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.