Language, Art, Myth, and Technology*. Yale University Press.

[2] https://ca.kentkc.org

Kent K. Chang, Danica Chen, and David Bamman. 2023a. Dramatic conversation disentanglement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4020–4046, Toronto, Canada. Association for Computational Linguistics.

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023b. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Kent K. Chang, Mackenzie Hanh Cramer, Anna Ho, Ti Ti Nguyen, Yilin Yuan, and David Bamman. 2026. Multimodal conversation structure understanding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7437–7458, Rabat, Morocco. Association for Computational Linguistics.

Kent K. Chang and Simon DeDeo. 2020. Divergence and the complexity of difference in text and culture. *Journal of Cultural Analytics*, 4(11):1–36.

Kent K. Chang, Anna Ho, and David Bamman. 2024. Subversive characters and stereotyping readers: Characterizing queer relationalities with dialogue-based relation extraction. In *Proceedings of the Computational Humanities Research Conference 2024*, Aarhus, Denmark.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Herbert H. Clark and Thomas B. Carlson. 1982. Hearers and speech acts. *Language*, 58(2).

John M Clum. 1999. *Something for the Boys*. St. Martin's Press, New York.

Jonathan Culpeper. 2001. *Language and Characterisation: People in Plays and Other Texts*. Longman.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

James E. Dobson. 2025. Beyond computational formalism or, architecture matters. *Journal of Cultural Analytics*, 10(3).

Micha Elsner and Eugene Charniak. 2008. You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Susan Ervin-Tripp. 1964. An analysis of the interaction of language, topic, and listener. *American anthropologist*, 66(6_PART2):86–102.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv [cs.CL]*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv [cs.CL]*.

Erving Goffman. 1963. *Behavior in Public Places*. The Free Press, New York.

Erving Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press, Philadelphia, PA.

Charles Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

David M Halperin. 2002. *How to Do the History of Homosexuality*. Univ. of Chicago Press, Chicago.

Andrew Halterman and Katherine A Keith. 2025. What is a protest anyway? codebook conceptualization is still a first-order concern in LLM-era classification. *arXiv [cs.CL]*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

David L. Hirst. 1979. *Comedy of Manners*. Methuen, London.

hooks, bell. 1992. *Black Looks: Race and Representation*. South End Press.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to Disentangle Interleaved Conversational Threads with a Siamese Hierarchical Network and Similarity Ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5).

Jonathan K Kummerfeld, Sai R Gouravajhala, Joseph J Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A Large-Scale Corpus for Conversation Disentanglement.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adrian Pastor López-Monroy, Fabio A. González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. 2018. Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1216–1225, New Orleans, Louisiana. Association for Computational Linguistics.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450–1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36.

Gerald Mast. 1975. *The Comic Mind: Comedy and the Movies*. Bobbs-Merrill, Indianapolis.

Robert McKee. 2016. *Dialogue: The art of verbal action for page, stage, and screen*. Hachette UK.

Tess McNulty and Laura Alice Chapot. 2025. Computation and form, reconsidered. *Journal of Cultural Analytics*, 10(3).

Franco Moretti. 2013. Operationalizing: or, the function of measurement in modern literary theory. *New Left Review*, 84:103–119.

Sik H Ng and James J Bradac. 1993. *Power in Language: Verbal Communication and Social Influence*. SAGE Publications.

Andrew Piper. 2016. There will be numbers. *Journal of cultural analytics*.

Andrew Piper. 2017. Think small: On literary modeling. *PMLA*, 132(3):651–658.

Kay Richardson. 2010. *Television Dramatic Dialogue: A Sociolinguistic Study*. Oxford University Press.

Marjorie Rosen. 1973. *Popcorn Venus; Women, Movies and the American Dream*. Coward, McCann and Geoghegan.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334. Association for Computational Linguistics.

Eve Kosofsky Sedgwick. 2003. *Touching Feeling*. Duke University Press.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago, IL.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*.

Kaitlyn Zhou, Haishan Gao, Sarah Li Chen, Dan Edelstein, Dan Jurafsky, and Chen Shani. 2025a. Rethinking word similarity: Semantic similarity through classification confusion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5803–5817, Albuquerque, New Mexico. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025b. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2021. Findings on Conversation Disentanglement. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 1–11, Online. Australasian Language Technology Association.

A Restoration Drama Corpus

The corpus underlying §6.2 draws on the Chadwyck-Healey English Drama collection, hand-transcribed by the original publisher with a bespoke XML schema and shared by the Stanford Literary Lab. The raw transcripts are processed through a two-pass agentic pipeline. In the first pass, Gemini reads each source file with an annotated view of its page-break tags and produces a structured manifest via a Pydantic schema: the manifest segments the file into plays by title, author, and page range, and recovers a cast map linking canonical character names to abbreviated speech tags. In the second pass, a LangGraph state machine walks each play page-by-page, maintaining the active play, the current speaker, and a short context window of previous lines; for each page, the model classifies lines as `SPEECH`, `STAGE_DIRECTION`, `ACT_HEADER`, `SCENE_HEADER`, `PROLOGUE`, `EPILOGUE`, or `PARATEXT`, normalizes speaker names against the cast map, and detects transitions between plays in multi-play source files.

The pipeline ran over 166 source files, producing 598 play-level segments. From these, 113 TSVs were retained for downstream analysis, 111 of which were readable, and 109 yielded at least one valid character episode under our preprocessing rules. The author first annotate a randomly sampled five plays, which informed the design of the pipeline described above, and then use Gemini 2.5 Flash to generate the character-episode archetype labels upstream. Cohen's κ between the Gemini and human labels was 0.71; broader human annotation remains future work.

An *episode* is the dialogue produced by one character in one play, kept in dramatic order, retained only if it contains ≥ 3 lines and has a majority archetype label among the four classes. Across 109 plays we obtain 1,283 episodes (median 49 lines per episode, maximum 1,110). Data is split by play, not by character, yielding 82 train, 11 development (for hyperparameter selection), 16 test plays, and 972, 116, 195 episodes, respectively.

The continued-pretraining data is constructed by concatenating each play's speaker-attributed lines in dramatic order, totaling 174,016 lines across 111 play-level documents. Continued pre-training uses causal next-token modelling on `google/gemma-3-1b-pt` for one epoch with block size 1024, learning rate 2×10^{-4}, per-device batch size 1, and gradient accumulation 16, without LoRA or 4-bit quantization. Because this stage uses unlabeled dialogue from the corpus as a whole, including held-out plays, it should be understood as unsupervised domain adaptation.

Memorisation Meets Compositionality in Natural Language Processing

Verna Dankers
University of Edinburgh
McGill University, Mila – Quebec AI Institute
vernadankers@gmail.com

Abstract

Memorisation in deep learning is undergoing a paradigm shift; it is increasingly recognised as a mechanism that can support, rather than hinder, generalisation. This is particularly relevant in NLP, where language combines compositional, generalisable structure with non-compositional expressions such as idioms, requiring memorisation from models and humans alike. My PhD work investigated memorisation in transformer models in generic terms, and through the lens of (non-)compositionality, from both data and model-internal perspectives. I analysed which training examples require memorisation, whether memorisation supports generalisation, and where memorisation occurs within model layers. I also studied how transformers process non-compositional idiom translations and how they balance compositional generalisation with non-compositional memorisation. Based on my findings, I stress that memorisation is an inherent part of learning *natural* language, can be beneficial, and is partially predictable. Yet it is not cleanly separable from generalisation, both at the level of data and of model parameters. Here, I summarise those findings and reflect on my PhD work.

1 Introduction

In deep learning, the perspective on memorisation of training examples is undergoing a paradigm shift. Previously linked to overfitting and poor generalisation, memorisation is now seen both as beneficial when it enhances deep neural networks' generalisation capabilities (e.g. Feldman, 2020; Feldman and Zhang, 2020; Zheng and Jiang, 2022) and as concerning when it involves examples that should not be memorised (e.g. Huang et al., 2022; Chang et al., 2023). This shift raises questions about how much models *can* even memorise, what they should or should not memorise, and how memorisation is implemented internally. Although these questions apply broadly, they are particularly relevant for language learning and NLP.

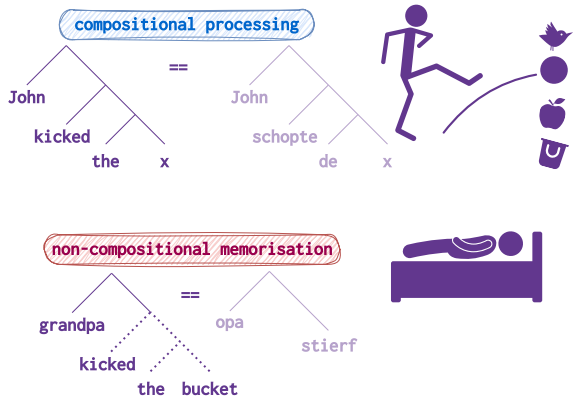


Figure 1: Illustration of the relation between compositionality and memorisation. "<person> kicked the <object>" is normally processed compositionally, yet "kicked the bucket" needs to be memorised as a single unit. This affects NLP tasks, such as translation.

Natural language itself requires both syntax-driven, generalisable meaning composition *and* memorisation, because it is simultaneously compositional and non-compositional – due to the prevalence of fixed formulaic expressions, among which proverbs, idioms and non-compositional noun compounds (Wray, 2002; Baggio, 2021). Many non-compositional expressions cannot be interpreted by composing the meanings of their parts, so they must be stored and retrieved holistically. Without such memorisation, humans and NLP models alike would default to literal readings (e.g. interpreting "grandpa kicked the bucket" literally rather than as "passed away" in Figure 1). The non-compositional side of language makes memorisation an essential complement to compositional processing in natural language understanding.

In my dissertation, I investigated memorisation in computational models of language, both as a

Home institute while conducting the PhD work.
Home institute at the time of submission.

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 144–159

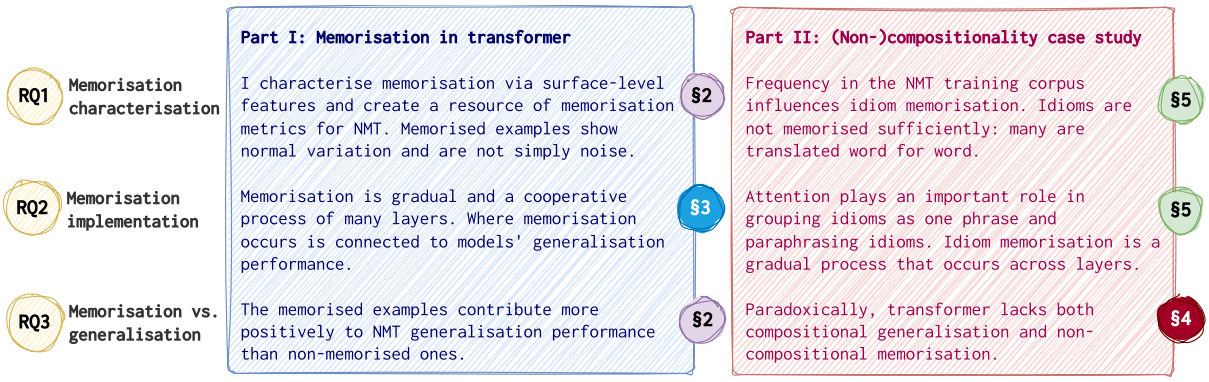


Figure 2: An overview of how the different sections address the three central research questions, both for memorisation in general (in blue) and for a (non-)compositionality case study (in red).

general phenomenon and through the lens of non-compositionality, examining the compositional vs non-compositional dichotomy as a memorisation–generalisation case study. In the process, both *neural machine translation* (NMT) and generic natural language understanding classification tasks were used. All analyses adopted the transformer architecture (Vaswani et al., 2017), albeit in different setups, varying training set sizes, model sizes and whether or not the model was pretrained. In this write-up presented to you here, I summarise my dissertation's work, highlighting findings and providing a retrospective.[1] This write-up is divided into two parts, each containing two sections with experiments and findings drawn from two previously published papers, together addressing the following research questions:

1. What characterises memorised examples? (§2 for generic memorisation patterns, §5 for idioms, specifically)
2. Which model-internal mechanisms enable memorisation? (§3 for memorised mislabelled examples, §5 for idioms)
3. To what extent are memorisation and generalisation at odds with one another? (§2 for how counterfactual memorisation benefits generalisation, §4 and §5 for the balance between compositional generalisation and non-compositional memorisation)

Lastly, §6 summarises the answers to these questions, which are also illustrated in Figure 2. Furthermore, Appendix A summarises lessons learnt during the PhD. Appendix B elaborates on PhD work that was not incorporated in the thesis.

[1]The dissertation extended the experiments conducted in the original papers. Findings that are the result of those added experiments are indicated in Sepia.

Part I
Memorisation in transformer

I will first focus on memorisation in generic terms: within a dataset, some examples require more memorisation than others. Which examples do models memorise, and where in these multi-layered networks can memorisation be localised?

2 A memorisation-generalisation continuum of data (Dankers et al., 2023)

When training neural networks, we aim for models to generalise rather than simply memorise training data. However, fitting natural language datasets inevitably requires memorising some of the data's idiosyncrasies (e.g. Feldman, 2020; Zheng and Jiang, 2022; Zhang et al., 2023). That memorisation is not always harmful, and can benefit generalisation had previously been established prior to the work discussed in this section, but primarily for artificial setups or classification tasks (Feldman and Zhang, 2020; Raunak et al., 2021; Zheng and Jiang, 2022).

Yet, is memorisation still beneficial in the *real-world*, noisy domain of NMT? Very little prior work has discussed memorisation in the context of NMT, with the exception of Raunak et al. (2021); Raunak and Menezes (2022). These works, however, focused on memorisation in a narrow sense, applied to a very small set of examples, and discussed it in relation to hallucinations. We, on the other hand, perform a very comprehensive analysis by constructing a multilingual resource of memorisation metrics, analysing which datapoint characteristics influence memorisation, and examining how memorisation relates to performance.

2.1 Experiments and findings

We treat memorisation as a graded phenomenon, quantified using the **counterfactual memorisation** (CM) metric (Feldman and Zhang, 2020):[2]

$$\texttt{CM}(x, y) = \underbrace{p_{\theta^{\text{IN}}}(y|x)}_{\text{training mem. (TM)}} - \underbrace{p_{\theta^{\text{OUT}}}(y|x)}_{\text{generalisation score (GS)}}$$

Here, θ^{IN} and θ^{OUT} represent models that have and have not seen (x, y) during training. The CM metric thus contrasts how a model performs on a training example to how a model *would have performed*, had the example not been in the training set; hence the 'counterfactual' nomenclature.

We compute an approximation of the TM, GS and CM metrics for 5M examples: 1M for 5 Indo-European language pairs (`En-Nl`, `-De`, `-It`, `-Fr`, `-Es`). We train 40 `transformer-base` (Vaswani et al., 2017) models from scratch per pair, on a parallel subcorpus constructed using OPUS data (Tiedemann and Thottingal, 2020). The 40 models have varying train and evaluation sets, such that TM, GS and CM can be computed for every example, averaging the quantities in the equation over outputs from multiple model instantiations. We put those 5M examples on a 'memorisation map', which we use to address the following sub-questions:

How do characteristics of datapoints relate to their position on the memorisation map? We compute 28 quantitative features and annotate a data subset manually using 7 additional features. We discuss how features such as source-target similarity, input and output length, token frequency and tokens' segmentation relate to the memorisation map. Figure 3 illustrates some key takeaways for different areas of the memorisation map, among which:

- As source–target overlap decreases, examples move down the diagonal. As a result, examples in the top right are near word-for-word translations, and misaligned examples are in the bottom left: truly misaligned examples are – even within a training regime with 1M examples – not memorised during training;
- Paraphrased and slightly inaccurate examples (in blue) can look alike according to a model;
- Examples with high CM scores tend to contain more infrequent words, be longer, and have higher BPE segmentation rates.

[2]This is a simplified representation of the metric. In practice, we replace probability with a geometric mean over target tokens' probabilities, to reduce length bias, and compute $p_{\theta^{\text{IN}}}$ and $p_{\theta^{\text{OUT}}}$ by averaging over results from various models.

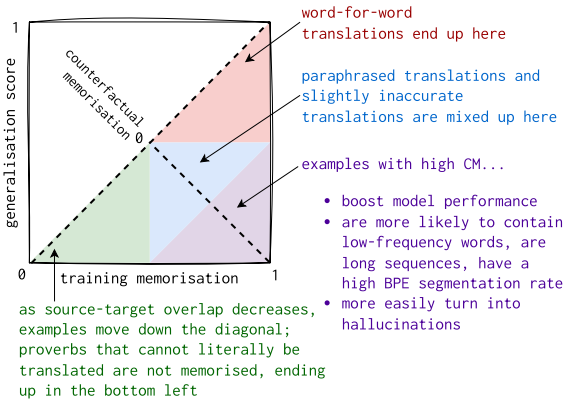


Figure 3: Illustrative summary of findings for different areas of the memorisation map. Counterfactual memorisation subtracts the y-coordinate from the x-coordinate. We detail some takeaways for areas of the map with high CM, low TM and low GS, high TM and high GS, and a similar TM and GS.

How do datapoints containing formulaic phrases stand out on the memorisation map? Although most experiments conducted here pertain to all training examples, we include an intermezzo to examine memorisation scores of source sequences that contain proverbs, idioms or non-compositional compounds. We would want these examples to be memorised *more*, yet, find that they are memorised *less*, emphasising that what is actually memorised is not necessarily what should be memorised.

Can we approximate memorisation metrics using datapoints' characteristics? Next, we use datapoints' characteristics to predict memorisation values with small regression models, allowing us to compare different language pairs and understand whether resource-intensive memorisation computation has cheaper approximates. We find that the regression models generalise cross-lingually, since characteristics' relation to memorisation is largely language-independent for our language pairs.

How does training on examples from different regions of the memorisation map change models' performance? Finally, we relate different parts of the map to the quality of NMT systems in terms of BLEU, COMET, targets' log-probability and models' hallucination tendency. Our results confirm that even in this real-world task, examples with high CM are most beneficial to model performance.

2.2 Conclusion and retrospective

En résumé, we contribute a valuable resource of memorisation scores, establish that memorisation

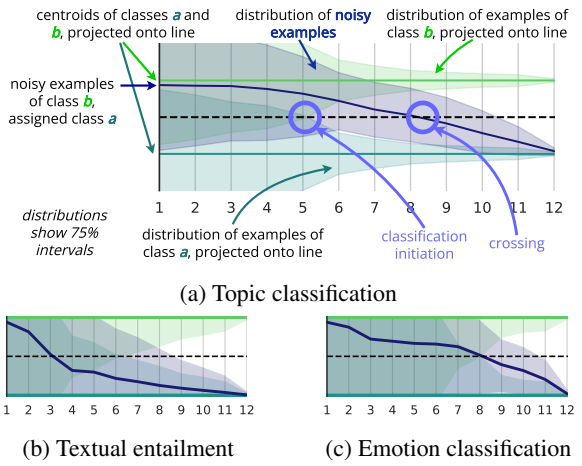
is not a mysterious phenomenon but a process that is predictable based on the features of data points and can positively benefit generalisation. While a valuable contribution, our work also had several limitations, among which were the computational expense of the experimental setup, and the focus on a set of five Indo-European language pairs. This was the consequence of a specific experimental design choice (parallel corpora across languages), which limits the generalisability of the findings.

Since the ideation of this project in 2022, the role of *large language models* (LLMs) in NMT has increased substantially. Because we train models from scratch – the default for NMT until ~2024 – it cannot assess how memorisation behaves during LLM fine-tuning or how pretraining might affect memorisation. Despite this, we consider the work a valuable contribution to the still-small body of research on CM in NLP (Raunak et al., 2021; Zheng and Jiang, 2022; Zhang et al., 2023). It was only the second to study CM at the scale of millions of examples (Zhang et al., 2023). Subsequent work has further examined memorised sequences in LLMs, echoing some of our findings (Prashanth et al., 2024). We hope that our per-datum memorisation estimates may serve as a benchmark in the future, for instance, benefiting the development of proxy metrics for CM.

It is also worth noting that, although this section takes the stance of CM being a generalisation benefit, not all types of memorisation are beneficial; memorisation is not a monolithic concept. In fact, in follow-up work (Dankers and Raunak, 2025), we show that knowledge distillation in NMT can increase extractive memorisation (a detrimental type of memorisation) and lead to more hallucinations.

3 Layer-based memorisation localisation (Dankers and Titov, 2024)

The previous section approached memorisation as a gradual phenomenon. Here, we instead focus on extreme memorisation of mislabelled examples to localise memorisation in LMs' layers. In *computer vision* (CV), studies tracing memorised, mislabelled examples often argue that lower layers learn generalisable features while deeper layers memorise (Cohen et al., 2018; Ansuini et al., 2019; Stephenson et al., 2021, i.a.). We refer to this as the *generalisation-first, memorisation-second* (GFMS) hypothesis. In NLP, localisation studies reach mixed conclusions when studying the memorisation of facts, idioms or verbatim sequences, pointing to the top (e.g. Dai et al., 2022; Zhao et al., 2024), early/middle (e.g. Meng et al., 2022), or lowest layers (e.g. Haviv et al., 2023; Stoehr et al., 2024). These conflicting findings may stem from varying experimental setups, and studying various memorisation types. Instead, we study the memorisation of mislabelled examples in NLP classification models to more directly compare with work from CV, putting the GFMS hypothesis to the test.



(a) Topic classification

(b) Textual entailment

(c) Emotion classification

Figure 4: Memorisation in action: an illustration of how memorised, mislabelled (noisy) examples move away from the centroid of their real class to their new class. Memorisation is gradual and non-local, but does shift to deeper layers for some tasks, such as topic classification.

3.1 Experiments and findings

We finetune four LMs (Devlin et al., 2019; Black et al., 2021; Zhang et al., 2022; Biderman et al., 2023) with a newly learned classification head across twelve datasets covering topic classification, sentiment analysis, hate speech detection, and generic NLU tasks (e.g. recognising textual entailment). We mislabel 15% of the examples and use them to study layer-based memorisation localisation, addressing the following sub-questions:

Can memorisation of mislabelled examples be localised to individual layers? Using four localisation methods – layer swapping, layer retraining, probing, and forgetting gradient analysis – we find that memorisation is not confined to single layers. Instead, multiple layers gradually shift mislabelled examples toward their assigned class, which is a process in which earlier layers are, relatively speaking, more important than later layers. To interpret this process, we introduce **centroid analysis** (Figure 4), which visualises how hidden representations of mislabelled examples change across layers.

How consistent is layer-based localisation across LMs and tasks? We observe subtle task differences linked to generalisation performance: when models generalise better, deeper layers contribute more to memorisation. Figure 4 illustrates this contrast. For instance, when comparing recognising textual entailment to topic classification, the latter generalises much better to test data than the former, while also having a 'crossing' of mislabelled examples in deeper layers in Figure 4. Comparing models with 12 and 24 layers shows that the lowest layers are not necessarily the most important in absolute terms, lowest is relative both with respect to model size and how 'deep' finetuning managed to change the pretrained model.

3.2 Conclusion and retrospective

Summarising, we localised memorisation by tracing mislabelled examples across layers. Memorisation was not confined to specific layers but emerged through cooperation across many layers, indicating that memorisation and generalisation are intertwined. However, layers contribute unequally: early layers play a larger role, as memorised examples begin to diverge there. These findings contradict the GFMS hypothesis. Because memorisation is distributed, local weight manipulations through editing or unlearning may alter behaviour without fully removing stored information.

Our work has several limitations. Mislabelled examples are only a proxy for real-world memorisation, and the localisation methods used are imperfect. Nonetheless, the agreement observed across LMs and methods suggests our conclusions are reliable. From a 2026 perspective, another limitation is our focus on 'traditional' fine-tuning with a task classification head. Yet, since the publication of our work, related studies have similarly argued that memorisation is distributed and intertwined with general language modelling abilities (Huang et al., 2024; Menta et al., 2025), suggesting our findings may extend beyond the specific setup studied.

Part II
(Non-)compositionality: a memorisation-generalisation case study

Let us now shift focus to (non-)compositionality. How does this reflect the tension between memorisation and generalisation, and how does memorisation of idioms affect models internally?

4 Evaluating (non-)compositional generalisation (Dankers et al., 2022a)

Compositionality plays an essential role in human language understanding, but whether neural networks exhibit this property has long been debated (e.g. Fodor and Pylyshyn, 1988; Smolensky, 1990; Marcus, 2003; Nefdt, 2020). Prior to 2022, studies of compositionality in NLP models mainly relied on synthetic datasets with simplified languages, where compositionality can be isolated and controlled (e.g. Lake and Baroni, 2018; Keysers et al., 2019; Hupkes et al., 2020; Kim and Linzen, 2020). These tests compute interpretations using a *local*, bottom-up notion of compositionality, ignoring that natural language contains exceptions such as idioms (see §1), which require more global sentence processing. For *natural* language, NLP systems must balance compositional and non-compositional processing. In this section, we analyse NMT outputs to explore this tension, contrasting compositional generalisation tests with the memorisation of non-compositional idioms. Before Dankers et al. (2022a), no datasets evaluated compositional generalisation in MT for models trained on natural language. We introduced new data to fill this gap and reformulated three theoretically grounded tests from Hupkes et al. (2020): systematicity, substitutivity, and overgeneralisation.

4.1 Experiments and findings

We train `transformer-base` (Vaswani et al., 2017) on a 1M, 8M, and 64M English-Dutch subset of the OPUS corpus (Tiedemann and Thottingal, 2020). We curate synthetic sentences in which we can control the lexical items and insert certain complex noun and verb phrases extracted from the OPUS data, yielding partially-natural, partially-synthetic evaluation data. We use the data and the models to answer the following research sub-questions:

How can we reformulate theoretically-grounded compositionality tests outside of toy task scenarios for NMT? We reimagine tests previously proposed by my co-authors and I (Hupkes et al., 2020) for English-Dutch data in an NMT setup:

- **Systematicity** evaluates the consistency of translations when recombining conjoined phrases with new phrases or when replacing words within a sentence (e.g. replacing "men" in "The girl sees that the men cry"). The recombinations are semantically unrelated and should not alter the translation when assuming

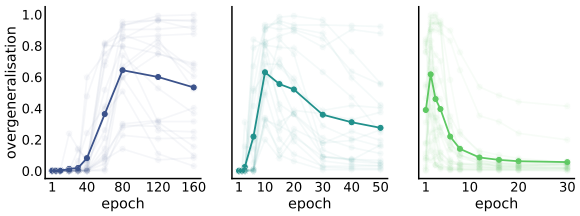


Figure 5: Overgeneralisation of idiomatic translations during training; an average is shown (bold) along with the trajectory of 20 individual idioms. From left to right, we show models trained on three training set sizes.

a locally compositional approach.

- **Substitutivity** evaluates translation consistency under synonym substitution, where two words in English map to the same Dutch word (e.g. ladybird/ladybug), using 20 synonym pairs. Since synonym substitution is meaning-preserving, the translations should not change.
- **Overgeneralisation** traces how 20 idioms, such as "out of the blue", are translated, assessing whether they are *overgeneralised* or *memorised*, based on the presence or absence of a literal translation of the idiom's keyword – i.e. translating "blue" as "blauw" would be overgeneralisation.

How compositional are NMT systems, and is the source of the errors natural language variation or model behaviour? Low systematicity and substitutivity consistency scores indicate that models often fail to behave compositionally under the strict local interpretation. We manually analyse 1800 inconsistent translation pairs, identifying that some inconsistencies reflect natural variation in language, but that the majority cannot reasonably be traced to linguistic ambiguity, underscoring NMT models' general volatility. This erratic behaviour highlights a lack of default reasoning, which can be problematic or even harmful in some cases, especially if faithfulness (Parthasarathi et al., 2021) or consistency is important to the end user.

How do NMT systems acquire non-compositional translations of idioms, and how does this align with generalisation performance? The third test demonstrates that models acquire idiomatic translations in two phases, as shown in Figure 5: early in training, the models learn to overgeneralise word-for-word translations, and later, they start to memorise paraphrases. Models' convergence based on memorisation does not appear to align with the other evaluation metrics; especially for the models trained on 1M and 8M data, more training would be needed to achieve memorisation. Interestingly, these models are simultaneously *not compositional enough* (as per the systematicity and substitutivity tests) and *too compositional* (as per the overgeneralisation test).

4.2 Conclusion and retrospective

Research on compositional generalisation often relies on artificial tasks that assume strictly local interpretations of compositionality. We argued that such interpretations overlook important aspects of natural language and proposed evaluating compositionality in NMT systems trained on natural data. We reformulated three compositionality tests showing that models simultaneously struggle with compositional generalisation and adequate memorisation of idioms. These findings highlight the difficulty of evaluating compositionality in natural language, where meaning composition is less clear-cut than in synthetic datasets. Following Baggio (2021), we suggest that human-like language use likely requires models to support both behaviours.

Our work also has limitations. Although we argue that compositional generalisation should ideally be evaluated using fully natural data, we rely on partially synthetic tests and do not propose direct solutions for improving compositional generalisation or non-compositional memorisation. Subsequent work, though, has leveraged our findings for actionable training techniques such as consistency regularisation (Yin et al., 2023) and novel dropout schemes (Niculae and Monz, 2023). Other studies have adapted our three tests to different languages, domains, and modalities (e.g. Liu, 2022; Li et al., 2024a; Liao et al., 2023; Kumon et al., 2024; Moisio et al., 2023). More broadly, we not only highlighted models' limitations but also explicitly encouraged the community to rethink how compositional generalisation is evaluated, rather than removing natural language variation for convenience. This call has been widely echoed (e.g. Zheng and Lapata, 2023; Sun et al., 2023; Chia, 2024; Chia et al., 2024; Fodor et al., 2025) and may be the most influential outcome of this research.

5 Mechanisms for idiomatic translations (Dankers et al., 2022b)

Having introduced the tension between compositional and non-compositional processing and framed idiom acquisition as a two-step process,

we now examine how pretrained models perform idiom translation. Idioms have long challenged NLP (e.g. Sag et al., 2002; Rayson et al., 2010; Shwartz and Dagan, 2019), particularly NMT systems (e.g. Barreiro et al., 2013; Isabelle et al., 2017; Constant et al., 2017; Avramidis et al., 2019). Not all *potentially idiomatic expressions* (PIEs) are figurative – e.g. consider "When I kicked the bucket, it fell over" – so correct translation depends on the context.[3] NMT systems must therefore learn to disambiguate usage, memorise the appropriate paraphrase, and generate it during decoding. Until the presentation of our work, the neural mechanisms underlying idiomatic translation remained poorly understood. Earlier work mainly examined how transformer-based LMs represent idioms (e.g. García et al., 2021a,b), but LMs merely need to detect figurativeness; they are not trained to explicate the idiomatic meaning. We present the first large-scale analysis of how transformers translate idioms, investigating whether models paraphrase or translate them word for word, and analysing their effects on self- and cross-attention as well as encoder hidden states.

5.1 Experiments and findings

We analyse `transformer-base` models (Vaswani et al., 2017) pretrained by Tiedemann and Thottingal (2020) for seven Indo-European language pairs (`En-Nl`, `-De`, `-Sv`, `-Da`, `-It`, `-Fr`, `-Es`) by comparing literal and figurative occurrences of PIEs. We address the following research sub-questions:

How can we perform analyses of NMT idiom processing at scale? Large-scale analyses of idiom translations suffer from a lack of parallel corpora (Fadaee et al., 2018). We therefore perform our analyses using data from the monolingual MAGPIE corpus (Haagsma et al., 2020), for which we devise a heuristic translation annotation method (similar to §4, but at a larger scale). We extract translations from our seven models and use a list of literal translations of idiom keywords to distinguish paraphrases from word-for-word translations. To validate the heuristic, we conducted human data annotation. Figurative PIEs should generally not be translated word for word due to their non-compositional meaning; however, only 20.7% of translations were paraphrased by the models. This

[3]Up to this point, we referred to idioms, but since this section considers both literal and figurative occurrences, we mainly use the term PIEs.

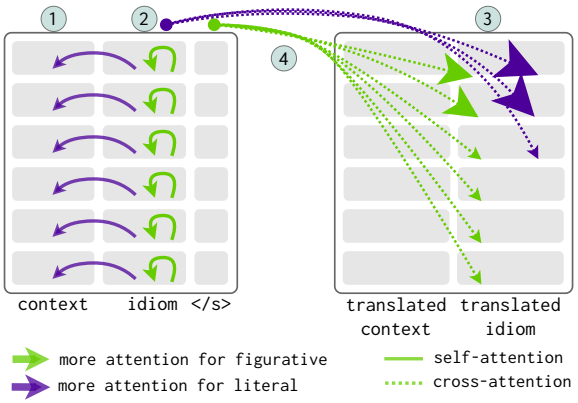


Figure 6: When comparing figurative, paraphrased PIEs to literal PIEs translated word for word, we find (1) less attention from PIE to context, (2) more attention within the PIE, (3) less cross-attention between PIE and its paraphrase, (4) more cross-attention from the paraphrase to `</s>`. The magnitude of the arrowhead indicates the effect size.

echoes findings from §4 that transformers can be *too* compositional, likely due to insufficient memorisation of idioms (§2). We further show that idiom frequency in the training corpus partially explain this.

How does idiomaticity and the paraphrasing of non-compositional idioms affect attention patterns and hidden representations? In the encoder, figurative PIEs are attended to more strongly as a single lexical unit than literal instances and interact less with the surrounding context. This aligns with prior work (e.g. Zaninello and Birch, 2020) showing that encoding idioms as single units improves translation. While paraphrasing PIEs, the decoder attends less to source tokens and more to the end-of-sentence token, temporarily detaching from the encoder input. Figure 6 visually summarises these findings. Hidden-state analysis confirms that these attention shifts affect the residual stream. The effects occur across layers: PIE representations gradually become more distinct, with figurative PIEs standing out from the first layer onward.

How do encoder-internal interventions affect non-compositional translations? We, lastly, intervene in the encoding of the PIEs using amnesic probing (Elazar et al., 2021), demonstrating that one can easily change non-compositional translations into compositional ones, underscoring that recalling memorised PIE paraphrases is a brittle process. When doing so, the transformer's self-attention changes such that the PIE components are grouped *less*, strengthening our previous findings.

Figure 7: Example of an overly compositional translation in Frisian ("chickens came home to roost" is included literally, but is not a Frisian idiom). Retrieved on March 15, 2026, from `GPT-5.3`.

5.2 Conclusion and retrospective

Summarising, we showed that idioms are often translated too compositionally and presented analyses of transformer mechanisms for paraphrasing idioms, pointing to a grouping mechanism in self-attention as key for treating idioms as single, non-compositional units. Our work had several limitations, most notably the focus on high-resource languages (due to the lack of high-quality idiom translations for low-resource languages) and the heuristic labelling of idiom translations. Nonetheless, it was among the first influential interpretability analyses of idiom translation. Closely related work includes Baziotis et al. (2023), Haviv et al. (2023), and Lim et al. (2024).

Although the lack of parallel idiom corpora remains a challenge, several multilingual idiom translation datasets have now been introduced (e.g. Amrhein et al., 2022; Stap et al., 2024; Lee et al., 2025), along with methods for improving idiom translations (e.g. Santing et al., 2022; Li et al., 2024b; Liu et al., 2023; Donthi et al., 2025). The largest translation quality gains since 2022, however, have come from scaling pretraining corpora for ever-growing LLMs. And yet, for the most powerful (non-)commercial translation systems and LLMs, researchers continued to echo our findings of overly compositional translations for a range of languages, such as German, Spanish and Japanese (Ferrando et al., 2023), Arabic (Obeidat et al., 2024), Urdu (Basit et al., 2024) and Indonesian (Dewayanti and Margana, 2024). In 2026, it still takes mere minutes to identify idiom translation failures for the most powerful models, particularly for low-resource languages (see Figure 7). Idiom translation has vastly improved, but it is not quite there yet!

6 Conclusion

Across the different sections, we investigated memorisation in neural models, its relationship to generalisation, and its connection to (non-)compositionality. Seven main lessons emerged with respect to the research questions I introduced in §1.

What characterises memorised examples? We find that *(1) memorisation is predictable rather than mysterious*. In §2, we introduced memorisation scores for 5M MT examples (with *Counterfactual Memorisation* being the core focus) showing that memorisation exists along a continuum. Much of the variation in these scores can be explained by surface-level features such as source-target overlap, and these patterns generalised across five language pairs. For idioms, a characteristic influencing memorisation of paraphrased translations was the frequency in the training corpus (§5). Second, *(2) what requires memorisation is not necessarily what models memorise under standard training regimes*. NMT systems often fail to memorise idioms and instead translate them compositionally (§2,4,5).

Which model-internal mechanisms enable memorisation? *(3) Memorisation is distributed across layers rather than localised*. In §3, localisation experiments on four transformer LMs across twelve tasks showed that memorisation of mislabelled examples emerges through cooperation across layers. Hidden representations gradually shift toward memorised labels rather than changing in a single layer, and deeper layers do not play a uniquely dominant role. Furthermore, *(4) idiom memorisation in translation involves grouping on the source side and reduced reliance on the encoder during decoding*. Attention analyses in §5 revealed that paraphrased idioms exhibit increased internal attention and reduced interaction with surrounding context, suggesting that they are processed as single units. During decoding, attention shifts away from idiom tokens toward the EOS token.

To what extent are memorisation and generalisation at odds with one another? *(5) Memorising atypical examples can support generalisation*. Experiments in §2 show that examples with higher CM scores benefit models' translation performance most, likely because examples with high CM are not merely noise, but representative of natural language variation in translations. Next, *(6) idiom acquisition follows a multi-phase process*. In §4,

tracing translations during training revealed an initial overgeneralisation phase followed by memorisation. For frequent idioms, models eventually produce memorised paraphrases, but many idioms remain in the overgeneralisation phase (§5), yielding overly compositional translations. Finally, *(7) transformers do not process language in a locally compositional manner*. By adapting synthetic compositional generalisation tests to natural-language MT data (§4), we identified that when we expect models to be locally compositional, they can actually be very volatile and inconsistent in their translations. This underscores a paradox: models are simultaneously not *compositional* and not *non-compositional* enough.

A final retrospective and outlook In summary, we learnt that transformer models memorise substantial aspects of their training data, which can support generalisation in natural language tasks. However, they often fail to memorise the types of formulaic expressions that require it, such as idioms. At the same time, transformers display both insufficient and excessive compositionality: non-local processing supports memorisation of idioms, yet harms compositional generalisation. Memorisation mechanisms emerge naturally but are distributed across layers and remain insufficiently adapted to natural language's formulaic nature.

Based on my findings, I propose several directions for future work. First, evaluations of compositional generalisation should not avoid non-compositional phenomena in natural language; methods that improve generalisation should also consider their implications for proverbs and idioms. Second, memorisation should be studied more holistically by focusing on memorisation circuits rather than individual neurons or layers. Accordingly, model editing and unlearning should move beyond layer-local approaches if the goal is true erasure of information. Third, memorisation can be beneficial and should be more explicitly incorporated into LLM design. Finally, as models are increasingly trained on their own generations, we risk losing subtle linguistic phenomena – such as idiomatic and proverbial expressions – that may become overgeneralised across languages. Which parts of natural language are we losing if transformer's own predictions become a part of that language? This warrants carefully crafted investigations, such as measuring the prominence of formulaic language in real and synthesised corpora.

Acknowledgments

While I refer the reader to my dissertation for the full acknowledgments, I'd like to repeat that I'm particularly grateful to my supervisor Ivan Titov and research mentor Dieuwke Hupkes for their support, expertise, guidance, insights, and enthusiasm throughout my PhD.

I, furthermore, thank Marius Mosbach and Cesare Spinoso-Di Piano for their feedback on the summary presented to you in this document.

Throughout the PhD I (Verna Dankers) was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

AI Assistant Usage

ChatGPT 5.3/5.4 was used to assist with the writing, primarily through shortening and proofreading. The AI assistant was not used to generate experimental results, to make scientific claims, or to determine conclusions. All experiments from the original papers, and all text therein, were produced without any assistance from LLMs.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. Aces: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513.

Jacob Andreas. 2018. Measuring compositionality in representation learning. In *International Conference on Learning Representations*.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32:6111–6122.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454.

Giosuè Baggio. 2021. Compositionality in a parallel architecture for language processing. *Cognitive Science*, 45(5):e12949.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Fernando Batista, and 1 others. 2013. When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, pages 26–33.

Abdul Basit, Abdul Hameed Azeemi, and Agha Ali Raza. 2024. Challenges in Urdu machine translation. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 44–49.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327.

Zheng Chia. 2024. *Exploring Optimal Settings for Machine Translation of Irony with Application to Multilingual Irony Detection*. Phd thesis, Kitami Institute of Technology.

Zheng Lin Chia, Michal Ptaszynski, Marzena Karpinska, Juuso Eronen, and Fumito Masui. 2024. Initial exploration into sarcasm and irony through machine translation. *Natural Language Processing Journal*, 9:100106.

Gilad Cohen, Guillermo Sapiro, and Raja Giryes. 2018. DNN or k-NN: That is the generalize vs. memorize question. *arXiv preprint arXiv:1805.06822*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022a. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175.

Verna Dankers and Christopher Lucas. 2023. Non-compositionality in sentiment: New data and analyses. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5150–5162.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

Verna Dankers and Vikas Raunak. 2025. Memorization inheritance in sequence-level knowledge distillation for neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 760–774.

Verna Dankers and Ivan Titov. 2022. Recursive neural networks with bottlenecks diagnose (non-)compositionality. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4361–4378.

Verna Dankers and Ivan Titov. 2024. Generalisation first, memorisation second? Memorisation localisation for natural language classification tasks. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14348–14366.

Verna Dankers, Ivan Titov, and Dieuwke Hupkes. 2023. Memorisation cartography: Mapping out the memorisation-generalisation continuum in neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8323–8343.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Desakh Putu Setyalika Putri Dewayanti and Margana Margana. 2024. The impact of contextual understanding on neural machine translation accuracy: A case study of Indonesian cultural idioms in English translation. *Englisia: Journal of Language, Education, and Humanities*, 12(1):223–236.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vitaly Feldman. 2020. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2881–2891.

Javier Ferrando, Matthias Sperber, Hendra Setiawan, Dominic Telaar, and Saša Hasan. 2023. Automating behavioral testing in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1014–1030.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Marcos García, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741.

Marcos García, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 279–287.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10711–10732.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intellgence Research*, 67:757–795.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, and 1 others. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *The Seventh International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. COGS: a compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105.

Tomasz Korbak, Julian Zubek, and Joanna Rączaszek-Leonardi. 2020. Measuring non-trivial compositionality in emergent communication. In *NeurIPS 2020 workshop on Emergent Communication.*

Ryoma Kumon, Daiki Matsuoka, and Hitomi Yanaka. 2024. Evaluating structural generalization in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13220–13239.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR.

Minjae Lee, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646.

Chuanhao Li, Zhen Li, Chenchen Jing, Yuwei Wu, Mingliang Zhai, and Yunde Jia. 2024a. Compositional substitutivity of visual reasoning for visual question answering. In *European Conference on Computer Vision*, pages 143–160. Springer.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024b. Translate meanings, not just words: IdiomKB's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Weiduo Liao, Ying Wei, Mingchen Jiang, Qingfu Zhang, and Hisao Ishibuchi. 2023. Does continual learning meet compositionality? New benchmarks and an evaluation framework. *Advances in Neural Information Processing Systems*, 36:33499–33513.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111.

Yutong Liu. 2022. Compositional generalization in machine translation for low-resource languages. Master's thesis, University of Edinburgh.

Gary F Marcus. 2003. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Tarun Ram Menta, Susmit Agrawal, and Chirag Agarwal. 2025. Analyzing memorization in large language models through the lens of model attribution. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10661–10689.

Anssi Moisio, Mathias Creutz, and Mikko Kurimo. 2023. On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 204–213.

Ryan M Nefdt. 2020. A puzzle concerning compositionality in machines. *Minds & Machines*, 30(1).

Vlad Niculae and Christof Monz. 2023. Joint dropout: Improving generalizability in low-resource neural machine translation through phrase pair variables. *MT Summit 2023*, page 12.

Mohammed M Obeidat, Ahmad S Haider, Sausan Abu Tair, and Yousef Sahari. 2024. Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic. *FWU Journal of Social Sciences*, 18(4).

Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3205–3227.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, and 1 others. 2024. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161.

Vikas Raunak and Arul Menezes. 2022. Finding memo: Extractive memorization in constrained sequence generation tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5153–5162.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begona Villada Moirón. 2010. Multiword expressions: Hard going or plain sailing? *Language Resources and Evaluation*, 44:1–5.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002, Proceedings*, pages 1–15. Springer.

Lukas Santing, Ryan Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij, and Riza Batista-Navarro. 2022. Food for thought: How can we exploit contextual embeddings in the translation of idiomatic expressions? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 100–110.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206.

Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and Sue Yeon Chung. 2021. On the geometry of generalization and memorization in deep neural networks. In *The Ninth International Conference on Learning Representations*.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

Kaiser Sun, Adina Williams, and Dieuwke Hupkes. 2023. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 274–293.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Lena Voita. 2024. Analysis methods for natural language processing.

Alison Wray. 2002. *Formulaic language and the lexicon.* ERIC.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, Jie Zhou, and Yue Zhang. 2023. Consistency regularization training for compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1294–1308.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3816–3825.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102.

Hao Zheng and Mirella Lapata. 2023. Real-world compositional generalization with disentangled sequence-to-sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1711–1725.

Xiaosen Zheng and Jing Jiang. 2022. An empirical study of memorization in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6265–6278.

A Advice from beyond the PhD

① Don't stress but plan ahead Questions I often received towards the end of the PhD were along the lines of "How do I ensure that my thesis has a cohesive storyline?" and "Did you plan for this story from the start?". Answering them is non-trivial: in a fast-paced field like NLP, planning three to six years ahead for a *top-down* approach to PhD research is nearly impossible, and many a thesis is constructed in a *bottom-up* manner, where the candidate bundles their papers towards the end, determining the storyline somewhat post-hoc. This works well for many, so do not worry too much if you are currently in a situation in which you can't completely see the dissertation's forest through your papers' trees yet. However, I personally very much enjoy the fact that my dissertation has clear, connected themes.

I believe that crucial reasons for why this is the case are (a) that I entered the PhD with a strong interest in (non-)compositionality and generalisation and didn't stray too far from that through the years, while having a flexibility regarding projects to conduct along the way, and (b) that I planned my projects ahead of time, supported by my university's administrative structures. During three intermediate PhD evaluations, I planned ahead. During evaluation 1, I sketched out research directions for the next year, during evaluation 2, I discussed how the thesis's storyline was starting to emerge, and what was missing, and during evaluation 3, I discussed a concrete thesis outline. This does not mean that I stuck exactly to what was planned (see lesson ②, below), but it was a massive help. If you're a student, and your university/supervisor does not enforce such evaluation and planning checkpoints, be proactive and schedule them yourself, at least once a year.

For me, the right approach was not working bottom-up, not working top-down, but adopting a *hybrid* approach, where most meetings are low-level, bottom-up project meetings, but some take the helicopter view of the PhD as a marathon, to study whether I was on track, and was headed in the right direction.

② Curate the thesis's storyline I can't say with certainty that I did it well, but I attempted to curate a very specific storyline in the dissertation. Three things I did to achieve that, are the following:

Selecting what to include. Not everything one does during a PhD has to be included in the thesis, and even if you include all first-author papers, you can often write multiple stories using the same papers. If you've read the summary of my thesis above, you will have noticed that it contains two primary parts, circled around 'memorisation vs generalisation' and 'compositionality vs non-compositionality'. Planning for the thesis writing, I originally thought I'd include a Part III – based on Dankers and Titov (2022); Dankers and Lucas (2023), as summarised in Appendix B – but I followed the advice of my UoE annual review board to omit it. More isn't always better.

Deciding where to include what. In my case (but perhaps not yours) the papers didn't have to be included in a chronological order, and I opted for putting the most recent two first to curate a narrative in which the second two papers (on (non-)compositionality) are considered a case study of the more general memorisation–generalisation paradox.

Going beyond the original papers. Every experimental chapter of my thesis contains experiments that were not in the original paper, either because I felt something was missing that could have made the conclusions stronger, or because seeing the individual papers in the light of the overall dissertation made me realise that certain experiments could provide inter-chapter connections.

③ Plan for reproducibility By the time the thesis comes around you might be getting back into 4-year-old codebases, to see whether you can recreate some graphs with old data – either just because the thesis would look nicer with updated graphs, or because you actually want to extend the experiments, like I did. At that point, you might realise that not everything is reproducible. Even with the greatest `README.md` out there, packages will change, clusters will change, and checkpoints you thought you had stored will no longer be there because the cluster admin decided a clean-up was in order. When you're now only one year into the PhD, I know you're likely not thinking of the dissertation, and you may want to move on from a project as soon as it is submitted, but please invest time into the reproducibility aspect. First and foremost because of others that might want to build upon your work, and second because you yourself may want to re-run those exact experiments. Think ahead: "Which model checkpoints would be crucial if I wanted to test a few more hypotheses for the thesis?", or

"What data would I need to store if I wanted to regenerate the figures, or run an additional statistical significance test on the results?", or "Should I ask my internship manager for approval for exporting these models, since I won't be an intern at the time of graduation?".

I wish I had curated a PhD-thesis-ready version of all of my papers on a dedicated hard drive, with the most interesting models, and the data required to regenerate tables and graphs. Although it was there for most of the chapters, some of it needed to be regenerated.

④ **Give your dataset a name** A very silly mistake I realised I made was that in Dankers et al. (2022a) I produced a compositional generalisation evaluation dataset that never received a name. Therefore, various papers referred to it under names they had come up with. The consensus was 'OPUS En-Nl', but we clearly should have assigned the evaluation set a name. Lesson learnt!

⑤ **Write a retrospective** Inspired by Voita (2024)'s sections entitled "Implications: View from the future", I wrote my own "Retrospective and outlook" sections, which were also highly recommended by my phenomenal supervisor Prof Titov. Taking my own work, and reflecting on (a) what impact this specific paper has had and (b) how the field has changed since this paper saw the light of day, was very informative for myself, and according to my thesis examiners. I can safely say they are my favourite thesis sections.

B The 'Lost' Part III: Quantifying (non-)compositionality

There are two papers that I initially planned to include in the thesis, but did not, in the end (Dankers and Titov, 2022; Dankers and Lucas, 2023). These articles focus on the notion of (non-)compositionality: Across sentences and even across subphrases of one sentence, there is variation in terms of how compositional phrases are. While there is a wealth of knowledge about how very specific types of non-compositional phenomena behave, quantifying the compositionality of phrases or sentences in general is an open and ill-defined problem. Both of these articles address this open problem from different angles by using either model-computed or human-derived quantifications.

Recursive model-based metric (Dankers and Titov, 2022) Gaining a better understanding of the challenges that the (non-)compositionality of natural language presents to neural models, requires metrics for quantifying that (non-)compositionality. While this had previously been investigated on a small scale, such as for idioms (Ramisch et al., 2016), for natural language bigrams (Andreas, 2018), or for artificial languages in the context of communication games (Korbak et al., 2020), the metrics proposed for those applications did not easily scale to natural language sentences.

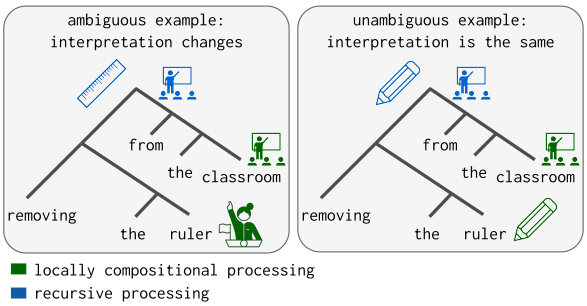


Figure 8: When processing this phrase, "the ruler" is interpreted differently when comparing recursive processing with local processing. We enforce local processing by equipping models with bottlenecks, and our **bottleneck compositionality metric (BCM)** then compares inputs' representations before and after compression through the bottleneck.

In this work, we extended one such metric, namely the *Tree Reconstruction Error* (TRE) by Andreas (2018), that expresses the distance between a model's representation of an input and a strictly compositional reconstruction of that representation. To do so, we used recursive neural networks, namely Tree-LSTMs (Tai et al., 2015), to process inputs according to their syntactic structure. We augmented Tree-LSTMs with bottlenecks to compute the meaning of an input with respect to a certain task in a more locally compositional manner. We used these models to distinguish more compositional examples from less compositional ones in a *bottleneck compositionality metric* (BCM). BCM compares a regular model trained to perform a task to a model augmented with bottlenecks. Under the assumption that non-compositional processing of an input requires more complex meaning representations, the bottlenecks will hinder examples for which the model finds a non-compositional solution most, as is illustrated in Figure 8. As such, the difference between a model without the bottlenecks and one with the bottlenecks acts as a metric.

We experimented with three types of bottlenecks,

namely a *deep variational information bottleneck* (DVIB) (Alemi et al., 2017), compressing representations through increased dropout or simply using smaller hidden dimensionalities. As a proof of concept, the BCM was applied in a controlled environment where non-compositional examples were manually introduced, by taking arithmetic expressions and making one vocabulary item ambiguous. Afterwards, we applied the BCM to the real-world example of sentiment analysis using the *Stanford Sentiment Treebank* (SST) dataset (Socher et al., 2013). For both tasks, we illustrated that compression through a bottleneck encourages local processing, and showed that the bottleneck can act as a metric distinguishing compositional from less compositional samples. We, furthermore, used the compositionality judgments for the SST data to demonstrate that (i) in a training data scarce scenario, compositional training examples yield models that generalise better to test data, and (ii) that when the test set contains non-compositional examples, performance is substantially lower compared to a test set of compositional examples.

Using human annotations for compositionality judgments (Dankers and Lucas, 2023) In this work, we used a different approach and focused on human data annotations to obtain quantifications of natural language phrases' (non-)compositionality. For such phrases, their meaning is often more than 'just' the compositional sum of their parts. In the context of sentiment analysis, the 'meaning' of a phrase is its sentiment, and even though sentiment computations are largely compositional, there are still exceptional patterns. We selected the task of sentiment analysis as a testbed for obtaining non-compositionality ratings for phrases because of this rather straightforward interpretation of 'meaning', which is much less well defined in other tasks or when obtaining task-generic compositionality judgments.

We first designed a protocol to obtain non-compositionality judgments based on human-annotated sentiment. Our methodology uses phrases from the SST dataset (Socher et al., 2013) and contrasts the sentiment of a phrase with that of control stimuli, in which one of two subphrases has been replaced. Phrases whose annotated sentiment deviates from what is expected are considered less compositional. This approach, along with the example of the non-compositional phrase "all the excitement of eating oatmeal", is depicted

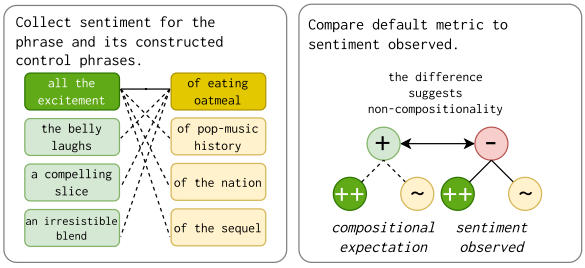


Figure 9: Illustration of how the non-compositionality ratings are obtained in NONCOMPSST: we contrast a sentiment's phrase to the sentiment of control phrases.

in Figure 9. We developed a resource of ratings for 259 phrases (dubbed NONCOMPSST) through sentiment annotations from 147 participants total, who provided us with more than 10,000 annotations via Prolific. Secondly, we presented an analysis of that resource, emphasising the higher non-compositionality ratings for figurative language, and use NONCOMPSST to evaluate computational models for sentiment analysis, focusing on conventional pretrained models such as ROBERTA, as well as models pretrained on sentiment-laden data, and models finetuned for sentiment analysis. Performance on conventional SST test data was markedly higher compared to performance on NONCOMPSST, underscoring that non-compositional phrases challenge models more. We suggest that NONCOMPSST can complement existing evaluation protocols for sentiment analysis models.

Author Index