

Analogy-Angle 2025

**The 2nd Workshop on Analogical Abstraction in Cognition,
Perception, and Language (Analogy-Angle II)**

Proceedings of the Workshop

August 1, 2025

©2025 Association for Computational Linguistics

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 979-8-89176-274-9

Preface by the Workshop Chairs

Welcome to the 2nd edition of the Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)!

Analogical abstraction is a fundamental human cognitive skill (Penn et al. 2008) (Penn et al. 2008; Hofstadter 2001) defined as the ability to perceive and utilize the similarities between concepts, situations, or events based on (systems of) relations rather than surface similarities (Holyoak 2012; Gentner and Smith 2012). Analogy enables creative inferences, explanations, and generalization of knowledge and has been used for scientific inventions (Dunbar 2012), solving problems (Gick and Holyoak 1980), and policy-making Houghton 1998. As such, it has been the subject of cognitive theories and studies about humans for standard processes, such as retrieving memories (Wharton et al. 1994) and problem-solving (Gick and Holyoak 1980).

Analogical tasks have gained considerable popularity in natural language processing (NLP) and artificial intelligence (AI), where they are often framed as tests of a model’s intelligence in comparison to human performance. These tasks typically involve so-called word-based proportional analogies of the form (A : B :: C : D). (Mikolov et al. 2013a; <https://arxiv.org/pdf/1301.3781>; Gladkova et al. 2016; Ushio et al. 2021) lend themselves well to large language models (LLMs) (Webb et al. 2023). However, controlling for association and memorization (Stevenson et al. 2023; Lewis and Mitchell 2024) or shifting toward more complex settings like narratives reveals limitations in scope, generalizability, and alignment with cognitive theories (Nagarajah et al. 2022; Wijesiriwardene et al. 2023; Sourati et al. 2024).

Inspired by the richness of analogical abstraction and the wide interest in this topic from computational linguistics, artificial intelligence, and cognitive psychology, ANALOGY-ANGLE II connects these communities and facilitates cross-disciplinary activities. ANALOGY-ANGLE II welcomes novel contributions in short, long, and review formats, as well as relevant papers accepted at top-tier venues over the past year (so-called *dissemination papers*).

Topics of interest belong to four major categories:

- 1. Cognitive modeling**, with the following sub-topics: analogy and abstraction; relation to conceptual metaphor, figurative language, sarcasm, and irony; cognitive frameworks of analogy; and cognitive psychological studies on analogy involving human participants.
- 2. Algorithms and methods**, featuring the sub-topics: studies of the analogical abilities of LLMs and multimodal models; algorithmic approaches to analogy; augmentation and verification of LLMs and multimodal models through analogy; neuro-symbolic AI architectures for analogical abstraction; and extracting analogies from knowledge bases.
- 3. Tasks and benchmarks**, with sub-topics: matching narratives and situational descriptions through narratives; novel tasks and benchmarks for evaluating analogies in text and vision; analogy in longer formats, e.g., narratives and videos; analogy and visual abstraction tasks; and analogical discovery in computational creativity.
- 4. Applications**, such as: personalization, explanation, and collaboration; novel applications of analogical abstraction; studies of the impact of analogy in specific applications and domains, including education, innovation, and law.

ANALOGY-ANGLE II aims to form a comprehensive picture of the current landscape of analogical abstraction, and point to standing challenges, evaluation methodologies, and emerging techniques of interest. To this end, the workshop features two prominent speakers with complementary perspectives, and paper sessions that cover a wide range of the envisioned topics. The multidisciplinary nature of

ANALOGY-ANGLE II is supported by the broad skills of the organization team and the program committee, as well as the diversity principle guiding the list of topics and the invited keynotes. After a successful first workshop at IJCAI 2024, we see ACL as an ideal venue for our second iteration. ACL is a primary venue for cutting-edge research on analogy and abstraction, facilitating discussions with leaders at the intersection of natural language processing, cognitive linguistics, and perception.

We received 14 novel submissions. 10 of them were submitted via direct submission, and the remaining 4 through the ARR commitment. Based on three reviews per submission, we accepted 9 novel papers, resulting in an acceptance rate of 64%. Additionally, 3 non-archival, cross-submissions were accepted based on two reviews per paper, assessing their relevance.

We are deeply grateful to the members of the Program Committee for their dedicated efforts. Our sincere thanks also go to our invited speakers, Melanie Mitchell and Ekaterina Shutova, for their valuable and inspiring contributions to this year's program. We thank Melanie Mitchell for her talk on abstraction in humans and AI systems and Ekaterina Shutova for her talk on metaphors.

Finally, we gratefully acknowledge the support of our funding bodies, in particular the NWO AiNed project "Human-Centric AI Agents with Common Sense", which funded Filip Ilievski, and the European Research Council's ABSTRACTION project (Grant agreement: ERC-2021-STG-101039777), which funded Marianna Bolognesi and Giulia Rambelli.

The Analogy-angle II 2025 Organizing Committee

Program Committee

Workshop Chairs

Marianna Bolognesi, University of Bologna
Filip Ilievski, Vrije Universiteit Amsterdam
Giulia Rambelli, University of Bologna
Ute Schmid, University of Bamberg
Pia Sommerauer, Vrije Universiteit Amsterdam

Reviewers

David M. Cerna, Czech Academy of Sciences Institute of Computer Science
Cas W. Coopmans, New York University
Sydelle De Souza, University of Edinburgh
Valentin Forch, Technische Universität Chemnitz
Diego Frassinelli, Ludwig-Maximilians-Universität München
Frank Guerin, University of Surrey
Mohammad Hossein Khojasteh, Vrije Universiteit Amsterdam
Gustaw Opielka, University of Amsterdam
Zeynep G. Saribatur, Technische Universität Wien
Hwanjun Song, Korea Advanced Institute of Science and Technology
Zhivar Sourati, University of Southern California
Eunice Yiu, UC Berkeley
Leonidas A. A. Doumas, University of Edinburgh, University of Edinburgh
Yifan Jiang, Information Sciences Institute, University of Southern California
Sundong Kim, Gwangju Institute of Science and Technology
Martha Lewis, University of Amsterdam
Antonio Lieto, University of Salerno
Henri Prade, CNRS, France and University of New South Wales
Philipp Wicke, Ludwig-Maximilians-Universität München
Alessandra Zarcone, Technische Hochschule Augsburg

Table of Contents

<i>Tore-Klose: Record Scorer, Goal Hunter, Machine? Human Association Norms for German Personal Name Compounds</i>	
Annerose Eichel, Tana Deeg, Andre Blessing, Milena Belosevic, Sabine Arndt-Lappe and Sabine Schulte Im Walde.....	1
<i>Using Large Language Models to Perform MIPVU-Inspired Automatic Metaphor Detection</i>	
Sebastian Reimann and Tatjana Scheffler.....	10
<i>Modeling Background Knowledge with Frame Semantics for Fine-grained Sentiment Classification</i>	
Muhammad Okky Ibrohim, Valerio Basile, Danilo Croce, Cristina Bosco and Roberto Basili .	22
<i>On choosing the vehicles of metaphors without a body: evidence from Large Language Models</i>	
Veronica Mangiaterra, Chiara Barattieri Di San Pietro, Federico Frau, Valentina Bambini and Hamad Al-Azary.....	37
<i>Prompting Metaphoricity: Soft Labeling with Large Language Models in Popular Communication of Science Tweets in Spanish</i>	
Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba and Gerardo Sierra ..	45
<i>HATS : Hindi Analogy Test Set for Evaluating Reasoning in Large Language Models</i>	
Ashray Gupta, Rohan Joseph and Sunny Rai.....	57
<i>Simulating Emotional Intelligence in LLMs through Behavioral Conditioning and Analogical Retrieval</i>	
G.Sai Linisha Reddy, Mounil Hiren Kankhara, Mridul Maheshwari, Swayam Bansal, Rishit Kapoor, Himesh Reddy M and Bagesh Kumar.....	81
<i>Can Stories Help LLMs Reason? Curating Information Space Through Narrative</i>	
Vahid Sadiri Javadi, Johanne Trippas, Yash Kumar Lal and Lucie Flek.....	92
<i>Testing Spatial Intuitions of Humans and Large Language and Multimodal Models in Analogies</i>	
Ivo Bueno, Anna Bavaresco, João Miguel Cunha and Philipp Wicke.....	108

Tore-Klose: Record Scorer, Goal Hunter, Machine? Human Association Norms for German Personal Name Compounds

Annerose Eichel¹, Tana Deeg¹, André Blessing¹, Milena Belosevic²,
Sabine Arndt-Lappe³, Sabine Schulte im Walde¹

¹Institute for Natural Language Processing, University of Stuttgart

²Faculty of Linguistics and Literary Studies, Department Linguistics, University of Bielefeld

³English Linguistics and Trier Center for Language and Communication, Trier University

{annerose.eichel, tana.deeg, andre.blessing, schulte}@ims.uni-stuttgart.de,

milena.belosevic@uni-bielefeld.de, arndtlappe@uni-trier.de

Abstract

We present a collection of human association norms to German personal name compounds (PNCs) such as *Tore-Klose* ('goal-Klose') and corresponding full names (Miroslav Klose), thus providing a novel testbed for PNC evaluation, i.e., analogical vs. contrastive positive vs. negative perception effects. The associations are obtained in an online experiment with German native speakers, analyzed regarding our novel intertwined PNC–person association setup, and accompanied by an LLM synthetic generation approach for augmentation.

1 Introduction

Personal name compounds (PNCs) such as *Tore-Klose* ('goal-Klose') and *Bierzelt-Söder* ('beertent-Söder'), i.e., nominal compounds representing modifications of personal names, are prevalent in social media, news, and political discourse. Recent work suggests that PNCs represent a rather frequent phenomenon across languages that establishes an evaluative function with regard to the reference person (Belosevic and Arndt-Lappe, 2021; Belosevic, 2022). More specifically, PNCs are generally perceived either as more positively or as more negatively in comparison to the perception of the corresponding name holder. Most recently, this evaluative nature of name compounds has been addressed and confirmed by regression analyses and large language models, identifying name and compound properties based on contextual data and valence norms (Eichel et al., 2024).

Our work approaches PNC evaluation from a novel perspective: based on the assumption that human associations provide a window into lexical-semantic representations (Kiss et al., 1973; Nelson et al., 2004; Schulte im Walde and Melinger, 2008; Schulte im Walde et al., 2008; de Deyne et al., 2019, i.a.), we propose to collect and utilize associations to PNCs and corresponding full names in order to

identify and measure salient evaluative characteristics which go beyond previously exploited knowledge sources such as contextual and valence information. I.e., we hypothesize that (i) the evaluative meanings of PNCs are reflected in free, spontaneous associations to the compounds, in analogy or in contrast to associations to the person names, and that (ii) the associations reveal reasons for using the specific compound modifiers, such as individual persons' characteristics or stances as well as specific events. For example, our associations include *elitär* ('elitist') and *abgehoben* ('aloof') to *Privatschul-Schwesig* ('private-school-Schwesig'); *grün* ('green') and *Privatflüge* ('private flights') to *Bonusmeilen-Özdemir* ('bonus-miles-Özdemir'), where the PNC *Privatschul-Schwesig* refers to the Social Democratic Party politician Manuela Schwesig who puts a political focus on child support and family affairs but decided to send her own son to a private rather than a public school; the PNC *Bonusmeilen-Özdemir* relates to the Green Party politician Cem Özdemir who was strongly criticized for using business-accumulated bonus miles for private travel.

The current paper presents our collection of association norms for 195 German PNCs and their corresponding full names, using PNCs from an existing target set and the domains *politics*, *sports*, *others* (Belosevic and Arndt-Lappe, 2021; Eichel et al., 2024) in an online experiment with German native speakers. For further reference points towards absolute and relative PNC evaluation, we also ask the experiment participants to provide (i) their familiarity with a given PNC or person name, and (ii) their subjective positive vs. negative person judgments. Our contributions can be summarized as follows:

- We collect and devise human free association norms that provide a window to the mind of a communicator for 195 German personal

name compounds and their respective 66 names. The norms are publicly available from <https://github.com/AnneroseEichel/pnc-association-norms>.

- We present a series of analyses demonstrating that our intertwined PNC–name association setup unmask salient analogical vs. contrastive assessments and thus expands research perspectives on the evaluative nature of PNCs beyond existing previous evidence.
- We outline LLM-based synthetic association generation experiments to augment the novel small-scale human norms.

2 Background and Related Work

Personal Name Compounds (PNCs) PNCs such as *Tore-Klose* (‘goal-Klose’) are nominal compounds consisting of a modifier, typically an appellative or onymic constituent such as *Tore* (‘goals’), and a head constituent referring to a first, last, or nickname (e.g., *Klose*) (Belosevic, 2022). PNC construction follows regular patterns and draws on contextual knowledge about the person they refer to (Belosevic and Arndt-Lappe, 2021). For instance, *Tore-Klose* refers to Miroslav Klose, the former German soccer player who holds the all-time top scorer title for Germany. This example demonstrates how the compound modifier contributes information about the person or events associated with them. In this case, *goal* implies a positive assessment, referring to remarkable athletic performance and extraordinary goaling.

Association Norms Under the assumption that human free associations (i.e., the first word(s) that come(s) to mind, such as *white*, *cold*, *winter* as associations to *snow*) provide a window into lexical-semantic representations and meaning components of the stimuli, association norms have been collected across research disciplines and languages for many decades. Prominent examples of English association norms are the Edinburgh Association Thesaurus as the first collection on a larger scale (Kiss et al., 1973), the long-term collection at the University of South Florida (Nelson et al., 2004), and the currently largest Small World of Words (SWOW) norms (de Deyne et al., 2019). Earlier collections for German include the first collection by Russell and Meseck (1959), word-class-specific verb and noun norms collected by

Schulte im Walde et al. (2008), and previous compound and constituent norms collected by Schulte im Walde and Borgwaldt (2015) – differently to the current collection they focus on general-language noun compounds and their constituents, and utilize the associations to assess implicit features of (non-)compositionality.

3 Data Collection and Post-Processing

Target Stimuli Selection We start out with 215 eventive PNCs and their corresponding 85 names devised by Eichel et al. (2024). Through a pilot study, we assess participant familiarity with the real-world people behind the PNCs and collect feedback regarding study setup and target presentation (see Appendix A for details on the pilot study). Based on the collected responses, we exclude 20 PNCs and their corresponding 19 names from the domain *politics*, because they were not known by a clear majority of participants. We randomly assign the remaining 195 PNC and 66 name stimuli to 5 PNC and 2 name experiment lists.

Study Setup Our study was carried out via Prolific¹ and Google Forms. We screen study participants only regarding their specified native language, which we require to be German. Participants are provided written study guidelines and an example with potential responses (see Appendix B for details on the guidelines and experiment interface). In the actual experiment, each trial consisted of a question regarding the participant’s familiarity with a PNC and corresponding name, or only the name, depending on the experimental list. If they knew the PNC/name, participants were asked to provide 3–5 associate responses to the given stimulus in form of a word or a phrase, and to indicate how they perceive the public figure on a scale from 0 (positive) to 5 (negative). If they had questions, participants were provided the option to contact the authors of the paper. Participants could complete the study flexibly within their own schedule and no time constraints. Average time effort per list was 26^{±3} minutes (PNC) and 26^{±1} minutes (name). Each participant could only submit one response set for a stimulus.

Post-Processing and Data Each submitted response provides us with a stimulus, an indication of familiarity regarding a given public figure, the associate responses in the given order, a

¹<https://www.prolific.com/>

score perception of the public figure as well as demographic information of the participant, e.g., *Miroslav Klose*; known; [*Fußballspieler, Nationalmannschaft, Salto*]²; 2; 30-40 years.

To allow for meaningful automatic processing of PNC and name associations, we perform manual post-processing of participant responses such as correcting typos, standardizing spelling variants, delimiters, and abbreviations, and eliminating responses where participants indicate that they do not know a PNC and/or name, as well as responses that are clearly comments such as “no more associations”. For this, we hire a German native speaker who has a background in computational linguistics and is provided detailed instructions and written guidelines regarding desired formatting and relevant corrections (cf. Appendix C for details).

For PNCs, we elicit a total of 980 associate words and 438 associate phrases³ provided as first associates, and 3,526 words and 1,353 phrases when taking into account all associations. The average number of first and all associates per PNC is 23.9 and 7.4, respectively. For full names, a total of 328 words and 127 phrases is collected as first associates, as well as 1,229 words and 468 phrases considering all associates. The average number of first and all associates per full name is 25.7 and 6.9, respectively.

Participants 90 German native speakers participated in the experiment. 67 and 30 participants provided valid responses for an average of 54^{±3} and 49^{±11} stimuli in the PNC and name parts of the study, respectively. 93% and 7% of participants provided valid responses for only one or both experiment parts, respectively.

4 Analysis of PNC and Name Associations

Distributions Table 1 presents frequency distributions, i.e., how often specific associations have been provided across target stimuli. For this, we distinguish between first-only as well as full association lists across PNCs and full names. While PNCs receive slightly more associates and single word responses overall (cf. §3), associates tend to differ from each other with maximum associate frequencies of 8 and 10 among first-only and all associations. In contrast, corresponding names seem to evoke more similar responses with maximum

FREQ.	PNC		NAME	
	1 st	all	1 st	all
1	0.865	0.878	0.721	0.800
2	0.091	0.077	0.127	0.099
3	0.031	0.020	0.068	0.040
4	0.001	0.013	0.042	0.012
5+	0.005	0.013	0.042	0.042

Table 1: Associate frequency distribution across target stimuli (proportions). Maximum frequency of 1st and all associates: 8 and 10 (PNC); and 9 and 12 (name).

PoS	PNC		NAME	
	1 st	all	1 st	all
NOUN	0.35	0.36	0.47	0.38
NAME	0.15	0.13	0.19	0.18
VERB	0.07	0.09	0.02	0.06
ADJ	0.09	0.09	0.14	0.10
ADV	0.11	0.14	0.09	0.16
OTHERS	0.23	0.19	0.08	0.12

Table 2: PoS tag distributions (proportions).

associate frequencies of 9 and 12 among first-only and all associations.

Targets with higher frequency responses among first associates usually evoke a strongly salient association connected to the PNC modifier or the person in general, such as the PNC *Knast-Hoeneß* (‘jail-Hoeneß’) eliciting the response *Steuerhinterziehung* (tax evasion, freq: 6), and a variety of single first associates related to the modifier including *Knacki* (‘jailbird’), *Steuerbetrüger* (‘tax fraudster’), *weil er für eine Steuerhinterziehung in den Knast musste* (‘because he had to go to jail for tax evasion’), *Steuern* (‘taxes’), and more generally to his person such as *FC Bayern* (‘FC Bayern’) and *ehemaliger Fußballspieler* (‘former soccer player’). In the case of politicians, first associates often correspond to political party membership, e.g., as first associates to the full name *Alexander Gauland* we find *AfD* (‘AfD’, freq: 8), *Politiker* (‘politician’, freq: 3), and *Rassismus* (‘racism’, freq: 1), while athletes are mainly connected to their sport or club, as first associates to the PNC *Vollgas-Vettel* (‘pedal-to-the-metal-Vettel’) show: *Formel 1* (‘formula 1’, freq: 7), *Rennfahrer* (‘racing driver’, freq: 2), and single responses such as *Autorennen* (‘car racing’), *Weltmeister* (‘world champion’), *Ferrari* (‘ferrari’).

Morpho-Syntactic Analysis We distinguish the collected associations with regard to the major PoS tags: nouns, verbs, adjectives, and adverbs.

²EN: ‘soccer player, national team, somersault’

³A phrase denotes all responses consisting of 1+ words.

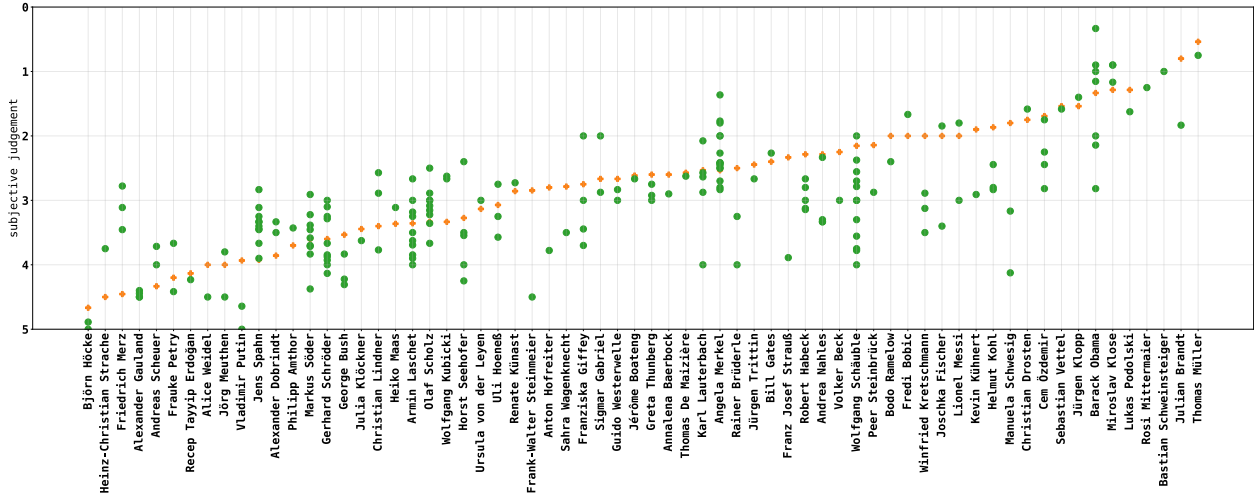


Figure 1: Overview of name (orange dots) vs. PNC (green dots) mean perception where 0 is positive and 5 is negative. 1+ PNCs can relate to one name, e.g., *Tore-Klose* (‘goal-Klose’), *Pokal-Klose* (‘trophy-Klose’), and *Salto-Klose* (‘somersault-Klose’) all referring to Miroslav Klose and more positively perceived than the name itself.

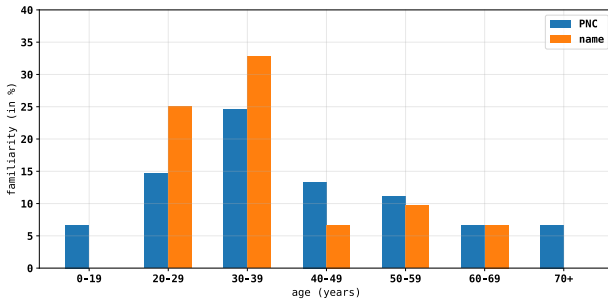


Figure 2: Average familiarity with target stimuli across age ranges. No participants below and above the age of 20 and 70 took part in the name part of the experiment.

For this, we remove all punctuation and obtain – possibly ambiguous– PoS tags for each word and phrase associate using spaCy (Honnibal and Montani, 2017). Table 2 presents PoS distributions as proportions. Across target stimuli and major PoS tags, participants provided noun and proper name associates in the clear majority of instances, followed by adverbs and adjectives, and finally verbs. We also report the proportion of the sum of all other PoS tags such as pronouns with higher prevalence in responses to PNC than full name targets.

Perception of real-world reference person For further reference points towards PNC evaluation, we ask experiment participants for their personal subjective positive vs. negative judgments regarding the target stimuli. Figure 1 shows collected judgments for PNCs and their corresponding full names on a scale from positive (0) to negative (5).

Regarding person names (orange dots), we generally observe domain-specific differences between politicians and athletes gathered more on the negative (left) vs. positive (right) part of the plot, respectively. The PNC-related evaluations (green dots) diverge in individual ways from the name-related evaluations, e.g., one of the green dots for Barack Obama below the orange name-related dot refers to the more negatively perceived PNC *Bomben-Obama* (‘bomb-Obama’); while one of the green dots above the orange name-related dot refers to the more positively perceived PNC *Hoffnungs-Obama* (‘hope-Obama’).

Participant age and familiarity with targets

Since many real-world people behind the PNCs and names are older than 30, we analyze optionally provided participant age indications to assess which PNCs and names are familiar across ranges. For this, we calculate the proportion of familiarity for every target, normalize by the total number of participants who provided responses, and show average proportional familiarity in Figure 2.

We observe that PNCs are mostly known among participants between 20 and 40, who also represent the largest proportion of annotators overall. Among participants aged above 40, PNCs are more likely to be known than a given name. This could be due to PNCs referring to well-known politicians such as Angela Merkel being over-represented in comparison to e.g., athletes such as Julian Brandt who are often aged younger (here: 28 yrs) than many politicians.

5 LLMs for Association Generation

To augment our small-scale human norms with synthetic data, we outline and perform pilot experiments for synthetic association generation using LLMs.

Models and Experimental Setup To provide a glance into possibilities and limitations of LLMs for generating associations to PNCs and corresponding names, we use the multilingual instruct fine-tuned LLM Mistral (Mistral-7B-Instruct-v0.2) (Jiang et al., 2023). Since we focus on German compounds, we also explore an LLM which is based on Llama-2 (Touvron et al., 2023) and fine-tuned for German specifically (leo-mistral-hessianai-7b-chat)⁴ to alleviate US-centric bias inherent to English data and improve model capabilities regarding German. We perform baseline experiments in zero-shot and few-shot prompting setups. When experimenting with prompts, we start with input comparable to the instructions for humans⁵ and reformulate prompts e.g., through shortening or simplification. For further details regarding the experimental setup and prompts, we refer to Appendix D.

Results Our findings indicate that (i) re-using instructions for humans does not yield desired outputs, while generation quality improves when reducing original prompt length and syntactic complexity; (ii) in the given settings and across prompts, the inspected LLMs seem to harness event knowledge similarly to humans, but vary wrt. prompt and PNC (cf. Appendix D for details). Further experiments using e.g., automatic prompt optimization tools such as dspy⁶ are warranted to explore LLMs for automatic association generation in more depth.

6 Conclusion

The presented dataset of association norms for personal name compounds and corresponding full person names provides a novel way of investigating the evaluative nature of personal name compounds, complementing previous text-driven approaches: The intertwined PNC–name target setup enables the direct comparison between perceptions of characteristics or events involving specific persons, in

⁴<https://laion.ai/blog/leo-lm/>

⁵Note that we neither expect nor believe that the tested LLMs show any kind of human-like behavior, and we clearly distance ourselves from any anthropomorphism.

⁶<https://github.com/stanfordnlp/dspy>

relation to perceptions of the persons per se. In this vein, associations and properties of associations will provide salient materials, and this novel perspective might also be useful for further lexical-semantic tasks with dependencies between target sets, where representation and interpretation of one set require those of the corresponding other set.

Limitations

While the presented association norms comprise a nascent amount of information on the characteristics and mental representation of the given target stimuli and thus a valuable testbed for PNC evaluation, we acknowledge the following limitations resulting from experiment design such as screening participants only regarding German as a native language. Since we provide available demographic information regarding participants, future work using the collected norms is equipped to balance associations, e.g., regarding age, self-reported participant gender, and ethnicity. This might however be connected with a potentially significant loss of data. Moreover, we did not require participants to reveal information regarding their political position or party membership. Hence, the presented associations are not suitable for analyzing the relationship between a PNC from the domain *politics*, the corresponding reference person, a provided response, and the political position of a participant.

Our experiments on LLMs for generating synthetic associations for comparison or augmentation of human associations are intended to provide first insights into potential possibilities and shortcomings. To draw reliable conclusions, we acknowledge that more in-depth experiments including a thorough human evaluation are required, possibly covering a larger range of models.

Ethics Section

In the context of the conducted experiment, we collected free associations from human participants. For this, the participants were provided an informed consent declaration with the name and the contact of the principal investigators; the title, purpose and procedure of the study; risks, benefits and compensation for participating in the study; confirmation of confidential anonymous data handling; and confirmation that participation in the study is voluntary. The informed consent declaration was signed by the participants before taking part in the study.

Participants were provided written guidelines including an example trial. In case of questions, participants had the option to contact the authors of the paper. The experiment task was carried out online in a remote setting using Prolific and Google Tables. Participants received compensation according to the authors' country minimum wage regulations for their effort.

Acknowledgments

This research was supported by the Hanns Seidel Foundation and the DFG Research Grant SCHU 2580/4-1 (*MUDCAT – Multimodal Dimensions and Computational Applications of Abstractness*). We also thank Laura Zeidler for the manual post-processing of the participant association responses (see Section 3), and the reviewers for useful feedback and suggestions.

References

- Milena Belosevic. 2022. [Veggie-Renate und Merci-Jens: Semantik und Pragmatik onymischer Personennamenkomposita](#). *Zeitschrift für germanistische Linguistik*, 50(2):289–319.
- Milena Belosevic and Sabine Arndt-Lappe. 2021. Merci-Jens and Lösch-Leyen. The Semantics of Personal Name Compounds in German. In *Third International Symposium of Morphology*, page 28, Toulouse, France.
- Simon de Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "Small World of Words" English Word Association Norms for over 12,000 Cue Words. *Behavior Research Methods*, 51:987–1006.
- Annerose Eichel, Tana Deeg, André Blessing, Milena Belosevic, Sabine Arndt-Lappe, and Sabine Schulte im Walde. 2024. Willkommens-Merkel, Chaos-Johnson, and Tore-Klose: Modeling the Evaluative Meaning of German Personal Name Compounds. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 17637–17650, Torino, Italy.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- George R. Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. [An Associative Thesaurus of English and its Computer Analysis](#). In *The Computer and Literary Studies*. Edinburgh University Press.
- Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida Free Association, Rhyme, and Word Fragment Norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.
- Wallace A. Russell and O.R. Meseck. 1959. Der Einfluss der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 6:191–211.
- Sabine Schulte im Walde and Susanne Borgwaldt. 2015. Association Norms for German Noun Compounds and their Constituents. *Behavior Research Methods*, 47(4):1199–1221.
- Sabine Schulte im Walde and Alissa Melinger. 2008. An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):89–128.
- Sabine Schulte im Walde, Alissa Melinger, Michael Roth, and Andrea Weber. 2008. An Empirical Characterisation of Response Types in German Association Norms. *Research on Language and Computation*, 6(2):205–238.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiofu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *Preprint*, arXiv:2307.09288.

A Details of Pilot Study

We perform a pilot study to assess how familiar participants are with the public figures behind the PNCs, and we collect their feedback regarding study setup and target presentation. In more detail, participants were first asked whether they are familiar with a PNC such as *Rolex-Chebli* ('Rolex-Chebli') and the corresponding name such as Sawsan Chebli. In case of familiarity, participants provided 3 to 5 associate responses to a given PNC or name in form of a word or a phrase and rated their personal perception of the given public figure on a scale from 0 (positive) to 5 (negative). We collected an average of 5 annotations per name stimulus and 2 annotations per PNC stimulus provided by 14 participants.

Participants were provided written guidelines including an example question with potential responses. In case of questions, participants had the option to contact the authors of the paper. The study was carried out online in a remote setting using Google Forms and Google Tables. We recruited 13 German-speaking and 1 near-native German speaking participant. The annotation task was completed by 2 authors and 12 externally recruited participants who have no connection to any of the authors’ affiliations. External participants could choose between donating their participation or receiving minimum wage compensation for their effort according to Germany’s minimum wage regulations. 7 and 4 participants donated and were compensated for their participation, respectively.

B Details of Main Study Setup

We collect responses through an online setup using Prolific and Google Forms. In Figure 3, the introductory text to the experiment is shown. In this case, the target stimuli are PNCs. We use a comparable introduction for the experiment part where responses are collected for name targets. As illustrated in Figure 4, participants are shown an example trial including an example PNC and sample responses. We use the exact same design for the actual trials.

C Manual Post-Processing

The full list of post-processing guidelines including the list of desired corrections is available from: <https://github.com/AnneroseEichel/pnc-association-norms>. In addition to the provided instructions, the authors and the annotator agreed on the additional correction of word class-specific capitalization, i.e., correcting capitalization if nouns including proper nouns and keeping all other word classes in lower-case.

The annotator is a native speaker of German with a background in computational linguistics and no connection to the presented work. The correction task could be completed fully remote and within 13 days. The required time was 7 hrs. The annotator received compensation according to Germany’s regulations regarding minimum wage.

D LLMs for Association Generation

D.1 Experimental Setup

We obtain model implementations through huggingface⁷ with standard parameter settings, and perform inference using 8 Nvidia GeForce RTX A6000 GPUs. For instruction fine-tuning with Mistral, prompts need be surrounded by the special tokens [INST] and [/INST]. In the case of LeoLM, a prompt format can be set. For Mistral, a message can be set to define a model role. We set the message following instructions as used for eliciting associations for humans and provide the same example trial:

```
messages = [ "role": "user", "content": "Hast du schon einmal von dem Spitznamen "Chaos-Johnson" gehört? Wenn ja, schreibe bitte 3-5 kurze Assoziationen auf. Bitte antworte auf Deutsch.", "role": "assistant", "content": "politisches Chaos, Partygate, Unordnung"]
```

When using LeoLM, we prepend each prompt with the recommended system prompt:

```
""""Dies ist eine Unterhaltung zwischen einem intelligenten, hilfsbereitem KI-Assistenten und einem Nutzer. Der Assistent gibt ausführliche, hilfreiche und ehrliche Antworten.""""
```

D.2 Prompting and Results

For brevity, we focus on the following example output generated using LeoLM. When experimenting with prompts, we start with prompts comparable to the instructions formulated for humans. Note that we neither expect nor believe that the tested LLMs show any kind of human-like behavior, and we clearly distance ourselves from any anthropomorphism. Prompts where we ask for associations to the PNC *Tore-Klose* (‘Goal-Klose’) would look like the following example:

```
(p1) "Hast du schon einmal von dem Spitznamen "Tore-Klose" bzw. Miroslav Klose gehört? Wenn du noch nie von dieser Person oder diesem Spitznamen gehört hast oder dir nicht erklären kannst, wie dieser zustande kam, antworte bitte mit "Kenne ich nicht". Wenn du Miroslav Klose und den Spitznamen Tore-Klose kennst: Was fällt dir spontan in Bezug auf "Tore-Klose" ein? Bitte gib 3-5 kurze, mit einem Komma getrennte Assoziationen aus."
```

⁷<https://huggingface.co/>

(p1 generation): [no output generated] Since prompts such as this example did not yield any results in the case of LeoLM, we reformulate prompts, e.g., by shortening the prompt and simplifying syntactic complexity as illustrated in the following example prompts:

(p2) Was fällt dir spontan in Bezug auf den Spitznamen "Tore-Klose" für Miroslav Klose ein? Bitte gib 3-5 kurze, mit einem Komma getrennte Assoziationen aus. Wenn du den Namen nicht kennst, gib bitte "Kenne ich nicht" aus.

(p2 generation): 1. "Kenne ich nicht" 2. "Wer ist das?" 3. "Der hat aber einen seltsamen Namen..." 4. "Ist das eine Art von Pizza?" 5. "Wieso nennen sie ihn "Tore-Klose"?"

(p3) Fallen dir Gründe dafür ein, warum Miroslav Klose der Spitzname "Tore-Klose" gegeben wurde? Bitte gib 3-5 Gründe an. Wenn du den Namen nicht kennst, gib bitte "Kenne ich nicht" aus.

(p3 generation): Miroslav Klose hat den Spitznamen "Tore-Klose" erhalten, weil er während seiner Fußballkarriere für seine unglaubliche Fähigkeit bekannt war, Tore zu erzielen.2. Der Spitzname wurde ihm verliehen, da Klose während seiner Zeit bei Bayern München und in der Nationalmannschaft konsequent hohe Zahlen an Toren pro Spiel erzielte und damit seine Kullanz in erzielten Toren zeigte.3. Ein weiterer Grund für Kloses Spitznamen könnte sein, dass er in jeder seiner Spielzeiten konstant hohe Zahlen an Toren pro Spiel erzielte und damit seine Kullanz in erzielten Toren zeigte.

In contrast to (p1), results can be obtained using (p2) and (p3). While generated associations are non-sensical in the case of (p2), (p3) yields meaningful output. This is, however, not always the case as we observe stark differences between generated associates for specific PNCs with output quality decreasing for more abstract or metaphoric modifiers and PNCs relating to quite specific events such as *Ferkel-Merkel* ('iglet-Merkel') or *Villen-Spahn* ('villa-Spahn').

Studie: Assoziationen zu Spitznamen bekannter Personen




Was ist Deine Aufgabe?

In diesem Experiment geht es um Spitznamen von bekannten Personen aus verschiedenen Bereichen des Alltagslebens (Politik, Sport, etc.), wie z.B. "Chaos-Johnson" in Bezug auf Boris Johnson. Die Personen haben die Spitznamen aus einem bestimmten Grund bekommen und wir möchten von dir wissen, was dir spontan dazu einfällt, **warum diese Person diesen Spitznamen bekommen hat**. Deine Aufgabe ist es daher, **3-5 mögliche Gründe** in einem Wort oder als kurzen Satz aufzulisten, wie z.B. "Premierminister, UK, politisches Chaos, Partygate, Tories" in Bezug auf "Chaos-Johnson".

Zeitaufwand: Das Ausfüllen der Studie dauert ca. 25 Minuten.

Infos und Einverständniserklärung: Bitte lies Dir die Informationen zur Studie durch und entscheide dann, ob Du an der Studie teilnehmen möchtest.

Du hast Fragen oder ein technisches Problem?

Solltest Du nicht weiterkommen, Fragen zur Studie haben oder auf ein technisches Problem stoßen, schreibe uns gerne jederzeit eine E-Mail . Wir melden uns so schnell wie möglich bei dir.

Vielen Dank und viel Spaß!

Figure 3: Screenshot of experiment introduction to collect associate responses to PNC targets.

Beispielfrage

Damit du dir besser vorstellen kannst, wie die Fragen aussehen und was deine Aufgabe ist, zeigen wir dir hier beispielhaft, wie Maxi Muster die Studie durchführen würde.

Nehmen wir an, Maxi Muster hat schon von Boris Johnson und dem Spitznamen "Chaos-Johnson" gehört. Maxi verbindet mit Boris Johnson z.B. die Punkte "Premierminister," "UK" und "politisches Chaos" und nimmt Boris Johnson im Allgemeinen eher negativ wahr.

Boris Johnson: Chaos-Johnson

Hast du schon einmal von dem Spitznamen "Chaos-Johnson" bzw. Boris Johnson * gehört? Wenn du noch nie von dieser Person oder diesem Spitznamen gehört hast oder dir nicht erklären kannst, wie dieser zustande kam, kreuze bitte an "Kenne ich nicht" und gehe zur nächsten Person weiter.

Kenne ich

Kenne ich nicht

Wenn du Boris Johnson und den Spitznamen Chaos-Johnson kennst: Was fällt dir spontan in Bezug auf "Chaos-Johnson" ein? Schreibe bitte 3-5 kurze Assoziationen auf.

Premierminister, UK, politisches Chaos, Partygate, Tories

Wenn du Boris Johnson kennst, wie nimmst du diese Person wahr?

1 2 3 4 5

positiv negativ

Auswahl löschen

Figure 4: Screenshot of an example trial to demonstrate how associate responses to PNC targets and personal judgments regarding the subjective perception of the PNC and person are elicited.

Using Large Language Models to Perform MIPVU-Inspired Automatic Metaphor Detection

Sebastian Reimann and Tatjana Scheffler

Ruhr University Bochum

Department for German Language and Literature

Bochum, Germany

{sebastian.reimann,tatjana.scheffler}@rub.de

Abstract

Automatic metaphor detection has often been inspired by linguistic procedures for manual metaphor identification. In this work, we test how closely the steps required by the Metaphor Identification Procedure VU Amsterdam (MIPVU) can be translated into prompts for generative Large Language Models (LLMs) and how well three commonly used LLMs are able to perform these steps. We find that while the procedure itself can be modeled with only a few compromises, neither language model is able to match the performance of supervised, fine-tuned methods for metaphor detection. All models failed to sufficiently filter out literal examples, where no contrast between the contextual and a more basic or concrete meaning was present. Both versions of LLaMa however signaled interesting potentials in detecting similarities between literal and metaphoric meanings that may be exploited in further work.

1 Introduction

Metaphors, according to the Conceptual Metaphor Theory (CMT) of Lakoff and Johnson (1980), fundamentally shape the way humans perceive the world. Metaphorical language like considering claims to be *indefensible* is, according to CMT, a way how conceptual mappings in human cognition may be observed on the surface. Several procedures have been developed to systematically identify such linguistic metaphors or metaphor-related words (MRWs) in text, the most famous out of which are the Metaphor Identification Procedure (MIP) by the Pragglejaz Group (2007) and its modified version, the Metaphor Identification Procedure VU Amsterdam (MIPVU) of Steen et al. (2010).

The automatic detection of metaphors has already received considerable interest in NLP, with two shared tasks (Leong et al., 2018, 2020) specifically devoted to it. One line of work in automatic metaphor detection is rooted in modeling aspects

of guidelines for manual metaphor identification with the help of (contextualized) language models (Mao et al., 2019; Choi et al., 2021; Babieno et al., 2022; Zhang and Liu, 2022). More recently, efforts were made to integrate the current generation of generative large language models (LLMs) such as GPT (Brown et al., 2020) and LLaMa (Touvron et al., 2023) into annotation processes (Tan et al., 2024). While previous MIP(VU)-motivated approaches to automatic metaphor detection only approximated the instructions of manual metaphor identification methods, prompting LLMs would theoretically allow for the direct instruction of the language models to conduct the steps required by metaphor identification procedures.

Moreover, the analogical reasoning capabilities of LMs have been a recent topic of interest, ranging from verbal, word-level analogies of the type “lawyer:defending :: teacher:educating” (Stevenson et al., 2023) to drawing analogies between high-level messages and more complex narratives (Sourati et al., 2024). Analogy and metaphor are strongly intertwined, since metaphor is often considered a subtype of analogy (Bowdle and Gentner, 2005) and even the most complex type of analogy (Wijesiriwardene et al., 2023). Determining similarities and analogies between two senses is also a key part of MIPVU since, in the MIPVU manual, Steen et al. (2010) explicitly state that two distinct senses are considered metaphor-related if they “capitalize on external or functional resemblances (attributes and relations) between the functions they designate”. Consequently, a detailed evaluation of the performance of LLMs on MIPVU, with a particular focus on the role of similarity, would provide further insights into the reasoning capacities of LLMs on complex analogies.

In this paper, we make the following contributions:

1. We present the first effort to adapt MIPVU for

its application with generative LLMs, which we achieve with only minor tweaks to the procedure.

2. We evaluate state-of-the-art generative LLM families on two large-scale datasets annotated with MIPVU, where LLaMa performed best but none of the models matched the performance of supervised, fine-tuned approaches.
3. We present an error analysis concentrating on the step where the models made the most mistakes. We find that judging the concreteness and distinctness of two senses is a larger problem than reasoning about similarity, where LLaMa showed interesting potential.

2 Previous Work

2.1 MIPVU

The starting point for our approach is the Metaphor Identification Procedure VU Amsterdam (MIPVU) by Steen et al. (2010), which we will outline in this section. After reading and understanding the entire text (Step 1) and dividing the text into lexical units (Step 2), MIPVU asks the annotator on a word by word basis to **identify the contextual meaning** (Step 3a) of each word. This meaning may or may not be found in a dictionary.

After the contextual meaning has been identified, MIPVU requires the annotator to **identify whether a more basic meaning exists** (Step 3b). Here, Steen et al. (2010) ask the annotator to consult the dictionary entry since they stress that a more basic meaning according to their definition is always conventionalized enough to be found in a dictionary. Moreover, they consider a meaning to be *more basic* if it is more concrete, more specific, or more human-oriented.

If such a basic meaning is found, then MIPVU asks the annotator to **decide if the two meanings are sufficiently distinct** and if they are **related by similarity** (Step 3c). As a shortcut for distinctness, Steen et al. (2010) consider two senses with two different numbered sense descriptions in a dictionary always to be sufficiently distinct. If only one sense description is available or if the contextual meaning is not represented in the dictionary (such as in the case of novel metaphors), this decision is up to the annotator.

Similarity is described by Steen et al. (2010) as “sharing external or functional resemblances (attributes and relations)”. They stress that similarity

distinguishes metaphor from metonymies like *The White House* for the US government, which expresses a part-whole relationship. If the two senses are similar, then the word should be marked as an MRW. Additionally, MIPVU leaves room for special cases, like MRWs that are part of direct comparisons (*[...]like an eagle*; “direct MRWs”), the replacement of MRWs by pronouns, personifications and borderline cases.

2.2 Automatic Metaphor Identification inspired by MIP(VU)

Mao et al. (2019) aimed to model MIP and Selectional Preference Violations (SPV) by Wilks (1975), which detect metaphors through clashes with their context. The MIP model uses GloVe and ELMo embeddings and a BiLSTM, whose hidden states represent the contextual meaning. The basic meaning is represented by the GloVe embedding only and a concatenation of contextual and basic meaning representation serves as input to a classifier. SPV is modeled with the same architecture and a concatenation of left- and right-context representations. Both models achieved F1-scores of around 74 points on the VUA18 dataset (Leong et al., 2018) and outperformed other metaphor identification approaches at the time.

Inspired by Mao et al. (2019), Choi et al. (2021) developed MeIBERT, which aims to model MIP and SPV with the help of contextual BERT embeddings. It uses two encoders, one for the entire sentence, and one for the word in isolation. MeIBERT models MIP through a concatenation of the contextual embedding and the embedding of the word in isolation. SPV, on the other hand, is imitated through a concatenation of the sentence embedding and the contextual word embedding. The output of the SPV and MIP layers is concatenated and fed into a linear classifier. Evaluated on VUA18, it outperformed Mao et al. (2019) with an F1-score of 78.5. On VUA20 (Leong et al., 2020), it achieved an F1-score of 73.9.

Several other authors presented further improvements of MeIBERT, mainly concentrating on the representation of the basic meaning. MisRoBERTaWiLDE (Babieno et al., 2022) uses a Wiktionary entry to represent the basic meaning, MisNet (Zhang and Liu, 2022) encodes an example sentence from the first sense in the dictionary entry of the target word and uses its contextual word embedding as basic meaning. Li et al. (2023) search the training set for non-metaphoric literal examples,

encode them and average the obtained embeddings to represent the basic meaning.

2.3 LLMs and Metaphor

One early contribution to evaluate the metaphor understanding of generative LLMs was made by Wachowiak and Gromann (2023). They asked GPT-3 to provide the source for examples from the Master Metaphor List¹ and the English and Spanish sections of the LCC Corpus (Mohler et al., 2016) with prompts that include the target domain of the metaphor and completed source-target mappings. They also included non-metaphoric examples from VUA to test if the model is able to distinguish metaphoric and non-metaphoric examples. For the simple sentences from the Master Metaphor List, the model predicted the correct source domain with an accuracy of 81.33%, which however went down drastically for the more complex LCC examples. Moreover, the non-metaphoric examples were only singled out with an accuracy of 42.11%.

Schuster and Markert (2023) investigated cross-lingual detection of metaphoric adjective-noun pairs. They compared the zero-shot cross-lingual transfer performance of various BERT and fastText-based classifiers with the performance of ChatGPT when given various prompts. Here, ChatGPT performed best when including the MIP guidelines in the prompt. However, already with little target language data, smaller models were outperforming ChatGPT.

Chen et al. (2024) extended automatic detection with an additional reasoning task, where the models are asked to justify why they considered a token literal or metaphoric. They prompted two variants of LLaMA3 (8B and 70B), Gemma-7B, and ChatGPT3.5 on detecting metaphoric tokens and reasoning. The metaphor detection performance mostly falls short in comparison to supervised approaches, particularly on VUA, where they only achieve F1-scores between 27 and 44 points on the binary metaphor detection.

TSI by Tian et al. (2024) represents the most elaborate approach to metaphor identification to date. They prompted GPT-3.5 with a series of questions inspired by CMT, MIP and SPV and filled knowledge graphs with the answers. If the graph surpasses a final comparison with the ideal knowledge graph pattern, the example is labeled as metaphoric. Their CMT approach performed best

¹<https://www.lang.osaka-u.ac.jp/moto/MasterMetaphorList/metaphors/index.html>

sugi-

and even outperformed several fine-tuned BERT approaches with an F1 of 82.59 on the MOH-X dataset and 66.07 on the TroFi dataset, with MIP ranking second with F1-scores of 79.39 and 65.60, respectively.

The approach of Boisson et al. (2025) also extracted entire source-target mappings with generative LLMs. They represented metaphors as mappings involving two concept terms from each a source and a target domain and aimed to extract such mappings from a collection of novels, poems, songs and speeches. For this, they provided LLMs (GPT-4, Llama-3 and Mixtral) with prompts containing a text and one of the four terms representing the different concepts. They reported mostly satisfactory accuracies, often over 60%.

3 Experiments

3.1 Adapting MIPVU for LLMs

In the previous section, we have already seen two approaches that integrated MIP(VU) into LLM prompts and achieved satisfactory performance on a small amount of data. However, in both Schuster and Markert (2023) and Tian et al. (2024) the metaphor identification procedure was implemented in a simple and rather counterintuitive way. Schuster and Markert (2023) only provided the model with the respective adjective-noun-pair without context, and Tian et al. (2024) left out similarity as a criterion.

For our application of MIPVU, we aim to replicate the steps outlined in Section 2.1 as closely as possible. We, however, do not take extra steps for any of the aforementioned special cases. We already deal with tokenized text, thus no prompts for the steps 1 and 2 are needed. Step 3a, identifying the contextual meaning, can be achieved in a straightforward manner by simply providing the LLM with the word in question and the text containing the word and prompting it to provide the meaning of the word in the given context.

Step 3b, the identification of a more basic meaning, on the other hand, is more complex. It is first necessary to decide on a resource of possible senses. For this, we ask the model to provide us with an entire dictionary entry for the word in question. After extracting the senses, we additionally prompt the model to identify if any of the present senses can be considered more concrete than the contextual meaning.

Here, we needed to make compromises. MIPVU

technically asks for *a more basic meaning* instead of *the most basic meaning*. Thus, in theory, we would need to conduct the next steps for all potential meanings that fulfill the “more basic” criterion, which would be very resource-demanding. Consequently, we ask the model to only extract the most concrete candidate and proceed with this as the more basic meaning. If no basic meaning is available, the word is considered non-metaphoric.

If the model identified a more basic meaning, then the two meanings need to be checked for sufficient distinctness (first part of step 3c). We simply ask the model whether the contextual and more basic meanings refer to the same concept. If yes, the example is considered non-metaphoric.

Otherwise, we proceed with the crucial question on similarity between the two senses (second part of step 3c). For this, we ask the model if the two senses in question share aspects, functions or features (the criteria for “similarity” as outlined by Steen et al. (2010)) and for a short explanation. We test a zero-shot prompt that only asks for similarity according to Steen et al. (2010) as well as a one-shot prompt that illustrates similarity according to Steen et al. (2010) with examples 1 and 2, the two senses for *journey* provided in the Longman Dictionary of Contemporary English (Longman, 2023).

- (1) an occasion when you travel from one place to another, especially over a long distance
- (2) a long and often difficult process by which someone or something changes and develops

However, given that word-sense disambiguation (WSD) represents an NLP task which non-generative language models perform well (Bevilacqua et al., 2021), we also test the combination of a BERT-based WSD model and a generative LLMs. For this, we used the fine-tuned model presented in Yap et al. (2020). Here, for step 3b of MIPVU, we thus extract the sense keys and their respective glosses from WordNet (Miller, 1994) instead of generating an entire dictionary entry and then ask the LLM to provide us with the more basic meaning among the extracted senses.

The fine-tuned WSD model comes into play when checking for sufficient distinctness. We have the WSD model predict the sense. If this predicted sense by the WSD model and the predicted more basic meaning by the LLM are different, then they are considered sufficiently distinct. The final question for similarity is then asked in the same way as

in the procedure without WSD. An overview over all prompts that we used is provided in Appendix B.

3.2 Models

We evaluate three commonly used families of LLMs in our experiments: LLaMa, Mistral and GPT. For LLaMa, we specifically use the 8B and 70B instruction-tuned versions of LLaMa 3.1. For Mistral, we use the also instruction-tuned Mistral-Small-Instruct-2409 with 22 billion parameters. We obtain the LLaMa and Mistral models via HuggingFace (Wolf et al., 2020). For GPT, given financial constraints, we only use the lightweight GPT-4o-mini, which we access via the OpenAI API. We used the default hyperparameters of all the models. We ran the 8B version of LLaMa 3.1 on NVIDIA A40 GPUs, the 70B version of LLaMa 3.1 on NVIDIA H100 SXM5 GPUs and the Mistral model was run on NVIDIA A30 GPUs.

3.3 Data

For evaluation purposes, we use two larger metaphor datasets where the annotation followed MIPVU very closely. On the one hand, we use the VUA dataset in the version that was used in the 2020 Metaphor Detection Shared Task (Leong et al., 2020) and which is based on the original application of MIPVU to the British National Corpus by Steen et al. (2010).

Moreover, we use the metaphor dataset of Reimann and Scheffler (2024) (“R&S” in the following), which consists of posts from Christian subreddits annotated for metaphor via MIPVU. Due to the limits of the OpenAI API and financial considerations, we only use a fraction of the test data, namely two reddit threads from R&S and one fragment from VUA in the experiments involving GPT-4o-mini. Table 1 presents an overview of the data that we used.

Dataset	Tokens	MRWs
VUA	22196	3982
VUA (short)	3960	821
R&S	14437	3170
R&S (short)	3562	555

Table 1: Overview of the data.

Model	Setup	R & S				VUA			
		P	R	F1	Acc.	P	R	F1	Acc.
LLaMa 3.1 8B	0-Shot	25	72	38	47	21	59	31	50
	1-Shot	25	79	38	44	21	64	32	53
	0-Shot + WSD	32	47	38	68	32	63	43	64
	1-Shot + WSD	32	63	43	64	28	59	38	65
LLaMa 3.1 70B	0-Shot	26	82	39	44	21	69	33	49
	1-Shot	26	84	40	43	22	70	33	50
	0-Shot + WSD	27	63	38	56	26	68	38	60
	1-Shot + WSD	28	67	39	56	26	68	38	60
GPT-4o-mini	Zero	20	19	20	76	19	22	21	60
	One	20	11	14	79	22	18	20	67
	0-Shot + WSD	18	11	13	78	27	10	15	74
	1-Shot + WSD	17	8	11	79	29	13	18	73
Mistral-Small	Zero	22	96	36	29	19	95	32	27
	One	20	72	32	35	18	74	29	35
	0-Shot + WSD	31	54	40	65	29	55	38	68
	1-Shot + WSD	29	45	35	65	28	49	36	69
MelBERT (Choi et al., 2021)		76	69	72	-	68	60	64	-

Table 2: Precision, recall and F1 for the metaphor class and accuracy on the two datasets for all LLM setups as well as the results of the supervised MelBERT approach reported in Choi et al. (2021) and Reimann and Scheffler (2024) for comparison. Best result for each metric in bold, second best in italics.

4 Results and Error Analysis

Table 2 shows our results. Overall, we can see that neither of the LLM is actually able to achieve satisfactory performance in any setting, with GPT-4o-mini in particular trailing behind the other two models in all metrics except for accuracy. The data is imbalanced (around 80% of tokens non-metaphorical), which means that if the model considers fewer tokens to be MRWs, then will automatically be higher.

A general trend for LLaMa is that model size had a positive impact on recall. The same can be said for providing LLaMa with an example in the last prompt since this resulted in increased recall for almost all setup scenarios. GPT-4o-mini, on the other hand, was not at all able to profit from the inclusion of an example and for the Mistral model, the inclusion of an example in the prompt even had the opposite effect and led to drastic drops in performance.

4.1 False Positives

One more particularly striking pattern in Table 2 is the fact that both LLaMa and especially Mistral produce a large amount of false positives, i.e. non-metaphoric examples that were considered to be MRWs, as seen in the low values for precision. The way MIPVU functions may provide an explanation:

the initial steps of MIPVU are designed to weed out cases where no meaning can be considered “more basic” according to the procedure or where the contextual and the more basic meaning are too similar. This ensures that in the last step, only two distinct meanings are compared. However, when the model fails to complete the previous steps correctly, it will eventually compare two meanings that refer to the same concept to each other and, as a logical choice, understandably consider them to be similar.

The example of *conversation* in Table 3 illustrates such cases, which represent the most common type of false positive. Here, both the explanation of the contextual meaning and the “more basic” meaning refer to the same concept; however, the model did not judge the meanings to be distinct. They were then inevitably judged as related by similarity and received the label *MRW*.

However, some cases of false positives were also similar to the example of *night* in Table 3. Here, the model already misinterpreted the contextual meaning of *night* (which was used literally as the time of day in the original sentence) as nonliteral. Based on this propagated error, the meanings were then judged as sufficiently distinct and, consequently, as *MRW*. The wider context for both examples in Table 3 is provided in Appendix A.

Word	Contextual	More Basic	Similarity?
conversation	In the context of this post, the word "conversation" refers to a meaningful and impactful exchange of thoughts, ideas, and emotions between two individuals, often leading to a deeper understanding, personal growth, and a sense of connection or revelation.	a talk between two or more people, especially one that is friendly and informal	The similarity between the two senses is that they both refer to a talk or exchange between two or more people. Sense 1 emphasizes the depth and impact of the conversation, while Sense 2 provides a more general definition that focuses on the formality and tone of the conversation.
night	In this context, the word "night" refers to a period of time, specifically a dark and potentially difficult or vulnerable moment, often symbolizing a state of need or crisis.	The period of time from sunset to sunrise, when it is dark outside.	The similarity between the two senses of 'night' is that they both refer to a period of time, specifically a time when it is dark outside. Sense 1 describes 'night' as a moment of need or crisis, while Sense 2 defines it as the time from sunset to sunrise.

Table 3: Examples for false positives, produced by the 8B version of LLaMa.

Replacing the prompt-based judgment on sufficient distinctness with BERT-based word sense disambiguation overall resulted in a stricter application of the *MRW* label. However, for all models, the improvements for precision are much smaller than the drops in recall. This may be because the glosses for WordNet senses may sometimes be not informative enough or two glosses may appear too similar for the model.

- (3) water falling in drops from vapor condensed in the atmosphere
- (4) drops of fresh water that fall as precipitation from clouds

The examples 3 and 4 for *rain* illustrate this. They are the glosses for two different senses in WordNet, however, it may be argued that they denote the same concept. The LLaMa models selected 4 as the more basic meaning and 3 was selected by the WSD model to be the contextual meaning. This led the model to not dismiss the example as metaphoric and in the last step, the meanings were considered similar and thus wrongly labeled as *MRW*.

4.2 False Negatives

This implementation of MIPVU with generative LLMs gives us, in the case of false negatives, the opportunity to track exactly where the decisive error was made. We make use of this to better interpret the results of Table 2. The results of this analysis are provided in Table 4.

Consistently, for both instances of LLaMa, deciding on distinctness appears to be the biggest problem, followed by deciding on similarity. In contrast, the small GPT model already produces the most false positives when prompted to decide on a more

basic meaning. Surprisingly, the model notably produced more false negatives in the distinctness step when evaluated on VUA, compared to the evaluation on R & S. Looking into the most frequent false negatives for VUA, we find a wide range of heavily conventionalized MRWs such as *make* among the most common false negatives. For these examples, the contextual and more basic meaning may appear too similar, which possibly explains why the models considered them to be not sufficiently distinct.

Mistral without the added WSD model on the other hand in general produced not many false negatives, which further highlights that it was not sufficiently strict to carry out MIPVU. A striking result, however, is that when provided with the WordNet glosses, the model wrongly sorted many MRWs out when checking for a more basic meaning. This may again be because the WordNet glosses were shorter and less informative than the generated definitions.

Table 4 also further exemplifies the improvement in recall for LLaMa. Here, we can see that for all LLaMa models, the number of wrong classifications as non-metaphorical in the last step drops notably in the one shot scenario. This suggests that LLaMa indeed was able to better reason on metaphoric similarity when provided with an example.

This impression is also confirmed by looking at the first two examples in Table 5, which compares the output of several models for the zero- and few-shot prompts for the last MIPVU step. The example of *sheep*, by the 70B version of LLaMa is particularly interesting as the model, when asked to describe the contextual meaning of the word, already mentions that it is used metaphorically. However,

Model	R & S					VUA			
		Basic	Dist.	Sim. (0-shot)	Sim. (1-shot)	Basic	Dist.	Sim. (0-shot)	Sim. (1-shot)
LLaMa-3.1-8B-Instruct	w/o WSD	193	366	250	53	170	1202	277	45
	w/ WSD	153	943	551	31	139	1203	574	45
LLaMa-3.1-70B-Instruct	w/o WSD	85	304	163	86	372	743	130	99
	w/ WSD	60	766	264	177	97	769	172	176
GPT-4o-mini	w/o WSD	345	83	19	66	526	92	5	55
	w/ WSD	417	64	9	24	549	106	3	27
Mistral-Small-Instruct-2409	w/o WSD	53	71	6	769	106	105	6	805
	w/ WSD	1405	0	8	296	1568	0	217	6

Table 4: Number of MRWs wrongly considered to be non-metaphorical across different steps.

when asked to reason about similarities between the contextual and basic usage of *sheep* in a zero-shot manner, it denies the question. A relation by similarity is often considered a defining feature of metaphor (Steen et al., 2010), which makes the model output thus contradictory. The model given the one-shot similarity prompt, on the other hand, answers with *yes* and provides extensive reasoning on the similarities between the metaphorical *sheep* in the sense of believers and the animal. The metaphorical example of *light* in Table 5 also illustrates how the few-shot prompt had the completely opposite effect for Mistral as it reasoned correctly without the example senses. However, when provided with an example, it still understood that one sense is abstract and one physical but failed to point out the similarity. The context for all examples discussed in Table 5 is provided in Appendix A.

4.3 Deliberate and Direct Metaphor

Finally, we look at different subtypes of metaphors and see if they were harder to detect for the models. For this, we use the more detailed annotations of Reimann and Scheffler (2024), which also indicate if an MRW is direct or indirect and if it is used in a potentially deliberate way (i.e. used "as metaphor") according to the Deliberate Metaphor Identification Procedure (DMIP) by Reijnierse et al. (2018). In addition to direct metaphor, this also includes novel metaphor and extended metaphor. Table 6 shows the recall for all potentially deliberate metaphors and direct metaphors in the most successful scenario with respect to the overall recall (no external WSD model and a one-shot prompt for similarity).

Interestingly, finding potentially deliberate

metaphors did not pose particular problems for the LLaMa and Mistral models, which contrasts with the results of supervised BERT-based approaches, that struggled with several types of potentially deliberate metaphors. This also extends to direct metaphor in particular, which is surprising at first glance, given that our prompts do not cover direct metaphor, as outlined in Section 2.1.

- (5) I just have to learn to wait on the lord as I feel like a *car* that was *Idling* for so long that my *engine* is just *ready to go* , and I need to be calm and wait on the lord .
- (6) In this context, a "car" is used metaphorically to describe a person's eagerness and readiness to take action, likening their built-up energy and anticipation to a car's engine idling and ready to move.

The direct MRWs in example 5 are in bold. When asked for the contextual meaning of the word *car*, the large LLaMa model provided the explanation given in Example 6. Here, although not explicitly asked, the model already generates an explanation for the direct metaphor. This makes the contrast between contextual and basic meaning in further comparisons apparent and demonstrates why direct MRWs do not seem to pose problems for this procedure.

5 Discussion

Putting our results into the wider context of automatic metaphor detection, we can, on the one hand, see that they matched or outperformed for the most part the approach of Chen et al. (2024), which also

MRW	Contextual	More Basic		Expl. 0-Shot	Expl. 1-Shot
deeper	In the context of the given post, "deeper" refers to a more profound or underlying level of understanding, issue, or problem, often requiring a more nuanced or complex analysis.	very and very low	wide not shallow	There is no similarity between the two senses. The first sense refers to a more in-depth or thorough exploration or understanding, often implying a greater level of insight or knowledge. The second sense refers to the physical characteristics of a river or lake, specifically its width and depth.	Sense 1 and Sense 2 are similar because in both senses, the word "deeper" refers to a greater extent or dimension, whether it's a more in-depth understanding or a wider and less shallow body of water.
sheep	In this context, "sheep" is a metaphorical representation of a lost or misguided individual who has strayed from their spiritual path and is being called back to a higher power or sense of self-worth.	a farm animal that produces wool and is often kept for its meat		Sense 1 is a metaphorical representation, while Sense 2 refers to the literal animal. There is no similarity in their meanings, as one is a figurative concept and the other is a physical entity.	Sense 1 and Sense 2 are similar because in both senses, the term 'sheep' refers to an entity that can be guided or directed, whether it's a farm animal being herded or a metaphorical individual being called back to a spiritual path.
light	"Light" in this context refers to one's positive influence, joy, or faith that can be shared with others regardless of circumstances.	the natural brightness that comes from the sun, moon, or stars		Yes, both senses of "light" involve the idea of illumination and positivity. In Sense 1, it's about illuminating someone's spirit or life with positivity, while in Sense 2, it's about illuminating the physical environment with natural brightness.	No. Sense 1 refers to an abstract, emotional concept, while Sense 2 refers to a physical phenomenon. They do not share any aspects, functions, or features.

Table 5: Examples wrongly considered to be non-metaphorical by the 8B (first), 70B (second) versions of LLaMa and Mistral (third) when prompted in a zero-shot fashion and considered to be MRWs when given an example in the prompt on similarity.

Model	R (Pot. Delib.)	R (Direct)
LlaMa 3.1 8B	78	72
LlaMa 3.1 70B	87	88
GPT-4o-mini	10	13
Mistral-Small	67	68

Table 6: Recall for potentially deliberate and direct metaphors when using the models in the scenario involving a few-shot prompt and without WSD.

used generative LLMs and which evaluated all of the VUA data. The approaches of Schuster and Markert (2023) and Tian et al. (2024) are much harder to compare with ours, as they used much smaller and more balanced data sets.

However, it also becomes clear that our LLMs emulating MIPVU massively underperform previous approaches with fine-tuned variants of BERT. Our results demonstrate that, as of now, identifying metaphors with a series of prompts inspired by the steps of MIPVU is not a realistic alternative for automatic metaphor detection. In order to cor-

rectly identify metaphor, the last step of MIPVU is heavily dependent on the correct implication of the previous two steps and propagated errors may result in heavy overuse of the *MRW* label.

Moreover, the models from the three families performed wildly differently. This makes it harder to draw final conclusions about the suitability of LLMs for this task. GPT-4o-mini overall failed, with neither recall nor precision achieving satisfactory results, and Mistral was unable to sufficiently filter out negative examples. However, the results of LLaMa, while far from perfect, show encouraging tendencies.

Neither LLaMa version was able to filter out negative examples in a satisfactory way due to the aforementioned reasons, even though the picture here is more nuanced in comparison to Mistral. However, the improvements for both versions when provided with an example and a qualitative inspection of the output suggested some capacities to reason on similarity and analogy in order to correctly answer the last step of MIPVU. Paired with better identification of the basic meaning and better filtering of negative examples due to insufficient distinctness, this may present potential to even-

tually be able to conduct MIPVU via generative LLMs.

6 Conclusion and Future Work

We replicated MIPVU with LLMs as closely as possible in two different ways. On the one hand, with a series of prompts and only using the respective LLM and, on the other hand, via a combination with WordNet as an external resource. We evaluated four LLMs from three popular model families on two large, realistic datasets that were manually annotated via MIPVU and conducted an extensive evaluation and error analysis of the output.

We found that our approach achieves competitive performance with other LLM-based approaches on the VUA data with LLaMa. However, it still falls short in comparison to models that were specifically fine-tuned on the metaphor detection task in either setup. In general, the models also differed in their behavior, making interpretation and final conclusions difficult. However, one point that all models had in common were struggles to select a basic sense and to decide on distinctness of word senses. In contrast to the others, the LLaMa models performed satisfactory on the last question of analogical similarity between literal and contextual word senses, especially when provided with an example mapping.

For future work, we suggest further investigation into the steps that were problematic for the models, namely perceiving concreteness and distinctness of word senses. For this, and for a better evaluation of the suitability of LLMs for MIPVU in general, we would need more gold labels and human judgments for all the steps of MIPVU instead of only the final labels. Our prompt and MIP(VU) in general is also relatively vague on its definition of *more concrete*. Thus, extending the concreteness prompt with a more comprehensive definition of concreteness or concreteness ratings such as Brysbaert et al. (2014) would be worth trying out. Moreover, as the one-shot prompt already led to some improvements, we suggest providing more examples and testing if different examples for the few-shot experiment would lead to different results.

Finally, we suggest exploring the analogical reasoning capabilities of LLaMa further in the context of automatic metaphor identification: In this work, we used a relatively simple prompt that asks for *similarity* according to the definition of Steen et al. (2010) only. However, the output for this question

could be evaluated in an even more systematic manner by, for example, explicitly providing the model with source and target domain terms or, for the last MIPVU question, additionally asking it to identify source and target terms like Boisson et al. (2025) did.

7 Acknowledgements

We thank the Paderborn Center for Parallel Computing (PC2) and HPC@RUB for granting us compute time on their HPC clusters and we thank the anonymous reviewers for their valuable comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958

References

- Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. [Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions](#). *Applied Sciences*, 12(4).
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. [Recent Trends in Word Sense Disambiguation: A Survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization.
- Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2025. [Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6692–6704, Abu Dhabi, UAE. Association for Computational Linguistics.
- Brian F Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological review*, 112(1):193.
- 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](#). *Preprint*, arXiv:2005.14165.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. [Concreteness ratings for 40 thousand](#)

- generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.
- Puli Chen, Cheng Yang, and Qingbao Huang. 2024. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5850–5860, Miami, Florida, USA. Association for Computational Linguistics.
- Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Univ. of Chicago Press, Chicago [u.a.].
- Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xi-anyang Chen. 2020. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.
- Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A Report on the 2018 VUA Metaphor Detection Shared Task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.
- Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.
- Longman. 2023. *Longman Dictionary of Contemporary English (Online Edition)*. Pearson Education Limited.
- Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.
- George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).
- Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.
- W. Gudrun Reijniere, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147.
- Sebastian Reimann and Tatjana Scheffler. 2024. Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.
- Jakob Schuster and Katja Markert. 2023. Nut-cracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.
- Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.
- Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Claire E. Stevenson, Mathilde ter Veen, Rochelle Choenni, Han L. J. van der Maas, and Ekaterina Shutova. 2023. Do large language models solve verbal analogies like children do? *Preprint*, arXiv:2310.20384.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.
- Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. 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 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

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](#). *Preprint*, arXiv:2302.13971.

Lennart Wachowiak and Dagmar Gromann. 2023. [Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. [ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. [Adapting BERT for word sense disambiguation with gloss selection objective and example sentences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.

Shenglong Zhang and Ying Liu. 2022. [Metaphor detection via linguistics enhanced Siamese network](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Posts

Table 7 provides the post containing the examples presented in Tables 3 and 5.

B Used Prompts

We provide the LLM prompts in Table 8.

I recently spent a night of "homelessness" the details of which do n't matter. All I know is walking barefoot, alone, in the **night**, in need of help, during christmas. The only person that found me in the middle of the night or would help, was a homeless man.

He gave me his shoes, and a sweater, he walked with me for miles, and talked with me. He gave me aid and comfort, and told me truth about my life that he could n't of known. I thought I was insane, I was n't. (because the things he told me he could n't know) I have more peace now than I ever had in my life and I just need to figure out how to help. I also know I am a Christian and can never deny Christ in my life, I know every blessing I have is not mine, its simply for me to use to help others. I went back to church, and told my tale to the pastor, and that very day I went out to help again and was opposed. Opposed in just the right way to make me fall back into my depression and hide away from the world. The only reason I did n't was a christian brother was there with me and saw everything, he prayed and told me exactly what I needed to hear. because what the Homeless person screamed at me, was exactly every doubt I had about going and telling the pastor my tale, she hit every insecurity I had about the **conversation**. My friend simply said at the time "thats how you know your doing the right thing" So, I have discovered I have a heart for the homeless.

So from there, my standard for evaluating anything became, "Does this facilitate wholeness in someone's life?" That's the lens I use to view LGBT issues, and after breaking away from some of the Evangelical propoganda, I've realized that whole, healthy homosexual relationships are 100% possible and not at all uncommon.

I believe there are also examples where transitioning is the healthiest choice for someone, but I feel like there is a ton of **deeper** rooted identity dysfunction present within the ideologies accompanying that "community / movement / not - sure - the - right - term", but I have zero interest in dictating anyone's behavior or telling anyone how they should or should n't live their lives.

I feel like our purpose as people is just to love ourselves and everyone else, live whole, healthy, happy lives, and help those around us.

Sadly marriage wo n't solve your issues with self-worth which is the root of your hook up lifestyle. You're looking for value in relationships when Jesus has already paid the highest price for you. He loves you and wants you to return to Him. You're a daughter in His eyes. Go read Luke 15. You're the lost **sheep** / coin / son. God wants you to come back to Him. Your created value comes from Him alone, not what others think of you. The guys you're hooking up with you just want what you can give them and see no value in you. To them you're a means to an end, which is why you may find value in the moment but regret it after. It's a vicious cycle. At least you recognize you ca n't have one or the other.

But thanks be to His Word, first spoken, then written, and then affirmed, confirmed, and fulfilled by The Living Word, one can truly state that they know "of" our Father Love

The Adam and Eve, the only 2 flesh humans that had a clear recollection of their true celestial origin / heritage in The **Light**

Table 7: Posts containing the examples discussed in Tables 3 and 5 with the respective MRWs in bold.

Step	Prompt
get the contextual meaning	In one sentence, describe the meaning of the given word in the context of the given post as general as possible. Word: [WORD], Post: [POST]
generate the dictionary entry	Write a dictionary entry in the style of the Longman Dictionary of Contemporary English that provides all possible senses of the given word with the given Part of Speech. Word: [WORD], Part-of-Speech: [POS]
decide on basic meaning	Decide if any of the dictionary senses can be considered more concrete than the example definition. If yes, output 'yes' and in a new line provide only the respective sense. If no, then just provide 'no'. Dictionary Senses: [SENSES]
decide on sufficient distinctness	Do the two senses express the same meaning or is sense 2 only a more specific version of sense 1? Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONT. SENSE], Sense 2: [MORE BASIC SENSE]
decide on similarity (0-Shot)	Can you see a similarity between the senses 1 and 2? 'Similarity' means that the two senses denote distinct concepts that share certain aspects, functions or features. Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONT. SENSE], Sense 2: [MORE BASIC SENSE]
decide on similarity (1-Shot)	Can you see a similarity between the senses 1 and 2? 'Similarity' may also mean that the two senses denote distinct concepts that share certain aspects, functions or features. The following example for the word 'journey' illustrates this: journey: Sense 1: "an occasion when you travel from one place to another, especially over a long distance" Sense 2: "a long and often difficult process by which someone or something changes and develops" Answer: Yes. Sense 1 and Sense 2 are similar because in both senses refer to something that takes a longer period of time. Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONT. SENSE], Sense 2: [MORE BASIC SENSE]

Table 8: Overview over the used prompts for each step.

Modeling Background Knowledge with Frame Semantics for Fine-grained Sentiment Classification

Muhammad Okky Ibrohim¹, Valerio Basile¹,
Danilo Croce², Cristina Bosco¹, Roberto Basili²

¹Dipartimento di Informatica, Università degli Studi di Torino, Italy

²Department of Enterprise Engineering University of Rome, Tor Vergata Roma, Italy

¹{muhammadokky.ibrohim, valerio.basile, cristina.bosco}@unito.it

²{croce, basili}@info.uniroma2.it

Abstract

Few-shot learning via in-context learning (ICL) is widely used in NLP, but its effectiveness is highly sensitive to example selection, often leading to unstable performance. To address this, we introduce BACKGEN, a framework for generating structured Background Knowledge (BK) as an alternative to instance-based prompting. Our approach leverages Frame Semantics to uncover recurring conceptual patterns across data instances, clustering examples based on shared event structures and semantic roles. These patterns are then synthesized into generalized knowledge statements using a large language model (LLM) and injected into prompts to support contextual reasoning beyond surface-level cues. We apply BACKGEN to Sentiment Phrase Classification (SPC), a task where polarity judgments frequently depend on implicit commonsense knowledge. In this setting, BK serves as an abstract representation of prototypical scenarios, enabling schematic generalization to help the model perform analogical reasoning by mapping new inputs onto generalized event structures. Experimental results with Mistral-7B and Llama3-8B demonstrate that BK-based prompting consistently outperforms standard few-shot approaches, achieving up to 29.94% error reduction¹.

1 Introduction

Few-shot learning has become a standard approach in NLP, enabling models to generalize from limited labeled data. In particular, *in-context learning* (ICL) (Brown et al., 2020) allows large language models (LLMs) to perform tasks without parameter updates, relying instead on a well-designed prompt that includes relevant examples (Dong et al., 2024; Liu et al., 2022a; Lu et al., 2022; Wu et al., 2023). However, ICL suffers from high variance due to its sensitivity to example selection (Zhang et al., 2022;

¹<https://github.com/crux82/BackGen>

Task: Determine the polarity (either 'positive' or 'negative') of the target phrase.

Input:

- Text: "The government phases out fossil fuels."
- Target Phrase: "phases out fossil fuels"

Model Output: negative

(a) Example of zero-shot SPC prompt.

Task: Determine the polarity (either 'positive' or 'negative') of the target phrase, using background knowledge if helpful.

Input:

- Text: "The government phases out fossil fuels."
- Target Phrase: "phases out fossil fuels"

Background Knowledge:

1. The fact that a public entity wants to remove something related to green initiatives is perceived negatively.
2. Public entities' intention to reduce non-renewable energy sources is seen as a positive step.

Model Output: positive

(b) Example of bk-shot SPC prompt with injected background knowledge.

Figure 1: Examples of prompts used for Sentiment Phrase Classification (SPC). The zero-shot prompt lacks external context, while the BK-injected prompt includes relevant background knowledge that helps disambiguate sentiment. These are abbreviated versions for illustration; full prompt templates are reported in Appendix B.

Köksal et al., 2023; Pecher et al., 2024a). Prior research has attempted to mitigate this issue by selecting examples based on informativeness (Liu et al., 2022a; Liu and Wang, 2023; Köksal et al., 2023), representativeness (Levy et al., 2023), or learnability (Song et al., 2023), but these methods often come at a high computational cost.

A complementary approach is *knowledge prompting*, where explicit background knowledge (BK) replaces example-based selection in prompts, enabling *analogical reasoning* by mapping new in-

puts onto abstracted conceptual patterns derived from semantically related instances.

Prior work has explored using LLM-generated knowledge for commonsense reasoning (Liu et al., 2022b) or integrating structured knowledge from external sources (Baek et al., 2023). In this paper, we hypothesize that background knowledge (BK) can be particularly useful for Sentiment Phrase Classification (SPC), where the goal is to determine the sentiment polarity of a target phrase in a given text. SPC becomes especially challenging when the sentiment of a phrase is context-dependent or ambiguous. Figures 1a and 1b illustrate such a case: given the sentence “*The government phases out fossil fuels*”, the target phrase “*phases out fossil fuels*” might be misclassified as negative in a zero-shot setting (Figure 1a), as “*phases out*” often conveys abandonment. However, in the context of environmental policy, the action of phasing out fossil fuels is typically seen in a positive light. Figure 1b shows how injecting BK into the prompt can guide the model toward the correct interpretation. Statements such as “*public entities’ intention to reduce non-renewable energy sources is seen as a positive step*” help contextualize the sentiment, enabling the model to move beyond surface-level heuristics. This example demonstrates how BK can resolve subtle ambiguities in sentiment interpretation and reinforces our motivation for replacing concrete examples with structured, generalizable knowledge.

ICL typically addresses these issues and mitigates the negative impact of missing context by injecting example sentences into the prompt. However, in tasks involving short texts, the relationship between a support example and the test instance may be weak or even nonexistent, reducing the effectiveness of example-based prompting. Instead, structured BK provides a more reliable alternative, as it captures the higher level generalizations that underpin sentiment-bearing expressions.

In scenarios where we have annotated examples but do not perform fine-tuning, an alternative approach is to transform these examples into structured knowledge statements that generalize beyond individual instances. The goal is to construct a BK repository where each entry captures recurring conceptual patterns that can support multiple examples from the original dataset.

To achieve this, we propose a methodology for clustering similar examples and extracting their underlying commonalities. Instead of selecting in-

stances arbitrarily, we group them based on shared semantic properties and identify the minimal conceptual structure that describes their sentiment polarity in both positive and negative contexts. The clustering process leverages Frame Semantics (Fillmore, 1985), as it provides a structured representation of situations by encoding events, participants, and their relationships. This enables us to generalize beyond lexical choices and focus on the core elements that shape sentiment interpretation. Once structured, the extracted knowledge is verbalized using an LLM, producing natural language statements that encapsulate the core sentiment-related concepts within each cluster. These statements are then injected into the prompt as BK, replacing explicit few-shot examples. This approach aims at mitigating performance variance due to instance selection (Zhang et al., 2022) and enhances the model’s ability to reason over sentiment phrases in context, particularly in ambiguous cases.

Experiments with two LLMs show that integrating BK into prompts systematically improves performance over zero-shot and few-shot learning, yielding a 26-29% error reduction. These results confirm that structured BK enhances sentiment classification by providing essential context and reducing misinterpretations.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes the proposed methodology, Section 4 presents experiments and results, and Section 6 concludes with future directions.

2 Related Works

Few-shot learning via ICL. The ICL (Brown et al., 2020) has an essential role in solving many NLP tasks as it allows the LLM to learn some examples via specific template (then, this technique is called as few-shot prompting) without updating the model parameters (Dong et al., 2024; Liu et al., 2022a; Lu et al., 2022; Wu et al., 2023). Unfortunately, the classical few-shot prompting is very sensitive to sample selection strategies (Zhang et al., 2022; Köksal et al., 2023; Pecher et al., 2024a). Despite many techniques that have been introduced to solve that problem (Liu et al., 2022a; Liu and Wang, 2023; Köksal et al., 2023; Levy et al., 2023; Song et al., 2023; Pecher et al., 2024b), most of them come at a high computational cost since the procedure to retrieve complex examples should be run for each instance, leading to a new ICL approach

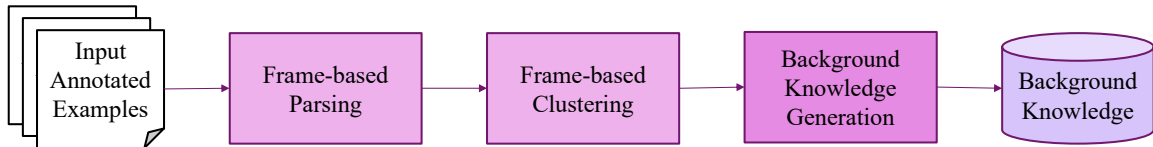


Figure 2: The BACKGEN pipeline.

called knowledge prompting.

Knowledge Prompting. A new approach of ICL was introduced to inject knowledge to the prompt where the knowledge is retrieved from a particular source or generated based on the instance. Guu et al. (2020) and Lewis et al. (2020) inject documents to LLM so that the model can retrieve answers from them. Baek et al. (2023) gives additional information to the LLM by retrieving knowledge graph triplet knowledge and converting it to strings to be injected to the prompt. Liu et al. (2022b) generate knowledge for each instance to be added to the prompt. In knowledge prompting via knowledge retrieval, a problem arises if the selected knowledge is not close enough to the instance. This can lead the model to a confusion and later it to give a wrong result. Meanwhile, the knowledge generation method proposed by Liu et al. (2022b) may produce hallucination since it simply asks the model to generate knowledge based on the instance only, without giving a context, thus leading the LLM to give a wrong answer because of misinformation. Moreover, as they generate knowledge for each instance, the computational cost of this approach is high. To address these issues, we propose a framework that generates background knowledge (BK) only once, as a preprocessing step. We then integrate this BK into the prompt for all relevant instances, eliminating the need to generate knowledge for each input and reducing computational overhead.

Background Knowledge Prompting. In contrast with the approaches described above, we propose to inject common-sense knowledge into the prompt. We postulate that this approach can be better than the classical prompting with few-shot in terms of the number of required examples, since it synthesizes several similar examples. As BK generalizes the information, the LLM can learn the reasoning from this generalization rather than focusing on a specific input-output pair. Moreover, our proposed method does not rely on specific knowledge sources as in the case of knowledge prompting via knowledge graph retrieval. The proposed BK

generation is inspired by Shah et al. (2017) and Basile et al. (2018) who propose to utilize frame semantics theory to build default knowledge by extracting frames from raw texts, cluster them, and finally extract the prototypical frame from that cluster. Nevertheless, our approach differs from theirs in that our goal is to synthesize the clustered frame into BK in the form of natural language via LLM prompting.

3 BACKGEN: A BK Generation Framework

The BACKGEN framework is a structured pipeline for generating Background Knowledge (BK) to support Sentiment Phrase Classification (SPC). As shown in Figure 2, it consists of three main steps: (i) **Frame-based Parsing**, where semantic frames and their elements are extracted from annotated examples; (ii) **Frame-based Clustering**, which groups similar frames to identify shared conceptual structures; and (iii) **Background Knowledge Generation**, where a generative model verbalizes the common information in each cluster into reusable BK.

Frame-based Abstraction for Background Knowledge. To generalize beyond individual examples, we rely on Frame Semantics (Fillmore, 1985), which models meaning through structured representations called *frames*. A frame encapsulates a conceptual scenario, consisting of a *Lexical Unit* (LU) and its associated *Frame Elements* (FEs), which define roles such as agents, attributes, or affected entities. Unlike lexical approaches, frames capture abstract relationships that recur across different linguistic expressions, enabling a more structured and reusable representation of meaning.

One of the key advantages of Frame Semantics is its ability to disambiguate lexical meaning based on conceptual structures. Consider the verb *reduce*, which can evoke different frames depending on the context: in “*The government is reducing coal power*”, it evokes the frame CAUSE CHANGE OF POSITION ON A SCALE, where an AGENT actively decreases a QUANTITY. In “*The army reduced enemy resistance*”, however, the verb belongs to

the frame CONQUERING, where a CONQUEROR overcomes a THEME rather than simply decreasing something. If we relied only on lexical similarity we would not be able to distinguish between these cases, whereas with frame-based parsing we can generalize meaning in a structured way that aligns with conceptual distinctions rather than surface word forms.

Beyond disambiguation, frames also facilitate generalization by capturing shared prototypical structures rather than simple text-level similarities, and analogical reasoning through schematic generalization at inference time by providing anchors to map unseen instances in a coherent symbolic structure. A key property of frames is their Frame Elements, which define the roles participating in an event. By clustering instances based on frames and their arguments (such as AGENT or ASSET) we can link sentences that share the same underlying linguistic primitive, regardless of the lexical items they use. For example, “*The government is phasing out coal power*” and “*Public authorities are limiting nuclear energy*” both evoke the CAUSE CHANGE OF POSITION ON A SCALE frame, despite differing in lexical selection. The presence of an AGENT (e.g., *government, public authorities*) and an ATTRIBUTE (e.g., *coal power, nuclear energy*) establishes a conceptual equivalence, allowing the method to identify structurally similar examples even when surface-level word similarity is low. Our aim is to go beyond traditional vector-space models, which primarily capture lexical and distributional similarity (Reimers and Gurevych, 2019), by leveraging frame semantics to identify deeper conceptual patterns.

Structuring Background Knowledge. A key step in our approach is clustering examples that evoke similar situations (frames), involve analogous participants (frame elements), and exhibit comparable role-filler relations which are obtained from the frame parsing process. The objective is to group instances based on deeper structural properties, ensuring that clusters capture prototypical conceptual structures rather than surface-level resemblances. To achieve this, we structure each parsed instance as a tree representation, as illustrated in Figure 3. In this representation, the frame serves as the root node, while frame elements and lexical units form intermediate nodes. The role fillers, which instantiate the semantic arguments of the frame, appear as terminal nodes.

This hierarchical encoding allows us to compare examples not merely by their lexical content but through their structural alignment within the frame-semantic paradigm.

Measuring the similarity between these structured representations requires a metric sensitive to both tree structure and semantic similarity of role fillers. We employ the Smoothed Partial Tree Kernel (SPTK) (Croce et al., 2011), which extends the Partial Tree Kernel (Moschitti, 2006) by incorporating distributed word representations into the kernel computation. This method evaluates the similarity of two trees by counting the number of shared substructures, while also weighting the contribution of lexically different but semantically related elements. In this way, two instances that share the same frame and structural configuration but differ in the lexical realizations of their role fillers will still be considered as similar. For example, the sentences “*The government is phasing out coal power*” and “*Public authorities are limiting nuclear energy*” both evoke the frame CAUSE CHANGE OF POSITION ON A SCALE, with an AGENT and an ATTRIBUTE: they are structurally analogous, and SPTK ensures that their similarity is preserved in the clustering process.

With a well-defined kernel function, we perform clustering using Kernel-based k-means (Dhillon et al., 2004), which embeds the tree structures in an implicit feature space where each dimension corresponds to a possible substructure. Unlike traditional k-means, which relies on explicit Euclidean distances, Kernel-based k-means operates in this high-dimensional space, ensuring that structurally similar examples are grouped together even if their surface forms differ significantly.

Since our task involves sentiment classification, we cluster positive and negative instances separately to maintain polarity coherence. To determine the number of clusters k , we follow a standard heuristic by setting it to the square root of the number of instances in each polarity group.

Background Knowledge Generation. The final step of BACKGEN is the generation of the structured BK from the clustered examples. At this stage, each cluster contains instances that share key semantic properties (such as the evoked frame, the roles of its participants, and the fillers of these roles) while allowing for lexical and syntactic variability. Given this structure, we employ a LLM to generate a concise generalization that synthesizes

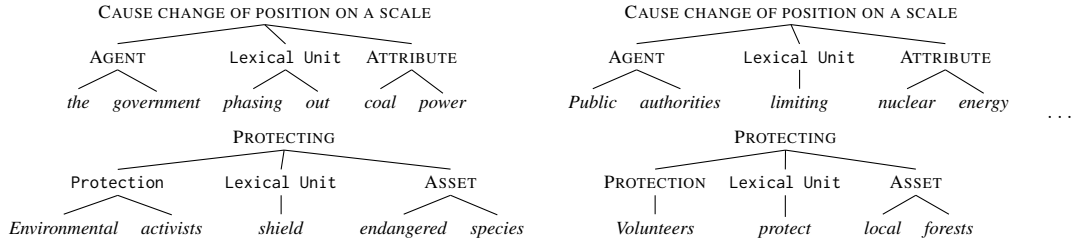


Figure 3: Examples of frame-semantic parse trees obtained from parsing. Each tree represents a frame (root node) with its frame elements (children) and lexical unit (LU).

the core meaning of each cluster.

The strong capabilities demonstrated by LLMs in summarization and abstraction (Liu et al., 2024) make them well-suited for this synthesis step. The prompt, exemplified in Figure 4, instructs the model to generate a general statement based on the provided examples, explicitly leveraging Frame Semantics. The input consists of clustered sentences along with the identified frames, their definitions, and the corresponding lexical units and role assignments. Additionally, the prompt enforces a sentiment constraint, ensuring that the generated BK aligns with the sentiment orientation of the cluster. By providing explicit semantic constraints (such as frame definitions, role structures, and example sentences directly extracted from the dataset) we also aim to mitigate the risk of hallucinations, a common issue in open-ended text generation. This controlled setting ensures that the generated BK remains grounded in the linguistic and conceptual structure of the dataset while still allowing for generalization. For the example shown in Figure 4, where the clustered sentences evoke the PROTECTING frame, the generated BK is: “*The efforts of environmental activists to protect wildlife from harm are viewed as a positive and crucial step toward conservation.*” The generated statements are then stored as BK, forming a knowledge base that can later be queried to enhance in-context learning.

Prompt Injection with BACKGEN’s Generated Knowledge. Once the BK base has been populated, the next challenge is determining how to retrieve relevant information when processing a new instance. Given a new example, the goal is to retrieve BK instances that offer useful generalizations and can be integrated into a prompt in a one-shot or few-shot learning setting. An efficient retrieval strategy is needed that allows selecting representative knowledge from the BK collection. Since the BK is structured into clusters, each con-

Write one sentence expressing general background knowledge based on the provided input sentences that are grouped by shared situations (or frames) modeled according to Frame Semantics Theory. Each input sentence explicitly indicates the Lexical Unit (evoking the frames) and the corresponding role. Definitions of the frames will also be provided to guide the generation. Ensure that the generated text conveys a positive sentiment.

Here are the definitions of the involved frame(s):

- Protecting: Some Protection prevents a Danger from harming an Asset.

Here are the input texts:

1. Environmental activists shield endangered species from extinction caused by poaching.
 - Protecting:
 - Lexical Unit (LU): shield
 - Roles: Asset(endangered species), Protection(environmental activists)
2. Volunteers protect local forests from the threat of wildfires by maintaining firebreaks.
 - Protecting:
 - Lexical Unit (LU): protect
 - Roles: Asset(local forests), Protection(volunteers)

Answer:

Figure 4: Example prompt for generating positive Background Knowledge (BK) from clustered instances, using frames, original text, and frame definitions. The full prompt is in Appendix A, with a simplified version shown here.

taining semantically related examples, retrieval can be efficiently performed by selecting the *medoid* of each cluster as an entry point. The medoid is the instance within the cluster that is closest to the *centroid* in the implicit space induced by the similarity measure (Dhillon et al., 2004), ensuring that it corresponds to a real example in the dataset. This choice allows selecting representative knowledge without needing to compare against all examples.

To retrieve the most relevant BK for a new input, we explore two alternative similarity-based approaches: one leveraging structural similarity through kernel functions and another using seman-

tic similarity in a dense embedding space. The first method is consistent with the clustering process used in BACKGEN as it relies on the same tree-structured representation of frames. Given a new input sentence, its frame representation is extracted and compared against each cluster medoid using the adopted tree kernel function (Croce et al., 2011), selecting those entry whose medoid maximizes the kernel function, i.e. the similarity. This approach captures fine-grained structural alignment between examples, reflecting similarities in event structures and role assignments. The main advantage is that it ensures coherence between the retrieved BK and the input instance. However, it requires parsing the new input according to FrameNet, which may introduce additional computational overhead, particularly in tasks where fast inference is required. An alternative retrieval strategy is based on text similarity. Instead of relying on structured frame representations, dense vector embeddings of both the new input and the BK entry points are computed using a pre-trained language model such as BERT (Reimers and Gurevych, 2019). The similarity between the new instance and each cluster medoid is then measured using cosine similarity, based on the original, unaltered text without frame labeling. This approach avoids the need for explicit frame parsing, making it more adaptable across different tasks, and captures broader contextual relationships beyond frame-level structures. Each retrieval method presents a trade-off between interpretability and efficiency. In our hypothesis, kernel-based retrieval maintains structural coherence, making it preferable when fine-grained semantic consistency is required. Embedding-based retrieval, however, provides a more flexible and computationally efficient alternative. In the experimental section, we evaluate both approaches in terms of their effectiveness in selecting useful BK for prompt augmentation and analyze their impact on task performance. This approach also keeps retrieval efficient, as the number of cluster medoids remains at most $O(\sqrt{n})$, where n is the number of original instances.

4 Experimental Validation

Evaluating a Background Knowledge (BK) repository typically involves assessing the factual accuracy of its statements with respect to real-world knowledge. However, such an evaluation is beyond the scope of this work. Instead, we assess the prac-

Attribute	Statistic	
# negative phrase	1,697	
# positive phrase	876	
avg. span length (# token)	neg. phrase	3.09
	pos. phrase	2.69
# tweets no sentiment phrase	198	
# tweets - total	1,500	

Table 1: Data overview of the aggregated dataset for the sentiment phrase layer.

tical utility of BACKGEN by measuring its impact on a downstream task-Sentiment Phrase Classification (SPC). Specifically, we examine whether integrating BACKGEN-derived BK into prompts improves the ability of a Large Language Model (LLM) to classify the sentiment polarity of a given phrase in context.

Experimental Setup. We created an SPC dataset for the environmental sustainability domain by extending the English dataset by Bosco et al. (2023) with additional language data from the social media platform X. The dataset consists of tweets discussing environmental and socio-political issues, where sentiment interpretation often relies on domain-specific background knowledge. Given the nuanced nature of these discussions, implicit assumptions and contextual understanding play a crucial role in correctly assessing sentiment polarity. The extended dataset follows the same data collection and annotation process as the original, ensuring safety regarding identifying individual people and absence of offensive content. Each message is annotated by three native English speakers from the crowdsourcing platform Prolific², at a rate of 9 GBP per hour, and the labels are aggregated by majority voting over sequence (Rodrigues et al., 2014). Personal information on the annotators is not disclosed in the final dataset. After filtering out the instances with no sentiment phrases, the dataset comprises 2,573 phrases (Table 1).

To parse the text with Frame Semantics, we employ LOME (Xia et al., 2021), a state-of-the-art parser for FrameNet that performs the full pipeline from lexical unit (LU) detection to complete semantic role labeling (SRL). For computing similarity between frame representations, we use the Smoothed Partial Tree Kernel (SPTK) (Croce et al., 2011), implemented within the KELP library (Filice et al., 2018), which also provides the kernel-

²<https://www.prolific.com/>

based k-means clustering algorithm (Dhillon et al., 2004). For generating BK, we use a LLM with a structured prompt following the example in Figure 4. The prompt template, detailed in Appendix A, is designed to extract generalizable knowledge from clustered examples by summarizing their common conceptual patterns. The binary task distinguishes positive and negative sentiment. Due to class imbalance, we report per-class precision, recall, and weighted F1-score. Experiments were run on an NVIDIA A-100.

Experiment and Results. We evaluate the effectiveness of BACKGEN using two state-of-the-art open-source models, Mistral-7B³ and Llama3-8B⁴ (Dubey et al., 2024). Each model is employed both for generating background knowledge (BK) and for performing sentiment phrase classification (SPC), ensuring a consistent evaluation across the entire pipeline. The evaluation follows a 5-fold cross-validation setup. For each fold, BACKGEN is applied to 4/5 of the dataset (training set) to generate a BK database, while the remaining 1/5 is used for testing. The models are tested under different prompting conditions. In the 0-shot setting, the LLM receives only the input text and target phrase, without additional context. In the few-shot setting, one (1-shot) or two (2-shot) examples from the training set are provided in the prompt, either selected randomly (R_{Rand}) or based on text similarity (T_{Sim}). The text similarity is computed via Sentence-BERT embeddings (Reimers and Gurevych, 2019) using all-MiniLM-L6-v2⁵. For background knowledge prompting, the examples are replaced with retrieved BK entries. The retrieval process selects entries based either on frame-based similarity (K_{Kernel}) or text similarity (T_{Sim}), the latter computed using the same Sentence-BERT model. In both cases, the number of BK entries matches the few-shot setting, with one or two retrieved statements included in the prompt. The specific templates used for 0-shot, few-shot, and BK-shot prompting are reported in Appendix B⁶. In all cases, greedy search is used for token genera-

tion to ensure reproducibility and robustness.

Tables 2 and 3 summarize the results in terms of per-class precision, recall, and F_1 score. The weighted F_1 score, which accounts for class imbalance, provides an overall measure of performance. As expected, few-shot prompting improves over 0-shot, with 2-shot generally outperforming 1-shot. Additionally, selecting examples based on their similarity to the test instance (T_{Sim}) leads to better performance than random selection (R_{Rand}), confirming that more relevant examples contribute to better predictions. The most significant improvement comes from replacing explicit examples with structured background knowledge. In particular, BK-based prompting consistently outperforms traditional few-shot methods, demonstrating that synthesized knowledge captures generalizable patterns that are more informative than individual training examples. The 2-BK T_{Sim} configuration achieves the best weighted F_1 scores across both models, with an error reduction relative to 0-shot of 29.94% for Mistral-7B and 26.76% for Llama3-8B. Compared to the stronger 2-shot R_{Rand} baseline, our proposed method yields a substantial relative error reduction, with Mistral-7B achieving 22.73% and Llama3-8B achieving 19.59%. While the absolute improvement in weighted F_1 over the best few-shot baseline may appear modest, these error reduction rates demonstrate the practical value of integrating structured background knowledge into the prompt. We anticipate that this benefit could be even more pronounced in settings where the zero-shot and few-shot baselines are less competitive.

Comparing the two BK selection methods, text similarity-based retrieval ($BK_{T_{\text{Sim}}}$) performs better than frame similarity-based retrieval ($BK_{K_{\text{Kernel}}}$). This suggests that text-based embeddings provide a more robust signal for retrieving relevant knowledge, while frame-based retrieval is more sensitive to parsing errors and the specificity of extracted structures. Compared to the classical few-shot approach, BACKGEN requires an additional step to generate background knowledge (BK); however, this process is performed only once as a preprocessing step. At inference time, the computational cost of retrieving BK entries for prompting is equivalent to that of selecting examples in a standard few-shot setting, when using the same retrieval strategy. Therefore, BACKGEN offers a practical advantage, achieving superior performance with respect to few-shot prompting while maintaining the same inference-time computational requirements.

³<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁵<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁶The model is expected to output *Positive* or *Negative* as the first word. If absent, the first occurrence of either label in the response is used; if neither is found, the instance is marked as unanswered, lowering recall.

Shot	Mistral-7B									
	Negative			Positive			Weighted F_1	Absolute Error	Relative Error Reduction to 0-shot	Relative Error Reduction to 2-shot _{Rand}
	Precision	Recall	F_1	Precision	Recall	F_1				
0-shot	0.966	0.923	0.944	0.886	0.911	0.898	0.928	0.072	-	-10.29%
1-shot _{Rand}	0.957	0.944	0.950	0.917	0.876	0.896	0.931	0.069	4.46%	-5.38%
2-shot _{Rand}	0.969	0.931	0.949	0.895	0.919	0.907	0.935	0.065	9.33%	-
1-shot _{TSim}	0.957	0.955	0.956	0.931	0.877	0.903	0.938	0.062	13.09%	4.15%
2-shot _{TSim}	0.969	0.939	0.953	0.910	0.918	0.914	0.940	0.060	16.30%	7.68%
1-BK _{Kernel}	0.964	0.947	0.955	0.909	0.925	0.917	0.942	0.058	19.50%	11.21%
2-BK _{Kernel}	0.963	0.949	0.956	0.913	0.922	0.919	0.943	0.057	20.89%	12.75%
1-BK _{TSim}	0.965	0.952	0.959	0.917	0.927	0.922	0.946	0.054	24.79%	17.05%
2-BK _{TSim}	0.968	0.956	0.962	0.922	0.930	0.926	0.950	0.050	29.94%	22.73%

Table 2: Results of 5-fold cross-validation for the Sentiment Phrase Classification (SPC) task using Mistral-7B. For each prompting strategy, we report Precision, Recall, and F_1 score for both negative and positive classes, as well as the overall weighted F_1 and Absolute Error. Relative error reduction is shown with respect to both the 0-shot and 2-shot_{Rand} baselines. Prompting methods compared include: zero-shot (no additional context), few-shot with randomly selected (_{Rand}) or Sentence-BERT text similarity-selected (_{TSim}) examples, and background knowledge (BK)-based prompting, where BK entries are retrieved either by frame-semantic kernel similarity (_{Kernel}) or text similarity (_{TSim}). Best overall scores are highlighted in bold.

Shot	Llama3-8B									
	Negative			Positive			Weighted F_1	Absolute Error	Relative Error Reduction to 0-shot	Relative Error Reduction to 2-shot _{Rand}
	Precision	Recall	F_1	Precision	Recall	F_1				
0-shot	0.894	0.922	0.908	0.854	0.731	0.787	0.867	0.133	-	-9.79%
1-shot _{Rand}	0.866	0.954	0.908	0.888	0.706	0.786	0.866	0.134	-0.15%	-9.96%
2-shot _{Rand}	0.881	0.949	0.914	0.884	0.749	0.810	0.879	0.122	8.92%	-
1-shot _{TSim}	0.867	0.958	0.910	0.901	0.707	0.792	0.870	0.130	2.40%	-7.16%
2-shot _{TSim}	0.882	0.955	0.917	0.900	0.751	0.819	0.884	0.116	12.89%	4.36%
1-BK _{Kernel}	0.890	0.942	0.915	0.873	0.767	0.816	0.881	0.119	10.87%	2.14%
2-BK _{Kernel}	0.887	0.951	0.919	0.893	0.759	0.820	0.885	0.115	13.87%	5.43%
1-BK _{TSim}	0.882	0.948	0.914	0.882	0.748	0.809	0.878	0.122	8.62%	-0.33%
2-BK _{TSim}	0.900	0.962	0.930	0.915	0.791	0.848	0.902	0.098	26.76%	19.59%

Table 3: 5-fold cross-validation results using Llama3-8B, following the same setup as in Table 2.

Overall, these results highlight the potential of structured background knowledge to enhance sentiment phrase classification. By capturing conceptual generalizations rather than relying on specific examples, BACKGEN mitigates the performance variability associated with example selection and provides a more stable and effective alternative to few-shot learning.

5 Error Analysis

To better understand the impact of BK on model predictions, we analyze cases where BK improves classification as well as those where it introduces errors. The goal is to identify patterns in both helpful and harmful BK selections. Given that Mistral-7B outperforms Llama3-8B, we conduct this analysis using Mistral-7B with the 2-shot BK selection based on text similarity.

BK is particularly useful when the sentiment polarity of a phrase depends on contextual understanding. For example, in the instance “big problems

may arise if your ductwork system is not installed correctly homeowners will encounter **discomfort poor indoor air quality inflated electricity bills periodic repairs and in some cases complete replacement**”, the 0-shot model incorrectly classifies the target phrase “big problems may arise” as positive. However, a retrieved negative BK statement, i.e., “**The constant increase in expenses for various reasons, such as pollution and gentrification, is a major issue that negatively impacts our lives.**”, helps the model correctly reclassify the phrase as negative by reinforcing the association between financial burdens and negative sentiment.

Errors in BK selection primarily arise when (i) the retrieved BK is not sufficiently similar to the test instance, (ii) the BK is too generic, or (iii) the BK is overly specific. In cases where the retrieved BK does not align closely with the input, the model struggles to integrate it into the classification decision. Although the BK may contain relevant commonsense knowledge, it fails to pro-

vide meaningful guidance due to its semantic distance from the test instance. This can lead to the model overriding a previously correct classification, sometimes defaulting to a neutral response such as “. . . *The background knowledge does not provide enough information to determine the polarity of the target phrase.*” This suggests that, beyond BK retrieval, there is potential value in using model uncertainty as a signal, if no sufficiently relevant BK is found, the test instance itself may be an outlier relative to the training data. Another failure mode occurs when the retrieved BK is too generic. This typically results from poor clustering, where multiple frames that are not semantically aligned are grouped together, leading to vague or uninformative statements. For example, a BK entry such as “*Changes in policies can have a significant impact on society*” lacks specificity, making it difficult for the model to determine sentiment in a meaningful way. Overly specific BK can also introduce bias, particularly when the generated knowledge repeatedly mentions the same entity across multiple instances. Consider the instance “*you do realize bill gates is heavily invested in animal agriculture right he has enormous feed crop landholdings for animal ag supplying factory farms amp feedlots he also he invests in gmo cow research*”, where the 0-shot model correctly classifies the target phrase “*heavily invested*” as positive. However, one retrieved BK statement, i.e., “*The fact that Bill Gates is involved in funding and promoting synthetic meat, despite Jeremy’s disdain for him, is a disappointing turn of events.*”, introduces a negative stereotype, leading the model to misclassify the phrase as negative. This suggests that the model is overfitting to entity-level associations rather than recognizing general sentiment cues. A potential solution is to refine the BK generation prompt to avoid explicit mentions of named entities, ensuring the generated knowledge remains applicable.

6 Conclusions and Future Works

We presented BACKGEN, a framework that leverages Frame Semantics to generate structured Background Knowledge (BK) as a principled alternative to example-based prompting. By clustering semantically related instances and synthesizing generalized knowledge via LLMs, BACKGEN enables abstract and robust reasoning. Applied to Sentiment Phrase Classification (SPC), where polarity often depends on implicit context, BK-based prompt-

ing significantly improves performance, achieving up to 29.94% error reduction over zero-shot and 22.73% over few-shot baselines.

While our main focus is on evaluating whether background knowledge (BK) improves SPC performance over few-shot prompting, we do not directly assess the quality of the generated BK. A systematic evaluation of its fluency and generalizability could inform prompt refinement and strengthen our approach, which we leave for future work.

Further directions include exploring the application of background knowledge (BK) in other tasks that require commonsense, analogical, and context-sensitive reasoning, such as question answering or stance detection. We will also test the cross-domain generalization of BACKGEN by evaluating its scalability and robustness when BK is generated from domains different from the target task. Another line of investigation will focus on identifying and mitigating potential biases in both the generated knowledge and the underlying dataset, with the goal of increasing fairness and control in BK construction. Additionally, we plan to compare BACKGEN directly with existing knowledge prompting methods, to assess whether it can serve as a viable alternative beyond the classical few-shot baseline explored in this paper. Finally, we see promising opportunities in connecting BK with explainability: by anchoring model predictions to frame-based abstractions, BACKGEN could support more interpretable and cognitively grounded language understanding.

Limitations

The applicability of the BK database produced in this study is currently limited to the environmental sustainability domain, and its effectiveness in other sentiment analysis tasks remains to be explored. Additionally, as BACKGEN relies on a frame parser, the quality of the generated BK is inherently dependent on the accuracy of the parser.

Another limitation is the lack of automatic analysis of the collected BK statements, which may unintentionally introduce biases. Since BK is generated from clustered instances, certain perspectives may be overrepresented, reinforcing pre-existing biases in the data. Future work should focus on developing methods for detecting and mitigating such biases, ensuring that the generated BK remains neutral and representative across different domains. Moreover, investigating how BK influences model

reasoning, i.e., particularly in tasks requiring explainability, could provide insights into its broader applicability beyond sentiment analysis.

Ethical Reflections

It is important to consider the potential risks of NLP tools like BACKGEN, particularly the possibility of generating biased or misleading background knowledge (BK). Without proper safeguards, BACKGEN could produce inaccurate, overly generalized, or even harmful statements that misrepresent real-world contexts, especially in sensitive areas like environmental sustainability. To mitigate this risk, prompt design should be carefully refined to encourage neutral and well-grounded knowledge generation. Additionally, a verification step should be implemented to detect and filter out problematic BK, ensuring that the generated content remains accurate, unbiased, and contextually appropriate.

References

- Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. [Knowledge-augmented language model prompting for zero-shot knowledge graph question answering](#). In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCH-ING 2023)*, pages 70–98, Toronto, ON, Canada. Association for Computational Linguistics.
- Valerio Basile, Roque Lopez Condori, and Elena Cabrio. 2018. [Measuring frame instance relatedness](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 245–254, New Orleans, Louisiana. Association for Computational Linguistics.
- Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile, and Indra Budi. 2023. How green is sentiment analysis? environmental topics in corpora at the university of turin. In *The 9th Italian Conference on Computational Linguistics (CLiC-it)*, volume 3596. CEUR-WS.
- 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, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. [Structured lexical similarity via convolution kernels on dependency trees](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. 2004. [Kernel k-means: spectral clustering and normalized cuts](#). In *KDD*, pages 551–556. ACM.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). *Preprint*, arXiv:2301.00234.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Simone Filice, Giuseppe Castellucci, Giovanni Da San Martino, Alessandro Moschitti, Danilo Croce, and Roberto Basili. 2018. [Kelp: a kernel-based learning platform](#). *Journal of Machine Learning Research*, 18(191):1–5.
- Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.
- Abdullatif Köksal, Timo Schick, and Hinrich Schuetze. 2023. [MEAL: Stable and active learning for few-shot prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 506–517, Singapore. Association for Computational Linguistics.
- Itay Levy, Ben Bogin, and Jonathan Berant. 2023. [Diverse demonstrations improve in-context compositional generalization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Hongfu Liu and Ye Wang. 2023. [Towards informative few-shot prompt with maximum information gain for in-context learning](#). In *Findings of the Association*

- for *Computational Linguistics: EMNLP 2023*, pages 15825–15838, Singapore. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022b. [Generated knowledge prompting for commonsense reasoning.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.
- Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. [On learning to summarize with large language models as references.](#) *Preprint*, arXiv:2305.14239.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Machine Learning: ECML 2006*, pages 318–329, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024a. [A survey on stability of learning with limited labelled data and its sensitivity to the effects of randomness.](#) *ACM Comput. Surv.* Just Accepted.
- Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2024b. [Automatic combination of sample selection strategies for few-shot learning.](#) *Preprint*, arXiv:2402.03038.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. [Sequence labeling with multiple annotators.](#) *Machine Learning*, 95(2):165–181.
- Avijit Shah, Valerio Basile, Elena Cabrio, and Sowmya Kamath S. 2017. Frame instance extraction and clustering for default knowledge building. In *CEUR Workshop Proceedings*, volume 10, pages 1–10.
- Chengyu Song, Fei Cai, Mengru Wang, Jianming Zheng, and Taihua Shao. 2023. [Taxonprompt: Taxonomy-aware curriculum prompt learning for few-shot event classification.](#) *Knowledge-Based Systems*, 264:110290.
- Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. [Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.
- Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. [LOME: Large ontology multilingual extraction.](#) In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. [Active example selection for in-context learning.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Prompts for Background Knowledge Generation

The BACKGEN framework employs two prompts to generate Background Knowledge (BK) from clusters of semantically similar instances. These clusters are formed by grouping examples that evoke the same semantic frames and share a common sentiment polarity, either positive or negative. Each cluster is then processed using the appropriate prompt:

- Clusters of positive instances use the **Positive Sentiment Knowledge Prompt** (Figure A).
- Clusters of negative instances use the **Negative Sentiment Knowledge Prompt** (Figure B).

Each prompt follows a standardized structure to ensure consistency in BK generation:

1. **Task Definition:** The prompt begins with an explicit instruction, guiding the model to generate a single sentence that captures general background knowledge from the clustered examples. This instruction specifies that the output should reflect a stereotypical generalization, either positively or negatively framed, depending on the sentiment of the cluster.
2. **Example Cluster:** The prompt includes an example cluster of semantically related instances, where each input sentence is annotated with its corresponding *frame-semantic structure*. This includes:
 - The **Lexical Unit (LU)** evoking the frame.
 - The **Frame Elements (roles)** present in the sentence.
 - The **Frame Definitions** to provide contextual understanding.
3. **Example BK Statement:** A correctly structured BK statement is provided as a reference, demonstrating the level of abstraction and generalization expected from the model. This serves as a guideline to ensure that the output captures high-level conceptual knowledge rather than instance-specific details.
4. **Target Cluster for BK Generation (“Your Turn”):** The final section of the prompt presents a new set of sentences from a different cluster (all sharing the same sentiment polarity and evoking similar frames). This part of the prompt contains placeholders (e.g., {frame_n}, {text_n}, {LU}, {arguments_of_frame}) that are dynamically populated based on the actual instances and frame annotations of the current cluster. The model is then instructed to generate a single BK statement that generalizes the semantic properties of these instances, mirroring the structure of the provided example.

Both prompts are designed to ensure that the model generates reliable, structured commonsense knowledge that can be effectively injected into prompts for downstream NLP tasks. Additionally, the framework supports variations of these prompts where the instruction is modified to generate a short paragraph instead of a single sentence, allowing for more detailed knowledge synthesis.

B Prompts for Sentiment Phrase Classification (SPC)

To evaluate different prompting strategies in Sentiment Phrase Classification (SPC), we employed three approaches:

- **Zero-shot** (Figure C): The model classifies the sentiment polarity (*positive* or *negative*) of a target phrase within a given text without additional context. The prompt explicitly instructs the model to provide a classification and a brief explanation.
- **Few-shot** (Figure D): The model is given one or two labeled examples (*1-shot* or *2-shot*) before classifying the target phrase. The examples are either selected randomly (R_{rand}) or based on text similarity (T_{sim}) with the input instance. The model cannot explicitly reference these examples in its explanation.

Write one sentence expressing general background knowledge that reflects stereotypical information, based on the input sentences provided. These sentences are grouped by shared situations (or frames) modeled according to Frame Semantics Theory. Each input sentence explicitly indicates the Lexical Unit (evoking the frames) and the corresponding role. Definitions of the frames will also be provided to guide the generation. Ensure that the generated text conveys a positive sentiment and the reason for the sentiment should be made explicit.

Example:

Here are the definitions of the involved frame(s):

- Cause_change_of_position_on_a_scale: This frame consists of words that indicate that an Agent or a Cause affects the position of an Item on some scale (the Attribute) to change it from an initial value (Value_1) to an end value (Value_2).

Here are the input texts:

1. if the tourism sector is serious about reducing its footprint they should choose real emission reductions and biodiversity protection even airlines are starting to move away from offsets for nature
 - Cause_change_of_position_on_a_scale:
 - Lexical Unit (LU): reducing
 - Roles: Attribute(its footprint)
2. moving away from capitalism green washing is not easy under the current systems political allegiances we live within so i commend for being bold enough to try but let us not forget that redistributing wealth and reducing consumerism must remain 1 priorities
 - Cause_change_of_position_on_a_scale:
 - Lexical Unit (LU): reducing
 - Roles: Attribute(consumerism)
3. india reduced emission intensity of its gdp by 24 per cent in 11 yrs through 2016 un via official pollution
 - Cause_change_of_position_on_a_scale:
 - Lexical Unit (LU): reduced
 - Roles: Agent(India),Attribute(emission intensity of its GDP),Difference(by 24 per cent),Speed(in 11 yrs),Time(through 2016),Means(un via official pollution)

Answer: Reducing material that is bad for the environment is a positive act.

Your Turn:

Here are the definitions of the involved frame(s):

- {frame_1} : {definition_of_frame_1}.
- ...
- {frame_n} : {definition_of_frame_n}

Here are the input texts:

1. {text_1}
 - {frame_label_of_text_1}:
 - Lexical Unit (LU): {LU_span_of_frame_label_of_text_1}
 - Roles: {arguments_of_frame_label_of_text_1}
 - ...
- n. {text_n}
 - {frame_label_of_text_n}:
 - Lexical Unit (LU): {LU_span_of_frame_label_of_text_n}
 - Roles: {arguments_of_frame_label_of_text_n}

Answer:

Figure A: Prompt for generating positive sentiment background knowledge. The input sentences are clustered based on shared semantic frames, and the model is instructed to generate a generalized knowledge statement that reflects a positive sentiment.

- **BK-shot** (Figure E): Instead of example-based prompting, the model receives background knowledge (BK) statements generated by BACKGEN. These statements, selected using either frame similarity (K_{frame}) or text similarity (T_{sim}), provide generalizable knowledge to guide sentiment classification.

Each prompt follows a structured format, including:

- **Task Definition:** the goal is to classify the sentiment polarity of a given target phrase.
- **Instructions:** Constraints are provided, including the requirement for a polarity label and an explanation, without explicit reference to examples or BK.
- **Input Information:** The given text and target phrase are explicitly stated.

Write one sentence expressing general background knowledge that reflects stereotypical information, based on the input sentences provided. These sentences are grouped by shared situations (or frames) modeled according to Frame Semantics Theory. Each input sentence explicitly indicates the Lexical Unit (evoking the frames) and the corresponding role. Definitions of the frames will also be provided to guide the generation. Ensure that the generated text conveys a negative sentiment and the reason for the sentiment should be made explicit.

Example:

Here are the definitions of the involved frame(s):

- Causation: A Cause causes an Effect.
- Destroying: A Destroyer (a conscious entity) or Cause (an event, or an entity involved in such an event) affects the Patient negatively so that the Patient no longer exists.
- Cause_to_end: An Agent or Cause causes a Process or State to end.
- Cause_to_amalgamate: These words refer to an Agent joining Parts to form a Whole.

Here are the input texts:

1. water pollution is putting our health at risk unsafe water kills more people each year than war and all other forms of violence combined here are six causes of water pollution as well as what we can do to reduce it
 - Causation:
 - Lexical Unit (LU): putting
 - Roles: Cause(water pollution),Effect(our health),Cause(at risk unsafe water kills more people each year than war and all other forms of violence combined)
2. i hope izzy one day understands that we can be against pollution in all it s forms which truly is destroying our environment and health but also be smart enough to see through the carbon emissions global warming shenanigans
 - Destroying:
 - Lexical Unit (LU): destroying
 - Roles: Cause(pollution in all it s forms),Cause(which),Patient(our environment and health)
3. extinction is forever amp for all we know we have lost what we will need to fix things when it becomes obvious we have to do something technology will not end pollution of the air water soil or the contamination of our food earth cycles themselves will be the only way out of it
 - Cause_to_end:
 - Lexical Unit (LU): end
 - Roles: Cause(technology),State(pollution of the air water soil)
4. water pollution is putting our health at risk unsafe water kills more people each year than war and all other forms of violence combined here are six causes of water pollution as well as what we can do to reduce it
 - Cause_to_amalgamate:
 - Lexical Unit (LU): combined
 - Roles: Parts(all other forms of violence)

Answer: The existence of pollution and other materials that cause damage and destroy our environment is very negative.

Your Turn:

Here are the definitions of the involved frame(s):

- {frame_1}: {definition_of_frame_1}.
- ...
- {frame_n}: {definition_of_frame_n}

Here are the input texts:

1. {text_1}
 - {frame_label_of_text_1}:
 - Lexical Unit (LU): {LU_span_of_frame_label_of_text_1}
 - Roles: {arguments_of_frame_label_of_text_1}
 - ...
- n. {text_n}
 - {frame_label_of_text_n}:
 - Lexical Unit (LU): {LU_span_of_frame_label_of_text_n}
 - Roles: {arguments_of_frame_label_of_text_n}

Answer:

Figure B: Prompt for generating negative sentiment background knowledge. The model generates a background knowledge statement that reflects the negative sentiment conveyed by the clustered examples.

- **Additional Context:** In few-shot prompting, examples are included; in BK-shot prompting, relevant background knowledge statements are injected instead.

- **Expected Output:** The model generates a classification followed by a justification.

Figures C, D, and E illustrate the complete templates for the zero-shot, few-shot, and BK-shot prompts.

Task: Determine the polarity (either 'positive' or 'negative') of the target phrase from the provided text. Then, provide a short explanation for your classification. The explanation should be clear and helpful for the user to understand the choice.

Instructions:

- The polarity output can only be 'positive' or 'negative'.
- The first word of your answer should be your final polarity classification, then followed by your explanation.

Input:

- Text: {text}
- Target Phrase: {target_phrase}

Answer:

Figure C: Prompt zero-shot for SPC.

Task: Determine the polarity (either 'positive' or 'negative') of the target phrase from the provided text. Then, provide a short explanation for your classification. You are also provided with some examples. The explanation should be clear and helpful for the user to understand the choice.

Instructions:

- Use the examples to help determine the polarity.
- Note the sentiment of each example as it may assist in your reasoning.
- The polarity output can only be 'positive' or 'negative'.
- The first word of your answer should be your final polarity classification, then followed by your explanation.
- The user is not aware of the examples, so you cannot refer to them explicitly in your explanation.

Input:

- Text: {text}
- Target Phrase: {target_phrase}

Examples:

1. {example_text_1}. Target Phrase: {example_target_phrase_1}. Sentiment: {example_polarity_1}
- ...
- n. {example_text_n}. Target Phrase: {example_target_phrase_n}. Sentiment: {example_polarity_n}

Answer:

Figure D: Prompt few-shot for SPC.

Task: Determine the polarity (either 'positive' or 'negative') of the target phrase from the provided text. Then, provide a short explanation for your classification. You are also provided with potentially useful sentences reflecting background knowledge. The explanation should be clear and helpful for the user to understand the choice.

Instructions:

- Use the background knowledge to help determine the polarity.
- Note the sentiment of each background sentence as it may assist in your reasoning.
- The polarity output can only be 'positive' or 'negative'.
- The first word of your answer should be your final polarity classification, then followed by your explanation.
- The user is not aware of the background knowledge, so you cannot refer to it explicitly in your explanation.

Input:

- Text: {text}
- Target Phrase: {target_phrase}

Examples:

1. {bk_text_1}. {bk_polarity_1}
- ...
- n. {bk_text_n}. {bk_polarity_n}

Answer:

Figure E: Prompt BK injection shot (bk-shot) for SPC.

On choosing the vehicles of metaphors without a body: evidence from Large Language Models

Veronica Mangiaterra¹, Chiara Barattieri di San Pietro¹, Federico Frau¹,
Valentina Bambini¹, Hamad Al-Azary²

¹Laboratory of Neurolinguistics and Experimental Pragmatics (NEPLab),
University School for Advanced Studies IUSS, Pavia, Italy

²Lawrence Technological University, Southfield, MI, USA
{veronica.mangiaterra, valentina.bambini}@iusspavia.it,
halazary@ltu.edu

Abstract

Since the advent of Large Language Models (LLMs), much work has been devoted to comparing the linguistic abilities of humans and machines. Figurative language, which is known to rely on pragmatic inferential processes as well as lexical-semantic, sensorimotor, and socio-cognitive information, has been often used as a benchmark for this comparison. In the present study, we build on previous behavioral evidence showing that both distributional and sensorimotor variables come into play when people are asked to produce novel and apt metaphors and examine the behavior of LLMs in the same task. We show that, while distributional features still hold a special status, LLMs are insensitive to the sensorimotor aspects of words. This points to the lack of human-like experience-based grounding in LLMs trained on linguistic input only, while offering further support to the multimodality of conceptual knowledge involved in metaphor processes in humans.

1 Introduction

Large language models (LLMs)’s outstanding abilities to solve increasingly complex linguistic tasks (Bubeck et al., 2023; Marvin and Linzen, 2018; Srivastava et al., 2022; Strachan et al., 2024; Webb et al., 2023) have given rise to a theoretical debate on how their skills compare to those of humans (Birhane and McGann, 2024; Mahowald et al., 2024; Mitchell and Krakauer, 2023).

One main point of debate is that, while human linguistic knowledge is acquired through multimodal (sensory, motor, emotional, verbal, etc.) sources (Barsalou, 2008), most LLMs are trained on huge amounts of text only (Bender and Koller, 2020; Bisk et al., 2020), questioning whether LLMs can be considered psychologically valid models of cognition (Bolhuis et al., 2024; Cuskley et al., 2024; Lake and Murphy, 2023).

Pragmatic skills, namely the ability to understand the context-dependent meaning of utterances, including non-literal expressions, have been taken as an ideal test bed to explore the high-level linguistic abilities of LLMs (Barattieri di San Pietro et al., 2023; Liu et al., 2022). Importantly, interpreting non-literal meaning might require skills beyond determining statistical regularities in language. Theoretical pragmatic accounts have highlighted the context-bound nature of pragmatic reasoning (Gumperz, 1982), the need for meta-representation of the speaker’s intention (Sperber, 1994), and the potential presence of imagistic processes alongside purely verbal processes (Carston, 2018; Gibbs Jr and Matlock, 2008; Paivio and Walsh, 1993). These claims are supported by neuroimaging experimental evidence that showed, for non-literal meaning processing, activations of brain circuits linked to mental imagery (Mashal et al., 2014) as well as Theory of Mind skills, namely the ability to attribute mental states to others (Bambini et al., 2011; Enrici et al., 2019; Van Ackeren et al., 2012).

Among pragmatic phenomena, figurative language and metaphors are of particular interest. Metaphors are non-literal uses of language that require a context-driven adjustment of the lexically encoded concepts to grasp the interlocutor’s intended meaning (Wilson and Carston, 2007). Most studies, including this one, focus on nominal metaphors that involve two terms – a topic and a vehicle – in the form “X is Y”. In a metaphor such as “Lawyers are sharks”, the concept evoked by the vehicle “sharks” is adjusted by dropping semantic features that are not salient based on the context – e.g., “swims” and “has a fin” – while promoting features leading to the intended meaning, such as “aggressive” and “ruthless”. Studies showed that, in addition to lexical-semantic processes, sensorimotor processes also have a role in metaphor processing (Battaglini et al., 2025; Al-

Azary and Katz, 2021). For example, Al-Azary and Katz (2021) showed that the bodily-action aspects of words are active when processing low-familiar metaphors.

The metaphoric processing abilities of LLMs have been widely investigated (Carenini et al., 2023; Ichien et al., 2024; Neidlein et al., 2020; Prystawski et al., 2022; Wachowiak and Gromann, 2023). LLMs exhibit high accuracy in identifying and interpreting metaphors, yet their performance may rely on different mechanisms compared to humans, as shown by different patterns of errors (Barattieri di San Pietro et al., 2023; Liu et al., 2022) or the need for psychologically informed paradigms to improve their responses (Prystawski et al., 2022).

Compared to metaphor identification and interpretation, metaphor production is a less investigated area, both in humans and machines. Examining how people construct metaphors, and which semantic features guide the process of metaphorical conceptualization, may shed further light on high-level language skills and the role of multimodality, and therefore is a fertile ground for comparison with LLMs.

Katz (1989) investigated metaphor production with a *vehicle selection paradigm*, (i.e., asking participants to produce a metaphor by selecting the metaphor vehicle word among a given set), and found that participants tend to choose concrete words with a moderate semantic distance from the topic. Expanding the study of the semantic features at work in metaphor production, Al-Azary and Katz (2023) investigated the role of semantic richness, namely the amount of semantic information carried by a word (Yap et al., 2012). Specifically, they used two variables reflecting different aspects of semantic richness: Semantic Neighborhood Density (SND) and Body-Object Interaction (BOI). SND was defined as high (dense) or low (sparse), based on the average distance between the word and its semantic neighbors (Buchanan et al., 2001). Body-Object Interaction is a normed variable, derived from human ratings, that indicates the ease with which a human body can physically interact with a word’s referent (Siakaluk et al., 2008), thus reflecting sensorimotor richness. Humans are more likely to experience physical interaction with the referent of a high-BOI word such as ‘umbrella’ rather than that of a low-BOI word such as ‘volcano’. They found that participants prefer vehicles with low SND and low BOI, like ‘cloud’ or ‘rainbow’

(rather than high-BOI high-SND words like ‘pillow’), resulting in metaphors such as ‘Boredom is a cloud’ or ‘Persuasion is a rainbow’. This indicates that humans are sensitive to both sensorimotor and linguistic richness, as expressed by high SND and high BOI, in metaphor production.

In this study, we tested LLMs in the same metaphor production task to examine whether distributional and sensorimotor features of words, which are relevant for human participants, drive metaphor production in LLMs as well, and whether LLMs show the same direction of effects.

2 Methods

We reproduced the behavioral experiments of the study by Al-Azary and Katz (2023) on LLMs by prompting four models within the open-source family of GPT2 models developed by OpenAI (Radford et al., 2019).

2.1 Material

Materials were taken from the original study. The set of stimuli included 36 topics and 48 potential vehicles. Topics were abstract words, as assessed by low concreteness ratings in the Brysbaert et al. (2014) dataset. Moreover, topics were balanced in terms of SND values extracted from the WINDSORS database (Durda and Buchanan, 2008), with half of them with low SND and half with high SND. Vehicles were concrete words, as assessed by high concreteness ratings in the Brysbaert et al. (2014) dataset and were balanced in terms of both SND values (Durda and Buchanan, 2008) and Body-Object Interaction (BOI) values, extracted from the databases by Bennett et al. (2011) and Tillotson et al. (2008), resulting in four combinations of 12 vehicles each. An example of each category of topic and vehicle is provided in Table 1.

2.2 Models

We employed four pre-trained transformer-based language models developed by OpenAI (Radford et al., 2019): GPT2 (124M parameters), GPT2-medium (355M parameter), GPT2-Large (774M parameters), and GPT2-XL (1.5B parameters). These models differ only in architecture scale: the number of layers, hidden dimensions, and parameters increases progressively, while the training data and objectives remain constant. Specifically, these models are trained on *WebText*, a large corpus of English created by scraping 45 million links from

Topic (abstract)	Low SND	Nostalgia
	High SND	Empathy
Vehicle (concrete)	High SND - High BOI	Seed
	High SND - Low BOI	Butterfly
	Low SND - High BOI	Umbrella
	Low SND - Low BOI	Lighthouse

Table 1: Examples of topics and vehicles presented in the experiment and their respective semantic conditions. Note: SND = Semantic Neighborhood Density; BOI = Body-Object Interaction.

Reddit. Differently from larger GPT models, GPT2 models provide access to the probability distributions over strings of words, a way of testing LLMs’ capabilities proven to be more reliable than prompting alone (Hu and Levy, 2023).

2.3 Procedure

In the original study, human participants were presented with the topic and the set of 48 vehicles, and they were asked to choose one of the vehicles to create a comprehensible and apt metaphor, resulting in 36 unique metaphors produced. To replicate this experimental paradigm, we prompted the model with the string "TOPIC is a/an" and we collected the likelihood scores of each candidate vehicle. For each topic, we then extracted, the normalized probability distribution over the whole set of vehicles. Probabilities were normalized using a softmax-like transformation with a temperature parameter $\tau = 0.05$ to control sharpness. We randomized vehicle order to reduce ordering bias and used a fixed seed (set prior to execution) to ensure reproducibility. The code was adapted from Carenini et al. (2023). Even if GPT models are not explicitly instructed to produce a metaphor, all possible candidates given the prompt "TOPIC is a/an" form a metaphorical expression. Thus, we expect that the models, when completing the prompt in the most likely way as they are trained to do, will provide us with what the LLM consider the most comprehensible and apt metaphor among the possible metaphorical combinations.

2.4 Statistical Analysis

To test whether LLMs choose vehicles according to the semantic features of both topics and vehicles, we fit a set of Linear Mixed Models using

lme4 and lmerTest packages (Bates et al., 2015; Kuznetsova et al., 2017) for each GPT2 model, separately. We consider the z-scaled probability of the vehicle as the dependent variable and Topic SND, Vehicle SND, and Vehicle BOI as interacting categorical predictors. A random intercept was added to account for the variability of individual vehicles. The resulting formula was: $lmer(\text{probability} \sim \text{vehicle-SND} * \text{vehicle-BOI} * \text{topic-SND} + (1|\text{vehicle}))$. Alpha level was set at .05.

3 Results

The fitted linear mixed models showed a main effect of SND of the vehicles for GPT2-Medium ($\beta = 0.64, t = 2.316, p = 0.025$), GPT2-Large ($\beta = 0.612, t = 2.264, p = 0.029$), and GPT2-XL ($\beta = 0.589, t = 2.159, p = 0.036$). These three models showed a higher probability of choosing low-SND vehicles compared to high-SND (Figure 1).

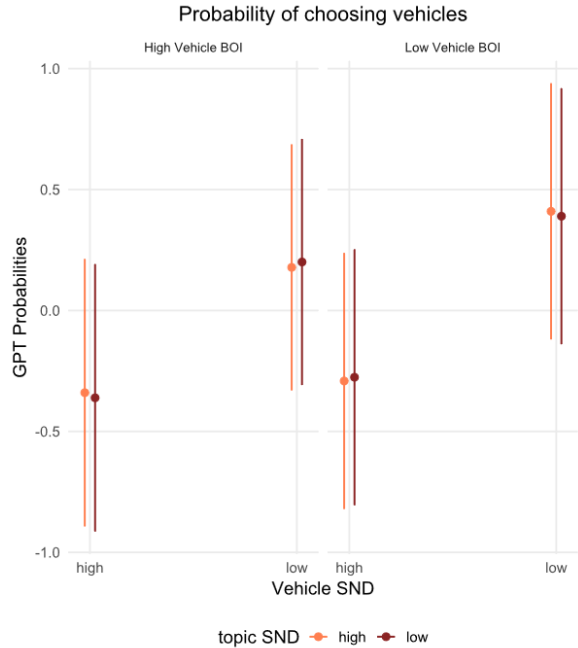


Figure 1: Effect of Vehicle SND on z-scaled probabilities of choosing vehicles extracted from GPT2-Large. Similar patterns are observed in the other GPT2 models.

Concerning GPT2, the linear mixed model showed an interaction between Topic SND and Vehicle SND ($\beta = -0.056, t = -2.321, p = 0.020$), meaning that GPT2 assigned higher probabilities to low-SND vehicles when completing metaphors with high-SND topics.

No effect of vehicle BOI was found for all four GPT2 models, regardless of their size. Some examples of the metaphors produced by models (and

by humans from the original [Al-Azary and Katz \(2023\)](#) study) can be found in Appendix A.

4 Discussion

In this work, we examined the role of distributional (SND) and sensorimotor (BOI) variables in metaphor production by LLMs. To do so, we prompted four models of the GPT2 family with a metaphor production task from [Al-Azary and Katz \(2023\)](#). In the original study, they found that human participants, when asked to choose a vehicle to construct an apt metaphor, tend to rely on both SND and BOI, preferring vehicles with low values in both variables. Differently, LLMs are not sensitive to BOI values, equally choosing low- and high-BOI words as vehicles, but showing a preference for low-SND vehicles, as humans do. In other words, LLMs align with human participants in choosing low-SND words such as ‘puzzle’, yet, they assign similarly high probabilities to metaphors like "Boredom is a vacuum" (high BOI) and "Persuasion is an eagle" (low BOI), showing no sensitivity to BOI values — in contrast to human participants, who preferred metaphors with low-BOI words such as "Boredom is a cloud" and "Persuasion is a butterfly".

In their original study, [Al-Azary and Katz \(2023\)](#) argued in favor of an advantage of low semantic richness in the emergence of metaphorical meaning. In line with that, [Al-Azary and Buchanan \(2017\)](#) showed that semantic richness, operationalized as a greater number of closer semantic neighbors (high SND), hinder the comprehensibility of metaphors, in that the adjustment process to derive the intended meaning in semantically rich concepts is more costly compared to semantically less-rich concepts, as many more features of the word need to be discarded ([Kintsch, 2000](#)), hence leading to a preference for low semantic richness (confirmed in studies on literary metaphors ([Reid et al., 2023](#); [Mangiaterra et al., 2024](#))).

Similarly, LLMs seem to adhere to the tendency toward low semantic richness exhibited by humans both in comprehension ([Al-Azary and Buchanan, 2017](#)) and production ([Al-Azary and Katz, 2023](#)). However, LLMs do so by relying only on the distributional features of words (SND) rather than their sensorimotor content (BOI). This pattern did not change across the four GPT2 models employed in this study, suggesting that the ability to rely on sensorimotor-relevant features is not significantly

enhanced by simply model scale. Although we did not find evidence of sensorimotor effects LLMs in this task, we cannot rule out sensorimotor effects in other tasks, which can be an area of future research. However, the lack of ability of the four GPT2 models to modulate experience-based aspects of vehicles is in line with previous work tackling (embodied) cognition in language models. [Xu et al. \(2023\)](#) found that, while language models have human-like representations of words in non-sensorimotor domains, they do not align with respect to words related to sensory and motor domains. Similarly, [Lee et al. \(2025\)](#) showed that even if models could approximate human perceptual ratings, they do so by relying more on linguistic cues. Even multimodal LLMs fall behind human performance in classical tasks exploring the activation of implied sensorimotor features of concepts in language processing ([Jones et al., 2024](#)). There is also specific evidence of a poor handling of sensorimotor experience in LLM’s processing of metaphor. For instance, [Barattieri di San Pietro et al. \(2023\)](#) found that ChatGPT performs better on metaphors expressing a psychological characteristic of the topic compared to metaphors capitalizing on physical features. All of these studies employed larger models compared to the GPT2 models used in the present work, confirming that even if scaling improves language performance, some sensorimotor aspects cannot fully emerge from training on textual input or input from visual modality only.

The role of both disembodied (SND) and embodied (BOI) sources of linguistic knowledge in metaphor production in humans ([Al-Azary and Katz, 2023](#)) supports the theoretical account of the “dual coding theory” ([Paivio, 1979](#)), namely the perspective according to which the nature of concepts is both symbolic and imagistic. Our results on LLMs may provide evidence in support of these claims about the nature of conceptual representation, as well. On the one hand, GPT2 models have access to verbal input, and this is reflected in their human-like behavior with respect to distributional aspects of words, meaning that a certain portion of linguistic abilities may have a purely disembodied nature and can be acquired with exposure to text only. On the other hand, a certain part of linguistic skills in humans is linked to sensorimotor experience, and the necessity of this experience for human-like linguistic behavior is reflected in the absence of these aspects in language models, which lack this source of knowledge. This implies that

even huge amounts of linguistic input cannot replace the multimodal sources from which linguistic knowledge is acquired in humans.

A suggestive hypothesis arising from the data is that LLMs may behave in a way similar to humans with low imagery and high vocabulary skills. Interestingly, electrophysiological and behavioral studies accounting for individual differences reported that different profiles emerge in metaphor processing (Battaglini et al., 2025) and that a greater reliance on the semantic/distributional route may also be present in humans. Shen et al. (2015) found that participants with low mental imagery abilities showed a greater neurophysiological response linked to semantic mismatch and no imagery activations when processing metaphors, while high-vocabulary participants are found to be less sensitive to sensorimotor features of words (Frau et al., 2025) but more sensitive to their semantic neighbors (Pexman and Yap, 2018).

This work confirms the importance of the study of figurative language and metaphorical abstraction to disentangle the subtle aspects of linguistic processing that distinguish between a formal and a functional human-like competence of language in large language models (Mahowald et al., 2024). The evidence reported in this study may add to the broader debate on which aspects of human language abilities LLMs are actually modeling (Dove, 2024). In particular, our results suggest that LLMs may be a valid model in cognitive science for those linguistic aspects relying on distributional features, while their limitations should be taken into account when considering language skills that require a multimodal conceptual representation.

5 Limitations

First of all, in the experiment, we chose to employ GPT2 models and, even if scaling does not necessarily improve human-like grounding (see Discussion), larger or multimodal models could provide different results for our metaphor production task. Moreover, all of the vehicles used in the task are concrete words and thus they potentially carry sensorimotor information, regardless of their BOI differences. So, the lack of sensitivity to this fine-grained feature does not exclude the possibility of a more general concreteness effect as found in Katz (1989). In addition, the use of only concrete words as vehicles did not allow us to explore the role of other experiential variables typically found in ab-

stract words (e.g., emotional and interoceptive properties) that may contribute to the processes at work in metaphor production for humans and machines.

6 Acknowledgments

This work received support from the European Research Council under the EU’s Horizon Europe program, ERC Consolidator Grant “PROcessing METaphors: Neurochronometry, Acquisition and Decay, PROMENADE” [101045733]. The content of this article is the sole responsibility of the authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

References

- Hamad Al-Azary and Lori Buchanan. 2017. Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory & Cognition*, 45:296–307.
- Hamad Al-Azary and Albert N Katz. 2021. Do metaphorical sharks bite? simulation and abstraction in metaphor processing. *Memory & Cognition*, 49:557–570.
- Hamad Al-Azary and Albert N Katz. 2023. On choosing the vehicles of metaphors 2.0: the interactive effects of semantic neighborhood density and body-object interaction on metaphor production. *Frontiers in Psychology*, 14:1216561.
- Valentina Bambini, Claudio Gentili, Emiliano Ricciardi, Pier Marco Bertinetto, and Pietro Pietrini. 2011. Decomposing metaphor processing at the cognitive and neural level through functional magnetic resonance imaging. *Brain research bulletin*, 86(3-4):203–216.
- Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2):379–400.
- Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.
- Chiara Battaglini, Federico Frau, Veronica Mangiaterra, Luca Bischetti, Paolo Canal, and Valentina Bambini. 2025. Imagers and mentalizers: capturing individual variation in metaphor interpretation via intersubject representational dissimilarity. *Accepted at Lingue e Linguaggio*. Preprint: <https://osf.io/n8x36>.

- Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.
- Stephen DR Bennett, A Nicole Burnett, Paul D Siakaluk, and Penny M Pexman. 2011. Imageability and body-object interaction ratings for 599 multisyllabic nouns. *Behavior research methods*, 43:1100–1109.
- Abeba Birhane and Marek McGann. 2024. Large models of what? mistaking engineering achievements for human linguistic agency. *Language Sciences*, 106:101672.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, and 1 others. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.
- Johan J Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. 2024. Three reasons why ai doesn't model human language. *Nature*, 627(8004):489.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, 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.
- Lori Buchanan, Chris Westbury, and Curt Burgess. 2001. Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, 8(3):531–544.
- Gaia Carenini, Louis Bodot, Luca Bischetti, Walter Schaeken, and Valentina Bambini. 2023. Large language models behave (almost) as rational speech actors: Insights from metaphor understanding. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.
- Robyn Carston. 2018. Figurative language, mental imagery, and pragmatics. *Metaphor and Symbol*, 33(3):198–217.
- Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.
- Guy Dove. 2024. Symbol ungrounding: what the successes (and failures) of large language models reveal about human cognition. *Philosophical Transactions B*, 379(1911):20230149.
- Kevin Durda and Lori Buchanan. 2008. Windsor: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40(3):705–712.
- Ivan Enrici, Bruno G Bara, and Mauro Adenzato. 2019. Theory of mind, pragmatics and the brain: Converging evidence for the role of intention processing as a core feature of human communication. *Pragmatics & Cognition*, 26(1):5–38.
- Federico Frau, Luca Bischetti, Lorenzo Campidelli, Elisabetta Tonini, Emiko J Muraki, Penny M Pexman, and Valentina Bambini. 2025. Understanding with the body? testing the role of verb relative embodiment across tasks at the interface of language and memory. *Journal of Memory and Language*, 140:104566.
- Raymond W Gibbs Jr and Teenie Matlock. 2008. Metaphor, imagination, and simulation: Psycholinguistic evidence. In Jr. R. W. Gibbs, editor, *The Cambridge handbook of metaphor and thought*, page 161–176. Cambridge University Press.
- John J Gumperz. 1982. *Discourse strategies*. Cambridge University Press.
- Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations. *arXiv preprint arXiv:2305.13264*.
- Nicholas Ichien, Dušan Stamenković, and Keith J Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.
- Cameron R Jones, Benjamin Bergen, and Sean Trott. 2024. Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, 50(4):1415–1440.
- Albert N Katz. 1989. On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and language*, 28(4):486–499.
- Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2):257–266.
- Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26.
- Brenden M Lake and Gregory L Murphy. 2023. Word meaning in minds and machines. *Psychological review*, 130(2):401.
- Jonghyun Lee, Dojun Park, Jiwoo Lee, Hoekeon Choi, and Sung-Eun Lee. 2025. Exploring multimodal perception in large language models through perceptual strength ratings. *arXiv preprint arXiv:2503.06980*.
- Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

- Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.
- Veronica Mangiaterra, Chiara Barattieri di San Pietro, and Valentina Bambini. 2024. Temporal word embeddings in the study of metaphor change over time and across genres: a proof-of-concept study on english. In *10th Italian Conference on Computational Linguistics, CLiC-it 2024*.
- Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.
- Nira Mashal, Tali Vishne, and Nathaniel Laor. 2014. The role of the precuneus in metaphor comprehension: evidence from an fmri study in people with schizophrenia and healthy participants. *Frontiers in human neuroscience*, 8:818.
- Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in ai’s large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.
- Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736.
- Allan Paivio. 1979. *Imagery and verbal processes*. Psychology Press.
- Allan Paivio and Mary Walsh. 1993. Psychological processes in metaphor comprehension and memory. In Andrew Ortony, editor, *Metaphor and Thought*, pages 2–307. Cambridge University Press.
- Penny M Pexman and Melvin J Yap. 2018. Individual differences in semantic processing: Insights from the calgary semantic decision project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(7):1091.
- Ben Prystawski, Paul Thibodeau, Christopher Potts, and Noah D Goodman. 2022. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Nick Reid, Hamad Al-Azary, and Albert N Katz. 2023. Cognitive factors related to metaphor goodness in poetic and non-literary metaphor. *Metaphor and Symbol*, 38(2):130–148.
- Zih-Yu Shen, Yi-Ting Tsai, and Chia-Lin Lee. 2015. Joint influence of metaphor familiarity and mental imagery ability on action metaphor comprehension: An event-related potential study. *Language and Linguistics*, 16(4):615–637.
- Paul D Siakaluk, Penny M Pexman, Christopher R Sears, Kim Wilson, Keri Locheed, and William J Owen. 2008. The benefits of sensorimotor knowledge: Body–object interaction facilitates semantic processing. *Cognitive Science*, 32(3):591–605.
- Dan Sperber. 1994. Understanding verbal understanding. *What is intelligence*, 179:98.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.
- Sherri M Tillotson, Paul D Siakaluk, and Penny M Pexman. 2008. Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, 40(4):1075–1078.
- Markus J Van Ackeren, Daniel Casasanto, Harold Bekkering, Peter Hagoort, and Shirley-Ann Rueschemeyer. 2012. Pragmatics in action: indirect requests engage theory of mind areas and the cortical motor network. *Journal of cognitive neuroscience*, 24(11):2237–2247.
- Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.
- Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.
- Deirdre Wilson and Robyn Carston. 2007. A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts, editor, *Pragmatics*, page 230–259. Palgrave Macmillan, Basingstoke.
- Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2023. Does conceptual representation require embodiment? insights from large language models. *arXiv preprint arXiv:2305.19103*.
- Melvin J Yap, Penny M Pexman, Michele Wellsby, Ian S Hargreaves, and Mark J Huff. 2012. An abundance of riches: Cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in human neuroscience*, 6:72.

A Appendix

Examples of metaphors with higher probability assigned by GPT2

Boredom is a vacuum (High BOI -Low SND)

Persuasion is an eagle (Low BOI – Low SND)

Prestige is a tiger (Low BOI – Low SND)

Destiny is a bicycle (High BOI – High SND)

Narcissism is a lighthouse (Low BOI – Low SND)

Sadness is a puzzle (High BOI – Low SND)

Examples of metaphors produced by humans

Boredom is a desert (Low BOI – Low SND)

Persuasion is a butterfly (Low BOI – Low SND)

Prestige is a rainbow (Low BOI – Low SND)

Destiny is a seed (High BOI – Low SND)

Narcissism is a volcano (Low Boi- High SND)

Sadness is a cloud (Low BOI – Low SND)

Prompting Metaphoricity: Soft Labeling with Large Language Models in Popular Communication of Science Tweets in Spanish

Alec Sánchez-Montero

alecm@comunidad.unam.mx

Gemma Bel-Enguix

gbele@iingen.unam.mx
gsierram@iingen.unam.mx

Sergio-Luis Ojeda-Trueba

sojedat@iingen.unam.mx

Universidad Nacional Autónoma de México

Abstract

In this paper, we explore how large language models (LLMs) can be used to assign soft labels for metaphoricity in Popular Communication of Science (PCS) tweets written in Spanish. Instead of treating metaphors as a binary yes/no phenomenon, we focus on their graded nature and the variability commonly found in human annotations. Through a combination of prompt design and quantitative evaluation over a stratified sample of our dataset, we show that GPT-4 can assign probabilistic scores not only for general metaphoricity but also for specific metaphor types with consistency (Direct, Indirect, and Personification). The results show that, while LLMs align reasonably well with average human judgments for some categories, capturing the subtle patterns of inter-annotator disagreement remains a challenge. We present a corpus of 3,733 tweets annotated with LLM-generated soft labels, a valuable resource for further metaphor analysis in scientific discourse and figurative language annotation with LLMs.

1 Introduction

Automatic metaphor detection has undergone a significant evolution over the last decades, transitioning from traditional rule- and knowledge-based methods to statistical and machine learning methods, including supervised, semi-supervised, and unsupervised techniques (Zayed, 2021). Most recently, due to considerable advances in deep learning, the utilization of large language models (LLMs) has shown promising results in various natural language understanding (NLU) tasks, including metaphor detection and figurative language processing (Wachowiak and Gromann, 2023; Tian et al., 2024; Jia et al., 2025; Xu et al., 2024; Lin et al., 2024). However, as we have previously noted (Sánchez-Montero et al., 2025), research specifically focused on metaphor detection in Spanish based on annotation disagreement remains significantly limited.

This study addresses the intricate nature of metaphor annotation from an exploratory LLM perspective, a task we found to be characterized by inherent subjectivity and consequent disagreements among human annotators. We assume that these disagreements not only reflect the complexity of the task, but may also be symptomatic of the gradable nature of “metaphoricity”, where expressions possess different degrees of metaphorical quality (Hanks, 2006). Particularly, we believe that advanced LLMs, such as gpt-4o and gpt-4.1, could be suitable for multi-label classification of metaphors, given their ability to handle datasets exhibiting an uneven distribution of multiple categories (Cloutier and Japkowicz, 2023; Kostina et al., 2025), allowing us to distinguish between different types or degrees of metaphoricity that contribute to variability in annotation.

To explore this variability of interpretation and the ability of LLMs to reflect it, we rely on a prompted-based methodology using optimized GPT-4 models, chosen for their demonstrated capabilities in annotating textual data (Yu et al., 2023, 2024; Yu, 2025). Our main objective is to investigate how different prompting strategies may influence the LLM’s ability to identify cases where metaphoricity is ambiguous or susceptible to multiple interpretations, paralleling the disagreements found in human annotation of Mexican Spanish Public Communication of Science (PCS) tweets. This approach deepens our understanding of metaphor gradability, a core concept in analogical reasoning, while also holding practical value for NLU in Spanish, where AI systems must grasp metaphor and figurative language to more accurately interpret and respond to human communication. A key contribution of our resource lies in its incorporation of soft labeling and the use of LLM-based reasoning to complement human annotation. This paper is structured as follows: Section 2 provides the necessary background on key concepts

and related work; Section 3 details the characteristics of the dataset used in this study; Section 4 outlines the methodology employed, including the prompt design, the experimental setup, and the results obtained, followed by concluding remarks in Section 5.

2 Background

2.1 Foundational Concepts

Linguistic Metaphor. According to Conceptual Metaphor Theory (CMT), linguistic metaphor is the manifestation in natural language of conceptual metaphors, where one conceptual domain (source) is used to understand another (target) through a structured mapping of entities and relationships (Lakoff and Johnson, 1980). Metaphors are not mere stylistic devices or figures of speech, but fundamental phenomena shaping human cognition and grounded in our bodily experiences.

Metaphor Annotation. Metaphor annotation presents challenges due to a lack of methodological consistency and variability in intuitions, making comprehensive corpora characterization and comparison across studies difficult (Veale et al., 2016). To address this, the Metaphor Identification Procedure Vrije Universiteit Amsterdam (MIPVU) (Steen et al., 2010) (developed initially by (Praggle-jaz, 2007) as Metaphor Identification Procedure, or MIP) offers a widely adopted systematic methodology for identifying potentially metaphorical linguistic units or metaphor-related words (MRWs), which encompass indirect, direct, and implicit metaphorical expressions, as well as explicit signals of metaphor and instances of personification. Beyond MIPVU, there are other approaches, such as the Deliberate Metaphor Identification Procedure (DMIP) (Reijnierse et al., 2017), focusing on deliberate metaphors from a semiotic and communicative perspective (Steen, 2008), and annotation schemes that extend identification to conceptual metaphors, annotating source and target domains (Shutova and Teufel, 2010).

Metaphoricity. The notion of metaphoricity refers to the gradual quality of a linguistic expression perceived as metaphorical, moving away from a strict binary categorization (Julich-Warpakowski and Jensen, 2023). This theoretical perspective recognizes the fuzzy boundaries between literal and figurative language, suggesting that some metaphors are “more metaphorical” than others (Hanks, 2006). The degree of metaphoricity can depend on factors

such as conventionality, the semantic or conceptual distance between source and target domains or conceptual frames (Bierwiazek, 2024), situational context, and inter-speaker variation (Julich-Warpakowski and Jensen, 2023). Understanding metaphoricity as a gradable phenomenon in NLP allows for modeling the subtleties and ambiguities that manifest in human annotation disagreement, derived from different interpretations of the potential metaphorical meaning of a linguistic expression.

LLMs and Prompt Engineering for Linguistic Annotation. Large language models (LLMs) represent a significant advancement in artificial intelligence, characterized by their ability to process and generate human-like text at scale, while becoming “the de facto baseline models to be used” in most NLP tasks (Zubiaga, 2024). Prompt engineering has emerged as a crucial technique for harnessing the capabilities of these models without extensive fine-tuning, involving the strategic design of textual inputs to guide desired outputs (Sahoo et al., 2024). Common strategies include zero-shot prompting (no examples), few-shot prompting (few examples) and instruction-based prompting. LLMs are increasingly being explored for linguistic annotation, including in tasks with significant human disagreement (Brown et al., 2025), and for processing phenomena such as figurative language and metaphor (Ichien et al., 2024). By providing LLMs with clear instructions and relevant context through well-designed prompts, researchers have shown that these models can perform various annotation tasks, sometimes achieving performance comparable to human annotators or outperforming them (Gilardi et al., 2023).

Learning from Disagreement Moving away from the traditional assumption of a single gold standard with hard labels and a single objective truth, the ‘learning from disagreement’ approach considers annotation discrepancies as valuable information, particularly for subjective linguistic tasks, such as figurative language annotation, where multiple interpretations coexist and intrinsic subjectivity generates variability (Uma et al., 2021). Rather than simply aggregating annotation disagreements into a single label and biasing models in favor of some linguistic theory, embracing disagreements allows for a richer representation of the inherent variability and gradability of subjective linguistic phenomena (Plank et al., 2014). Capturing this variability requires going beyond traditional hard

labels, using soft labels that represent the distribution or degree of human judgment. As probabilities or degrees of belief, soft labels can capture the inherent uncertainty and gradience of human annotation. Linguistic annotation, particularly for semantic interpretation and figurative language like metaphor, is inherently subjective due to variations in annotators' backgrounds, interpretations, and biases. As observed in our previous work on annotating metaphor in Spanish PCS tweets, this subjectivity resulted in significant inter-annotator disagreement (Sánchez-Montero et al., 2024, 2025). Unlike traditional hard metrics (e.g., F1, accuracy), soft evaluation metrics (e.g., cross-entropy, Manhattan distance, Euclidean distance, Jensen-Shannon divergence) are designed to compare probability distributions (Rizzi et al., 2024). This makes them suitable for evaluating models that produce soft or probabilistic outputs, which are necessary to capture the variability and gradable nature of subjective linguistic phenomena.

3 Dataset

As discussed by (Sanchez-Mora, 2016), Public Communication of Science (PCS) is a multidisciplinary field that encompasses a range of scientific disciplines and media platforms. It prioritizes accessibility and relevance for non-specialist audiences, often relying on metaphors to communicate complex ideas (Taylor and Dewsbury, 2018; Cormick, 2019). Our focus on Mexican Spanish PCS tweets stems from the scarcity of resources in this variety and genre. There is a limited pool of active science communicators on Twitter/X in Mexico, which necessarily constrains corpus size but also defines a domain that is underexplored and culturally meaningful.

The dataset utilized in this study comprises a corpus of 3733 Mexican Spanish tweets from the domain of Public Communication of Science (PCS), specifically annotated for metaphor detection from a multi-label annotation system. We have compiled this dataset from the timelines of 19 science communicators based in Mexico (January 2020 - May 2023). The information collected from these user accounts was obtained without targeting any specific scientific domain.

To our knowledge, this corpus is the first publicly documented effort to annotate linguistic metaphors specifically in Mexican Spanish PCS tweets. Although there are limited resources for metaphor

detection in Spanish, such as the CoMeta corpus (Sanchez-Bayona and Agerri, 2022), and previous work has explored other variants or domains (Martínez Santiago et al., 2014; Richi Pons-Sorolla, 2020; Alvarez Mouravskaia, 2020; Uribe and Mejía, 2024), there is a gap in publicly available linguistic metaphor corpora for Mexican Spanish that are suitable for exploring nuances and variability in metaphor annotation beyond simple binary classification of metaphorical expressions. While we acknowledge the existence of larger corpora in English, our decision to focus on Mexican Spanish is both intentional and necessary. We consider Spanish PCS tweets to offer a unique intersection of scientific and colloquial registers that pose rich and often ambiguous metaphorical constructions, ideal for examining gradable metaphoricity and inter-annotator variation in this language variant.

The dataset was annotated by six native Mexican Spanish-speaking linguistics students (2 female and 4 male), who independently labeled each tweet for the presence and type of metaphor. We divided the dataset into two halves, annotated independently by groups of three annotators each. Given the absence of a public adaptation of MIPVU for Spanish, we developed specific annotation guides for PCS tweets in Mexican Spanish, incorporating concepts from CMT (source/target domains, mapping). Annotators applied labels for the three types of metaphor that we established: Direct Metaphor (3 labels: source unit, target unit, signal), Indirect Metaphor (1 label: source unit) and Personification Metaphor (2 labels: personified object, personifier). Non-metaphorical tweets were indicated by saving them without annotations. Both our annotation guidelines and dataset are publicly available on our [GitHub repository](#). The principles of the Belmont Report were followed in the data annotation process (Belmont, 1978).

In a previous study (Sánchez-Montero et al., 2025), we presented a binary soft-labeled dataset of PCS tweets (metaphorical vs. non-metaphorical). In this paper, we introduce a new layer of analysis through fine-grained metaphor type soft annotation (Direct, Indirect, and Personification) and LLM-generated soft labels with reasoning traces. This expanded dataset provides a complementary perspective, which enriches the original binary setup with gradable metaphor judgments and interpretability signals from LLMs. A more detailed explanation of our multi-label annotation schema can be found

in Sánchez-Montero et al. (2024).

The inherent subjectivity of metaphor identification led to varying degrees of agreement among the annotators. To capture the spectrum of agreement and disagreement at the individual tweet level, we generated soft labels. For each tweet and each metaphor type (Direct, Indirect, Personification), a soft label was calculated as the proportion of annotators who assigned that label at both the binary and multi-label levels.

- For the **binary level**, the soft label represents the probability of a tweet being metaphorical, calculated as the proportion of annotators who labeled it as such (ranging from 0/3 to 3/3).
- For the **multi-class level**, for each tweet and each metaphor type (Direct, Indirect, Personification), a soft label was calculated as the proportion of annotators who assigned that specific label (ranging from 0/3 to 3/3). This allows for tweets to potentially have soft labels across multiple metaphor categories, reflecting the possibility of containing more than one type of metaphor.

Table 1 presents the distribution of tweets according to annotators' levels of agreement, represented by the soft labels, for both binary classification (metaphorical vs non-metaphorical tweets) and specific metaphor types. Our findings reveal a remarkable level of disagreement among annotators in all rating categories. Looking at the binary level, we see that almost half of the tweets (1780 tweets or 47.7%) showed some level of disagreement among annotators (1229 with 1/3 agreement and 551 with 2/3 agreement), in contrast to 1953 tweets (52.3%) where there was perfect consensus, although class 0 (non-metaphorical) is the most prevalent.

When examining specific metaphor types, the data in the 'Disagreement' columns (Soft Labels 0.33 and 0.66) further highlight the difficulty in consistently identifying and classifying Direct (with disagreement in 291 tweets, or 7.8%), Indirect (1340 tweets, or 35.9%, showed disagreement), and Personification (disagreement in 597 tweets, or 16.0%) metaphors. Furthermore, the relatively low number of tweets where annotators achieved perfect positive agreement (Label 1.0) for Direct (only 8), Indirect (99), and Personification (15), especially when considering the total size of the dataset, underscores how challenging it is to reach complete

consensus on the specific type of metaphor in each tweet.

Notably, across all metaphor categories analyzed, the distribution of inter-annotator agreement follows a consistent pattern from highest to lowest frequency in the soft labels: $0.0 > 0.33 > 0.66 > 1.0$. This distribution, where perfect positive agreement (Label 1.0) is consistently the least frequent outcome, may suggest that instances of what would be unanimously considered a clear metaphor are relatively rare in this corpus. Taken together, these distributions could support the perspective that metaphoricity exists along a spectrum of gradability, rather than conforming to strict, boundary-defining categories. This rich information about human disagreement, captured by our soft labels, served as the basis for our prompting experiments with LLMs.

4 Soft Metaphor Detection through Prompting

Our research employed a multiphase experimental methodology to explore and model the gradability of metaphoricity and conceptual mapping in the context of linguistic annotation with LLMs. This graded approach allowed us to refine LLM interaction strategies prior to large-scale evaluation and final corpus annotation, aiming to understand the LLM's ability to perform nuanced analogical abstractions on metaphor.

4.1 Qualitative Exploration and Prompt Design

The initial phase of our experimental approach consisted of a qualitative exploration of various prompting strategies to assess their potential ability to elicit GPT-4o responses that reflected the complexity and variability of metaphoricity. For this purpose, we selected a reduced set of 30 tweets from our dataset and applied the following prompt settings to the GPT-4o model:

- **Zero-Shot (ZS):** We asked the LLM for a binary classification (metaphorical/non-metaphorical tweets) with probability and reasoning process without prior definitions or examples.
- **One-Shot with Definition for Binary Classification (1S-Def-Bin):** We included a metaphor definition (source-target domain connection) and an example, requesting binary classification with probability and reasoning process.

Metaphor Category	Perfect Agreement			Disagreement		
	Label 0.0	Label 1.0	Total Tweets	Label 0.33	Label 0.66	Total Tweets
Binary Classification	1753	200	1953	1229	551	1780
Direct (D)	3434	8	3442	236	55	291
Indirect (I)	2294	99	2393	1000	340	1340
Personification (P)	3121	15	3136	498	99	597

Table 1: Soft-Label Distribution by Levels of Inter-Annotator Agreement in the Dataset

- Few-Shot with Definitions for Multi-label Classification (FS-Def-Multi): We provided definitions and examples for metaphor types (Direct, Indirect, Personification), requesting multi-label classification with probabilities for metaphorical tweets and reasoning process.
- Few-Shot with Definitions and Chain-of-Thought for Metaphorical Tweets (FS-Def-CoT-M): We added a step-by-step reasoning protocol (chain-of-thought) for multi-label classification, requesting the LLM to perform this process on tweets assumed to be metaphorical. The CoT sought to break down and guide the steps of LLM analogical reasoning for the fine classification of metaphor types.
- Few-Shot with Definitions and Chain-of-Thought for Binary and Multi-label Classification (FS-Def-CoT-BM): Combined the initial binary classification with multi-label classification by metaphor type, asking for probabilities and reasoning process, while applying the chain-of-thought protocol for metaphor identification. This integrated strategy sought to simulate a more complete analogical reasoning process (similar to that of human annotators), from the binary identification of the concept mapping to the detailed categorization of its type.
- Few-shot with Definitions, Chain-of-Thought, and Human Simulation (FS-Def-CoT-Sim): A variation of the previous prompt where an explicit instruction was added to the LLM to simulate the average of three human annotations when determining the binary probability and, if metaphorical, to follow the CoT process for multi-label classification with probabilities considering the same simulation.

The evaluation in this phase was primarily qualitative. We manually reviewed the LLM’s responses

to observe how it interpreted the instructions, its ability to identify potential metaphorical language and classify metaphor types according to the provided definitions, the quality and structure of its reasoning processes, and its capacity to assign probabilities that seemed to reflect uncertainty or degrees of metaphoricality.

Qualitatively, the ZS prompt showed some ability to assign non-binary probabilities and note subtle or conventionalized metaphors (e.g., assigning 0.6 to “plasticidad cerebral” [neuroplasticity] or 0.5 to “agujero de gusano” [wormhole], with justifications acknowledging their metaphorical origin or technical use). However, without explicit guidance, the consistency and alignment with our specific theoretical framework were less assured. The 1S-Def-Bin prompt appeared to guide the LLM more directly towards applying the provided definition based on source-target domain connection. Yet, interestingly, this prompt seemed to introduce more ‘doubt’ in the model for some tweets perceived as clearly non-metaphorical by human annotators (Soft Label 0.0), leading it to assign small but non-zero probabilities more frequently than the ZS prompt. For instance, in the case of “agujero de gusano,” which had perfect human agreement as metaphorical (Soft Label 1.0), the ZS prompt assigned a higher probability (0.5) and a justification more open to the term’s metaphorical origin than the 1S-Def-Bin prompt, which assigned a lower probability (0.3) arguing its technical use made it less metaphorical. This suggests that while a definition provides structure, it may sometimes override other signals the LLM captures in a zero-shot setting that are relevant to human judgment, leading to unexpected deviations.

For multi-label classification, the FS-Def-Multi prompt successfully elicited probabilities across different metaphor types, demonstrating the LLM’s capacity for multi-label soft assignment and for differentiating between distinct forms of analogical manifestation. The addition of a Chain-of-Thought (CoT) protocol in subsequent few-shot prompts

(like FS-Def-CoT-M and FS-Def-CoT-BM) generally led to more structured and detailed explanations, where the LLM explicitly broke down its reasoning based on domain identification and type characteristics, simulating the steps involved in the analogical reasoning for the classification of conceptual mappings by type. The most comprehensive strategy, FS-Def-CoT-BM, showed promise in simulating a multi-stage annotation process, reflecting a more complete analogical process.

For instance, for the tweet “Es como si solo tuvieras 93 tipos diferentes de piezas de Lego y con ellas pudieras armar todo el universo” (“It’s as if you only had 93 different types of Lego pieces and with them you could assemble the entire universe”) (Human labels: D=0.66, I=0.66), the FS-Def-Multi prompt assigned high probability to D (1.0) but low to Indirect (0.2), while a Few-Shot + CoT variant (referring to FS-Def-CoT(BM) here) assigned slightly lower to D (0.9) and higher to I (0.6), more closely reflecting the human annotators’ equal emphasis on both types. On another example, “Por primera vez, los científicos detectan los «gritos» de las plantas cuando son cortadas” (“For the first time, scientists detect the ‘screams’ of plants when they are cut”), there was a qualitative difference in interpretation: while human annotators saw a strong Indirect Metaphor and no Personification (Human labels: I=1.0, P=0.0), the LLMs (using FS-Def-Multi, FS-Def-CoT(BM), and FS-Def-CoT-Sim prompts) consistently assigned high probability to Personification (1.0, 1.0, 0.95 respectively) and low to Indirect (0.2, 0.4, 0.05), highlighting a divergence in how the models perform this specific analogical mapping compared to the human consensus in this instance.

The FS-Def-CoT-Sim prompt showed particular promise in its attempt to model the outcome of collective human judgment. Qualitatively, it sometimes produced binary probabilities that reflected intermediate levels of human disagreement. For instance, for the tweet “Cuando nace una estrella sigue agregando materia de la nube que se formó...” (“When a star is born it continues to add matter from the cloud that formed it...”), which had a human binary agreement of 0.66, this prompt assigned a binary probability of 0.40, providing a score within the disagreement range. Furthermore, this prompt’s multi-label assignments sometimes aligned well with human multi-label distributions even when the binary was intermediate. For the tweet “La dopamina interfiere en la función de tu

reloj interno...” (“Dopamine interferes with the function of your internal clock...”), while the human binary was 1.0, this prompt assigned 0.75; however, its multi-label score for Indirect (0.85) aligned closely with the human score (1.0), suggesting it could capture the specific type of analogical mapping even when its overall certainty differed.

To complement the qualitative exploration of prompting strategies, we calculated the Mean Absolute Difference (MAD) between the soft label assigned by the LLM and the corresponding human soft label for each tweet, averaging this value across the set of tweets tested with each prompt. This simple metric gives us an initial indication of the LLM’s closeness to human judgments on these examples. It is crucial to emphasize that these results are based on very small samples and are not generalizable to the full corpus. Table 2 presents the MAD for the prompting strategies evaluated in this phase, for both binary classification and the multi-label categories. A lower MAD indicates better preliminary alignment with human soft labels for that category and prompt strategy on the tested samples.

According to preliminary results, for binary classification, the strategy incorporating the Human Simulation instruction shows the lowest MAD (0.157), suggesting it may capture the overall presence/absence judgment with potentially better alignment to human consensus levels in this preliminary sample. For multi-label classification, the picture is more nuanced across categories. Looking at the Average MAD (Multi-label) across all three types, the “Few-shot with Definitions + CoT (Binary & Multi-label)” prompt shows a slightly lower average MAD (0.139) compared to the “Few-shot with Def. + CoT + Human Simulation” prompt (0.160) and the simpler multi-label prompts without CoT. While these results provide initial quantitative justification for selecting the most promising prompting strategies for larger-scale evaluation, an important consideration when implementing complex strategies like Chain-of-Thought (CoT) is the increased token consumption. This, in turn, translates to higher computational cost. However, the qualitative observation of more structured reasoning and the logical appeal of guiding the LLM through complex classification steps strongly suggest that CoT could lead to a more robust and interpretable model in its analogical processing, particularly for capturing the nuances and variability of metaphoricity. Similarly, the human simulation

Prompting	Binary	Direct	Indirect	Personif.	Avg. Multi
ZS	0.285	—	—	—	—
1S-Def-Bin	0.250	—	—	—	—
FS-Def-Multi	—	0.388	0.378	0.321	0.362
FS-Def-CoT-M	—	0.118	0.337	0.351	0.269
FS-Def-CoT-BM	0.198	0.084	0.185	0.149	0.139
FS-Def-CoT-Sim	0.157	0.198	0.180	0.102	0.160

Table 2: Preliminary quantitative results comparing the Mean Absolute Deviation (MAD) between LLM predictions and human annotators across prompting strategies.

strategy showed potential for eliciting responses that more closely approximated patterns of human agreement/disagreement.

4.2 Quantitative Evaluation on a Larger Sample

Following the qualitative exploration and preliminary quantitative analysis, the second phase of our methodology focused on conducting a more rigorous quantitative evaluation of promising prompting strategies, model configurations, and parameters on a larger sample of the corpus. The primary objective was to obtain statistically more robust metrics to assess the LLM’s ability to generate soft labels that could align with human annotation, capture disagreement patterns, and model the gradability of metaphoricity, with the aim of informing the selection of the final approach for full corpus annotation.

For this phase, a stratified random sample of 750 tweets (ca. 20% of the corpus) was selected from the total 3733 tweets. Stratification ensured that the sample represented the distribution of soft labels observed in the full dataset, reflecting the varying levels of human agreement encountered in the data, from clear cases to instances of significant disagreement. We conducted several experiments by applying different configurations to this sample, including baseline zero-shot prompting, few-shot prompting with and without human simulation instruction and a brief reasoning protocol, prompts that included more extensive elements from the annotation guide and additional few-shot examples, as well as model and temperature tuning. For all experiments in this phase, only the probabilistic soft labels (binary and multi-label) were requested as output from the LLM; reasoning processes were not included in the output.

For each category (Binary, Direct, Indirect, Personification), we computed the Mean Absolute Difference (MAD), Pearson Correlation, and Binary

Cross-entropy between the LLM’s soft labels and the corresponding human soft labels across the 750-tweet sample. Table 3 presents these metrics for all tested configurations.

Analysis of the metrics reveals that replicating human judgments varies significantly across configurations and metaphor categories. While several few-shot configurations achieved low MADs and Binary Cross-entropy for Direct metaphor (indicating good average alignment), the Pearson Correlation across all categories and configurations remains relatively low. This highlights the challenge in getting an advanced LLM to replicate the specific tweet-level patterns of human disagreement.

Overall, the gpt-4.1 (Few-shot + Gradable Examples + Human Simulation) configuration stands out in terms of capturing the overall linear trend and variability of human judgments, particularly for the crucial binary classification (highest Pearson Correlation). While some gpt-4o configurations, especially with temperature tuning, show competitive or slightly better MAD and CE for certain categories, the superior binary correlation of the gpt-4.1 configuration makes it the most promising for modeling the gradability of metaphoricity and aligning with human soft labels. Given the importance of the binary decision as a precursor to multi-label classification, and the potential for better capturing the spectrum of agreement, we selected the gpt-4.1 configuration for the final corpus annotation.

4.3 Corpus Annotation with LLM

After deciding on the LLM configuration identified and validated in the previous phase, we instructed GPT-4.1 to annotate the full corpus of 3733 PCS tweets. The objective was to generate a comprehensive dataset annotated with LLM-assigned soft labels for metaphoricity, capturing both binary presence and multi-label classification across different types, while also incorporating elements to facil-

Prompt Strategy / Model	Pearson Correlation (\uparrow)				MAD (\downarrow)				Binary Cross-entropy (\downarrow)			
	Bin	Dir	Ind	Per	Bin	Dir	Ind	Per	Bin	Dir	Ind	Per
gpt-4.1 (Few-shot + Gradable Ex + HumSim)	0.392	0.289	0.216	0.185	0.216	0.048	0.165	0.073	4.589	0.797	3.529	1.877
gpt-4o + Few-shot + Guide Details	0.293	0.133	0.168	-0.018	0.236	0.031	0.245	0.077	2.290	0.863	1.848	1.986
gpt-4o Few-shot - Human Simulation	0.214	0.304	0.121	0.015	0.252	0.031	0.214	0.073	4.857	0.849	3.603	1.994
gpt-4o Few-shot + Human Simulation	0.240	0.291	0.137	0.024	0.244	0.031	0.211	0.076	4.511	0.860	3.666	1.985
gpt-4o temp 0.2 Few-shot + Human Simulation	0.276	0.145	0.109	-0.016	0.232	0.036	0.193	0.073	4.134	0.845	3.132	2.002
gpt-4o temp 0.5 Few-shot + Human Simulation	0.293	0.221	0.141	0.088	0.229	0.034	0.184	0.075	4.014	0.819	2.985	1.926
gpt-4o temp 0.7 Few-shot + Human Simulation	0.303	0.161	0.158	0.045	0.227	0.034	0.183	0.073	3.888	0.833	3.183	1.981
gpt-4o Zero-shot	0.254	0.013	0.150	0.033	0.242	0.034	0.179	0.073	5.791	0.887	3.951	1.982

Table 3: Quantitative Soft Evaluation Metrics Comparison (LLM vs Human Soft Labels on 750 Tweets)

itate potential semi-supervised refinement in the future.

The design of the final prompt, refined through experimentation in previous phases, aimed to improve the LLM’s sensitivity in automatic metaphor detection, particularly for nuances in Direct and Personification metaphors (which were the most difficult to identify consistently during the previous phases). We also sought to model responses that reflected the inter-annotator variability observed in the human soft labels by incorporating intermediate examples and the explicit simulation instruction.

The optimized Few-shot prompt, including clear definitions and examples for Direct, Indirect, and Personification metaphors, along with the human simulation instruction and an internal structured reasoning process, was applied to each tweet. The model was also instructed to simulate the average of three human annotations and provide a binary probability between 0 (non-metaphorical) and 1 (metaphorical). If the binary probability was ≥ 0.5 (classified as metaphorical), soft probabilities between 0 and 1 for each of the three metaphor types were also requested. To facilitate potential future analysis or semi-supervised manual review, a brief justification for the classification was included in the output only for tweets with binary probability ≥ 0.3 , corresponding to the lowest probability

for considering a tweet as having some degree of perceived metaphoricity by human annotators.

The resulting LLM-annotated corpus consists of 3733 tweets, each associated with a binary soft label, multi-label soft labels (if classified as metaphorical), and a brief reasoning text (for tweets with a perceived metaphoricity ≥ 0.3). For the chosen gpt-4.1 configuration, the metrics comparing LLM predictions to human soft labels were:

- **Binary:** Pearson Correlation: **0.382**, MAD: **0.215**, Cross-Entropy: **4.229**
- **Direct:** Pearson Correlation: **0.295**, MAD: **0.053**, Cross-Entropy: **0.769**
- **Indirect:** Pearson Correlation: **0.279**, MAD: **0.165**, Cross-Entropy: **3.322**
- **Personification:** Pearson Correlation: **0.124**, MAD: **0.069**, Cross-Entropy: **1.735**

These metrics indicate that while the LLM’s soft labels show a degree of alignment with human soft labels (particularly low MAD for Direct and Personification, and the highest correlation for Binary), its ability to precisely replicate the tweet-level variability and complex patterns of human disagreement remains limited, as evidenced by the low Pearson correlations across all categories. Direct metaphors showed the best average alignment

(lowest MAD) and lowest probabilistic error (lowest CE). However, based on our evaluation, Direct and Personification categories, while sometimes having low MAD, presented significant challenges for the LLM in achieving high correlation with human judgments, indicating difficulty in consistently capturing the nuances of disagreement for these specific types.

5 Conclusions and Future Work

This study explored the use of large language models (LLMs) to generate soft labels for metaphoricity in Public Communication of Science tweets written in Spanish, aiming to capture gradability and reflect human annotation variability. Through a phased approach involving prompt engineering, model evaluation, and annotation of a 3733-tweet corpus, we demonstrated that LLMs can effectively produce probabilistic soft labels for binary metaphoricity and specific types (Direct, Indirect, Personification Metaphors).

Prompt design significantly impacted performance. Quantitative evaluation revealed that while LLMs achieve reasonable average alignment with human soft labels, they face challenges in consistently replicating the tweet-level patterns of human disagreement. Pearson correlations were relatively low across all categories, highlighting this limitation in modeling human variability. Despite this, the resulting LLM-annotated corpus is a valuable resource for analyzing metaphor and metaphoricity in scientific discourse.

Future work should focus on strategies to improve LLM alignment with the precise patterns of human disagreement, potentially through advanced prompting techniques, fine-tuning on soft-labeled data, or leveraging ensemble annotation strategies. Further analysis and application of the annotated corpus to downstream tasks, such as studying metaphor trends or enabling semi-supervised annotation pipelines, remains a promising direction. This research validates LLMs as a scalable tool for complex linguistic annotation, and can serve as a basis for exploring semi-supervised approaches or future research on LLM capabilities in complex linguistic annotation tasks related to analogical mapping.

Although existing work has shown that LLMs often reflect dominant or surface-level views while failing to capture minority or nuanced perspectives (Santurkar et al., 2023; Sourati et al., 2025), our

findings suggest that figurative language presents a more complex challenge than a simple majority/minority opinion divide. Metaphor understanding involves analogical reasoning, cultural grounding, and subjective interpretation—dimensions that do not always align with demographic or opinion group boundaries. Nonetheless, the broader concern about the homogenizing tendencies of LLMs resonates with our observation that LLMs often struggle to model fine-grained human disagreement. As such, we see metaphor annotation as a compelling testbed for probing alignment, interpretability, and diversity in LLM behavior, and advocate for more work at the intersection of linguistic theory, annotation practices, and model development, particularly for figurative understanding in languages beyond English.

Limitations

This study encountered several limitations inherent in the application of large language models (LLMs) to complex linguistic annotation tasks, particularly in replicating the nuances of human soft labels for metaphoricity. A primary limitation is the LLM’s demonstrated difficulty in consistently capturing the fine-grained patterns of human disagreement and variability at the tweet level. While quantitative evaluation showed that the LLM could achieve reasonable average alignment with human soft labels for certain metaphor categories (indicated by low Mean Absolute Difference and Binary Cross-entropy), the relatively low Pearson correlation coefficients across all categories highlight that the model did not accurately replicate the specific instances of high or low human consensus for individual tweets.

Furthermore, the performance varied across metaphor types. While Direct metaphors generally showed better average alignment, capturing the variability for both Direct and Personification categories proved challenging, with particularly low correlation observed for Personification. Indirect metaphors also presented difficulties in achieving strong alignment across metrics. This differential performance suggests that certain types of analogical mapping may be harder for current LLMs to model in a way that fully reflects human cognitive processing and social consensus.

Another limitation lies in the inherent constraints of the prompting approach. While prompt engineering significantly influenced the LLM’s perfor-

mance, the specific instructions, examples, and simulation requests used may not fully capture the multifaceted cognitive processes and contextual factors that contribute to human metaphorical judgment and inter-annotator variability. The reliance on a specific family of LLMs (GPT models) and the characteristics of the scientific tweet dataset also represent potential limitations to the generalizability of our findings. Future work should address these limitations by exploring alternative methodologies, models, and datasets to improve the replication of human disagreement patterns in LLM-based linguistic annotation.

Aknowledgements

This research was supported by UNAM through the PAPIIT project IG400325 and CONAHCYT (SE-CIHTI) CF-2023-G-64. Additional support was provided by the Sistema Nacional de Investigadoras e Investigadores (SNII) through a Research Assistant Scholarship.

References

- Kevin Alvarez Mouravskaia. 2020. Metaphor identification for spanish sentences using recurrent neural networks. Master’s thesis, Pontificia Universidad Católica del Perú.
- Informe Belmont. 1978. Principios éticos y directrices para la protección de sujetos humanos de investigación. *Estados Unidos de Norteamérica: Reporte de la Comisión Nacional para la Protección de Sujetos Humanos de Investigación Biomédica y de Comportamiento*.
- Bogusław Bierwiazzonek. 2024. [On the gradability of metaphor](#). *Studies in Logic, Grammar and Rhetoric*, 69(1):31–56.
- Megan A. Brown, Shubham Atreja, Libby Hemphill, and Patrick Y. Wu. 2025. [Evaluating how llm annotations represent diverse views on contentious topics](#). *arXiv preprint*.
- Nicolas Antonio Cloutier and Nathalie Japkowicz. 2023. [Fine-tuned generative LLM oversampling can improve performance over traditional techniques on multiclass imbalanced text classification](#). In *2023 IEEE International Conference on Big Data (Big-Data)*, pages 5181–5186. IEEE.
- Craig Cormick. 2019. *The science of communicating science: the ultimate guide*. CSIRO Publishing.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30).
- Patrick Hanks. 2006. [Metaphoricity is gradable](#), page 17–35. Mouton de Gruyter.
- Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. [Large language model displays emergent ability to interpret novel literary metaphors](#). *Metaphor and Symbol*, 39(4):296–309.
- Kaidi Jia, Yanxia Wu, Ming Liu, and Rongsheng Li. 2025. [Curriculum-style data augmentation for llm-based metaphor detection](#). *Preprint*, arXiv:2412.02956.
- Nina Julich-Warpakowski and Thomas Wiben Jensen. 2023. [Zooming in on the notion of metaphoricality: Notions, dimensions, and operationalizations](#). *Metaphor and the Social World*, 13(1):16–36.
- Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. [Large language models for text classification: Case study and comprehensive review](#). *Preprint*, arxiv:2501.08457 [cs].
- George Lakoff and Mark Leonard Johnson. 1980. *Metaphors we live by*. University of Chicago press.
- Yujie Lin, Jingyao Liu, Yan Gao, Ante Wang, and Jinsong Su. 2024. [A dual-perspective metaphor detection framework using large language models](#). *Preprint*, arXiv:2412.17332.
- Fernando Martínez Santiago, Miguel Ángel García Cumbreiras, Manuel Carlos Díaz Galiano, and Arturo Montejó Ráez. 2014. Etiquetado de metáforas lingüísticas en un conjunto de documentos en español. *Procesamiento del Lenguaje Natural*, (53):35–42.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. [Linguistically debatable or just plain wrong?](#) In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Pragglejaz. 2007. [MIP: A method for identifying metaphorically used words in discourse](#). *Metaphor and Symbol*, 22(1):1–39.
- W. Gudrun Reijniere, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2017. [Dmip: A method for identifying potentially deliberate metaphor in language use](#). *Corpus Pragmatics*, 2(2):129–147.
- Mateo Richi Pons-Sorolla. 2020. [Analizador de lectura fácil 4.0: identificación de metáforas](#).
- Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. [Soft metrics for evaluation with disagreements: an assessment](#). In *Proceedings of the 3rd Workshop on Perspective Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. [A systematic survey of prompt engineering in large language models: Techniques and applications](#). *arXiv preprint*.
- Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. [Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Alec Sánchez-Montero, Gemma Bel-Enguix, and Sergio-Luis Ojeda-Trueba. 2024. [Evaluating the development of linguistic metaphor annotation in Mexican Spanish popular science tweets](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 59–64, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. [Disagreement in metaphor annotation of Mexican Spanish science tweets](#). In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.
- M. Carmen Sanchez-Mora. 2016. [Hacia una taxonomía de las actividades de comunicación pública de la ciencia. 1824 - 2049](#), pages 1–9.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Ekaterina Shutova and Simone Teufel. 2010. [Metaphor corpus annotated for source - target domain mappings](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. [The shrinking landscape of linguistic diversity in the age of large language models](#). *Preprint*, arXiv:2502.11266.
- Gerard Steen. 2008. [The paradox of metaphor: Why we need a three-dimensional model of metaphor](#). *Metaphor and Symbol*, 23(4):213–241.
- Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. [A Method for Linguistic Metaphor Identification: From MIP to MIPVU](#), volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company.
- Cynthia Taylor and Bryan M. Dewsbury. 2018. [On the problem and promise of metaphor use in science and science communication](#). *Journal of Microbiology & Biology Education*, 19(1):19.1.46.
- Yuan Tian, Nan Xu, and Wenji Mao. 2024. [A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning](#). 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 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Stephany Nieves Uribe and Jorge Mauricio Molina Mejía. 2024. [Hacia una extracción semiautomática de metáforas conceptuales en un corpus de economía a partir del procesamiento de lenguaje natural](#). *Estudios de Lingüística Aplicada*, (76):81–109.
- Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. [Metaphor: a computational perspective](#). Number 31 in *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers.
- Lennart Wachowiak and Dagmar Gromann. 2023. [Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032. Association for Computational Linguistics.
- Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. [Exploring chain-of-thought for multimodal metaphor detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.
- Danni Yu. 2025. [Towards LLM-assisted move annotation: Leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports](#). *English for Specific Purposes*, 78:33–49.
- Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023. [Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology](#). *arXiv*. Publisher: arXiv Version Number: 5.
- Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. [Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology](#). *International Journal of Corpus Linguistics*, 29(4):534–561.

Omnia Zayed. 2021. *Metaphor processing in tweets*.
Master's thesis, NUI Galway.

Arkaitz Zubiaga. 2024. *Natural language processing
in the era of large language models*. *Frontiers in
Artificial Intelligence*, 6.

HATS : Hindi Analogy Test Set for Evaluating Reasoning in Large Language Models

Ashray Gupta^{*†}, Rohan Joseph^{*†}, & Sunny Rai[◊]

[†]Mahindra University, [◊]University of Pennsylvania
ashray20ucam008@mahindrauniversity.edu.in
rohan18545@mechyd.ac.in

Abstract

Analogies test a model’s ability to infer implicit relationships between concepts, making them a key benchmark for evaluating reasoning capabilities. While large language models (LLMs) are widely evaluated for reasoning in English, their abilities in Indic languages remain understudied, limiting our understanding of whether these models generalize across languages. To address this gap, we introduce a new Hindi Analogy Test Set (HATS), comprising 405 multiple-choice questions sourced from Indian government exams. We benchmark state-of-the-art multilingual LLMs using various prompting strategies and introduce a grounded Chain of Thought approach that leverages cognitive theories of analogical reasoning. This approach improves model performance on Hindi analogy questions. Our experiments show that models perform best with English prompts, irrespective of the prompting strategy. Our test set addresses the lack of a critical resource to evaluate LLM reasoning capabilities in Hindi. The test set is publicly available for research purposes here https://github.com/Inequilazitive/HATS-Hindi_Analogy_Test_Set.

1 Introduction

Self-supervised learning enabled language models to learn the notion of *similarity* and *relatedness*. However, abstraction and conceptualization as in analogies, are still a challenge. Growing research on common reasoning tasks including analogies (Ushio et al., 2021; Czinczoll et al., 2022; Bhavya et al., 2022), Winograd Schema Challenge (Liu et al., 2022; Emami et al., 2018), figurative text processing (Joseph et al., 2023; Bogireddy et al., 2023), reflects the trend to teach and evaluate LLMs on these tasks.

^{*}These authors contributed equally to this work.

Assessing reasoning abilities of LLMs in *low-resource languages* remains challenging (Robinson et al., 2023), primarily due to the scarcity and poor quality of available linguistic data (Khade et al., 2024), as well as the need for improved evaluation methodologies (Valmeekam et al., 2022; Wijesiriwardene et al., 2023; Bender and Koller, 2020). In this paper, we address this resource and knowledge gap by:

- Introducing **HATS**, a test set of 405 of in-situ semantic analogies curated from national and state-level administrative examinations and their preparatory material.
- Benchmarking state-of-the-art multilingual LLMs (see Sec 3.1) with diverse prompting strategies to evaluate LLMs’ reasoning abilities in Hindi.
- Proposing a grounded Chain of Thought prompting technique that leverages cognitive theories of analogical reasoning and improves model performance on Hindi analogy tasks (see Sec 3.5.2).

Existing datasets of Hindi analogies are primarily developed by translating English analogies and comprise only syntactic relations (Abdou et al., 2018; Grave et al., 2018). The translated analogies are used to test the quality of Hindi word embeddings (Gaikwad and Haribhakta, 2020) and LLMs trained on Hindi corpus (Kakwani et al., 2020). These datasets lack samples illustrating semantic relations between concepts specific to the Hindi language. This reflects the urgent need for resources to evaluate common reasoning in LLMs in the Indic language.

In this paper, we focus on *proportional analogy* comprising four words of the form $A : B :: C : D$ that is, A is to B as C is to D. Prior works introduced word-family based analogies exploiting

syntactic relations such as *singular-plural* (Abdou et al., 2018). We focus on semantic analogies.

2 HATS: Hindi Analogy Test Set

We scraped 405 analogy questions from national and state-level administrative service examinations and preparatory materials, including those for UPSC, SSC, PSC, Clerk, Defense, Railway, and Banking exams, using *BeautifulSoup* (Richardson, 2024). These analogies are designed to assess the aptitude and reasoning abilities of candidates.

Example

भोपाल (Bhopal): मध्य प्रदेश (Madhya Pradesh) :: भुवनेश्वर (Bhubaneswar): ?

A गुजरात (Gujarat)

B उड़ीसा (Odisha)

C राजस्थान (Rajasthan)

D अरुणाचल प्रदेश (Arunachal Pradesh)

Correct Answer: उड़ीसा (Odisha), since भुवनेश्वर (Bhubaneswar) is its capital, just as भोपाल (Bhopal) is the capital of मध्य प्रदेश (Madhya Pradesh).

The original multiple-choice questions appeared in varied formats. We standardized them to the $A : B :: X : Y$ structure and replaced Y with a question mark for model input. We also provide four options that were originally provided with these questions in examinations (See above example).

3 Benchmarking LLMs on HATS

3.1 Models

We evaluated three state-of-the-art multilingual LLMs: **Aya-expansive-8B** (Dang et al., 2024), **Llama-3.1-8B** (Grattafiori et al., 2024), and **Gemma-2-9B** (Team et al., 2024). These models were selected for their strong performance on multilingual and general-purpose language understanding benchmarks, and their accessibility for academic research (Cohere For AI Team, 2024).

3.2 Task A: Find the Most Likely Answer

We create a low-demand (i.e., forced-choice over a fixed set of answer options) task similar to (Hu and Frank, 2024) by presenting the model with an analogy truncated at the last colon ($A : B :: X :$). We select the most likely option as the answer using direct probability measurement. Since we avoid met-

alinguistic judgment, we chose non-instruct variants of models for this task.

We measured the accuracy of the models using normalized success rates (see Table 1). LLaMA outperforms Aya by 7.46% and Gemma by 6.85%. Overall, model performance in this setting remains suboptimal.

3.3 Prompt Design and Evaluation for Generation-Based Tasks

This section outlines the shared design principles and evaluation methodology used across Tasks B and C, both of which involve analogy completion using LLMs. The tasks differ in their prompting strategies but rely on a common structure, a system and user prompt template where we present the task-specific instructions and incomplete analogy with multiple-choice options. For these instruction-centric task settings, we utilize instruction-tuned model variants (see Appendix A for model specifications and prompt details).

Setting: To assess the impact of language on reasoning, prompts are evaluated under three configurations: (i) *Hindi-only* (both system and user prompts are in Hindi), (ii) *English-only* (both system and user prompts are in English), and (iii) *Mixed* (English system prompt and Hindi user prompt).

Evaluation: To mitigate positional bias in multiple-choice evaluations, we apply a cyclic rotation of the answer options. For a question with n options (typically $n = 4$), we generate n variants, each with the options shifted in position. The model answers all n variants, and the final answer is determined by majority voting across its n responses. A question is marked correct only if the majority-selected answer matches the ground truth; otherwise, it is considered incorrect. Detailed results are discussed in Section 3.6.

3.4 Task B: 0-Shot Prompting

Recent surveys and empirical studies highlight zero-shot prompting as a standard baseline for LLM evaluation, often used to benchmark models before exploring few-shot or fine-tuned settings. (Li, 2023). In the experiments carried out by (Reynolds and McDonnell, 2021), the authors show that well-crafted zero-shot prompts can, in fact, surpass the performance of few-shot prompts.

This baseline setting mimics the original exam-style format of the test set. For this task all the instructions were presented in the system prompt.

Model	Llama 3.1–8B	Aya Expanse–8B	Gemma 2–9B
Accuracy	46.17	42.96	43.20

Table 1: Accuracy (%) on Task A across all HATS samples. Each score represents the percentage of instances where the model correctly identified the answer option with the highest predicted likelihood.

Sys + User	Prompting	aya-expanse-8B	Llama-3.1-8B-instruct	gemma-2-9b-it
Hi+Hi	0-Shot	62.71	67.90	73.08
	0-Shot CoT	62.71	67.40	74.81
	Grounded 0-Shot-CoT	60.74	64.93	75.31
	Grounded FS-CoT	56.04	62.96	76.54
En+Hi	0-Shot CoT	63.70	64.69	76.05
	Grounded 0-Shot-CoT	61.23	64.93	75.80
	Grounded FS-CoT	59.50	65.67	75.31
En+En	0-Shot	65.67	71.85	78.77
	0-Shot CoT	65.43	66.91	78.52
	Grounded 0-Shot-CoT	65.43	74.56	79.75
	Grounded FS-CoT	61.72	74.07	77.28
	FS Translate-CoT	62.46	72.83	77.04

Table 2: Accuracy (%) across prompting strategies grouped by language setting. CoT = Chain-of-Thought, FS = Few-Shot. Best scores per setting are bolded. Refer to Section A.2.1 for prompt details. Accuracy is calculated only for valid analogies.

Mixed setting was not evaluated separately, as the prompt content is equivalent to English only in practice.

3.5 Task C: Chain of Thought Prompting

Prior work shows that prompting the model to reason step-by-step enhances LLM performance (Brown et al., 2020; Wei et al., 2023; Zhang et al., 2025).

3.5.1 0–Shot Chain of Thought

For this task we have taken a similar approach to (Kojima et al., 2023), and appended "Let's think step by step" at the end of the prompt.

3.5.2 Grounded 0-Shot Chain of Thought

We build on the (Wang et al., 2023) approach to guide the model's reasoning by presenting a fixed sequence of steps to solve analogies in the prompt. The steps are grounded in cognitive theories of analogical reasoning. Drawing on the (Minnameier, 2010) framework, the prompt integrates abductive structure identification, inductive concept mapping, and adequacy-based evaluation.

3.5.3 Grounded Few Shot Chain of Thought

Previous works use few shot examples for prompt based grounding (Mialon et al., 2023). In this

task we use the same prompt as in section 3.5.2 with 5 worked out examples. We guided *Claude-3.7-Sonnet** to generate Hindi examples, solved using our Grounded CoT instructions. The examples were verified and corrected by an expert of the Hindi language.

3.5.4 Few Shot Chain of Thought (with Translation)

Following the benchmark results, which showed LLMs performed best in English-only settings (see Table 2), we explored whether a translation-based approach could further improve performance on Hindi analogy tasks. Specifically, we implemented a three-step Chain of Thought (CoT) prompting strategy in English (see Sec 3.3):

- **Translation:** Convert the Hindi analogy and options into English.
- **Solution:** Solve the analogy using the method in Section 3.5.2.
- **Mapping:** Identify the correct Hindi option based on the English solution.

*<https://www.anthropic.com/news/claude-3-7-sonnet>

We included 5 worked out examples in the prompt. The examples were created using the process described in Section 3.5.3 with updated instructions.

3.6 Results

The accuracy scores are presented in Table 2. Prompts in English-only settings consistently led to the highest overall performance. Transitioning from baseline 0-Shot CoT to Grounded 0-Shot CoT resulted in an average improvement of +0.27 points across all models and settings. Gemma was the top performer, achieving the highest accuracy of 79.75% with Grounded 0-Shot Chain-of-Thought prompting (see Sec 3.5.2) in the English-only setting. LLaMA also performed best with Grounded 0-Shot CoT in the English-only setting, reaching an accuracy of 74.56%. In contrast, Aya was the weakest performer, with its highest score being 65.67%, obtained using 0-Shot prompting (see Sec 3.4) in the English-only setting. Some models struggled to follow instructions in Hindi, resulting in better performance with simpler 0-Shot CoT prompts compared to the more complex Grounded CoT setup.

4 Discussion

Gemma consistently outperformed other models by an average margin of 11.46 points across all tasks and exhibited minimal performance drop across different prompt settings. All models performed best when both system and user prompts were in English. Chain-of-Thought (CoT) reasoning boosted accuracy, especially in Few-Shot settings.

- While models reliably identified analogical pairs (A : B), they often failed to transfer the relation correctly to (C : D), highlighting limitations in structured reasoning.
- In the translation task, models like *aya-expansive-9b* and *LLaMA-3.1-8B-IT* frequently mistranslated critical terms. For example, the analogy फूल : माला :: ईट : ? (Flower : Garland :: Brick : ?) was misinterpreted as Flower : Garland :: Eat : ?, confusing ईट (Brick) with "Eat" due to phonetic similarity. This error was consistent across all 10 sampled failures.
- Models occasionally defaulted to "I don't know" or "None of the above," even when correct options were available.

- See Table A6 for model response languages across different task settings.

5 Conclusion

We introduced a test set **HATS** comprising 405 semantic analogies in Hindi. The benchmarking code and prompts for all tasks will be made publicly available. We designed five tasks to evaluate LLMs reasoning abilities in the Hindi language. These tasks assessed the reasoning abilities of LLMs in natural language and the usability of *translation* in creating low-resource language resources. Our experiments reveal the subpar performance of state-of-the-art LLMs when tested on HATS, highlighting the need to evaluate multilingual models on native language resources to better gauge their usability for non-English languages.

Limitations

In this study, we utilized smaller versions of the model (8B to 9B) due to resource and hardware constraints, and we anticipate models with higher parameters to perform better.

Ethics Statement

The test set is built from publicly available national level QPs and preparatory material. This ensures that the data is free from (a) anonymity concerns, (b) obscenities and (c) any stereotyping or bias. We have provided a Hindi language resource to evaluate the reasoning abilities of LLMs with the goal to make AI technology accessible to a wider population. We have not performed model training/finetuning and therefore, no significant carbon footprints were generated. We have chosen open source models for this work.

References

- Mostafa Abdou, Artur Kulmizev, and Vinit Ravishanker. 2018. Mgad: Multilingual generation of analogy datasets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.
- Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2022. Analogy generation by prompting large language models: A case study of instructgpt. *arXiv preprint arXiv:2210.04186*.

- Neha Reddy Bogireddy, Smriti Suresh, and Sunny Rai. 2023. I'm out of breath from laughing! i think? a dataset of covid-19 humor and its toxic variants. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1004–1013.
- Tom 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*, 33:1877–1901.
- Cohere For AI Team. 2024. A deepdive into aya expand: Advancing the frontier of multilinguality. <https://huggingface.co/blog/aya-expand>. Accessed: 2025-06-08.
- Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. *arXiv preprint arXiv:2211.15268*.
- John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermiş, Ahmet Üstün, and Sara Hooker. 2024. *Aya expand: Combining research breakthroughs for a new multilingual frontier*.
- Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. *arXiv preprint arXiv:1810.01375*.
- Vijay Gaikwad and Yashodhara Haribhakta. 2020. Adaptive glove and fasttext model for hindi word embeddings. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 175–179.
- 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 Esiohu, 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, Laurent 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 Kam-badur, 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, Shao-liang 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 Conguet, Virginie Do, Vish Vogeti, Vitor 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, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Del-pierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papanikos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boe-

senberg, 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, Asaf 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, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymmer, 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, Laverander 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, Xiaojuan 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*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Jennifer Hu and Michael C. Frank. 2024. [Auxiliary task demands mask the capabilities of smaller language models](#).

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai. 2023. Newsmet: A ‘do it all’ dataset of contemporary metaphors in news headlines. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlp suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2024. [Challenges in adapting multilingual llms to low-resource languages using lora peft tuning](#).

- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Large language models are zero-shot reasoners](#).
- Yinheng Li. 2023. [A practical survey on zero-shot prompt design for in-context learning](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. [Augmented language models: a survey](#).
- Gerhard Minnameier. 2010. *Abduction, Induction, and Analogy*, pages 107–119. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*.
- Leonard Richardson. 2024. [Beautiful soup documentation](#). Accessed: 2025-06-08.
- Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshov, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Letícia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kupala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv: 2408.00118*.
- Asahi Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*.
- Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can’t plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv: 2201.11903*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. Analogical—a new benchmark for analogy of long text for large language models. *arXiv preprint arXiv:2305.05050*.

Yufeng Zhang, Xuepeng Wang, Lingxiang Wu, and Jinqiao Wang. 2025. Enhancing chain of thought prompting in large language models via reasoning patterns. *arXiv preprint arXiv: 2404.14812*.

A Appendix

A.1 Model Specifications

The model specifications are provided below. We use the pre-trained models.

- Aya Expans 8B : We set the $max_new_tokens = 1200$, $torch_dtype = torch.float16$, $device_map = "auto"$, $do_sample = False$. The model was loaded using the HuggingFace API with the model name 'CohereForAI/aya-expans-8b'. The model runs in evaluation mode, which disables gradient updates for inference.
- Llama-3.1-8B-Instruct : We set the $max_new_tokens = 1200$, $torch_dtype = torch.float16$, $device_map = "auto"$, $do_sample = False$. The model was loaded using the HuggingFace API with the model name 'meta-Llama/Llama-3.1-8B-Instruct'. The model runs in evaluation mode, which disables gradient updates for inference.
- Gemma-2-9b-it : We set the $max_new_tokens = 1200$, $torch_dtype = torch.float16$, $device_map = "auto"$, $do_sample = False$. The model was loaded using the HuggingFace API with the model name 'google/gemma-2-9b-it'. The model runs in evaluation mode, which disables gradient updates for inference.

A.2 Tables

A.2.1 Prompts

Prompts for Analogy Tasks

Task B: 0-Shot Prompting (from Sec 3.4)

Models: Gemma-2-9B-it, Llama-3.1-8B-Instruct, Aya-Expans-8B

Hi-Hi Setting

System Prompt:

सादृश्य पूरा कीजिए :

आप अपना उत्तर इस प्रकार समाप्त करेंगे: ###अंतिम उत्तर: <आपके द्वारा चुना हुआ विकल्प>

User Prompt:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

En-En Setting

System Prompt:

Complete the analogy:

You will end your answer with: ###Final Answer: <Your chosen option>

User Prompt:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

Table A1: Prompts for Task B (0-Shot)

Task C: Chain of Thought Prompting (0-shot) (from Sec 3.5.1)

Models: Gemma-2-9B-it, Llama-3.1-8B-Instruct, Aya-Expans-8B

Hi-Hi Setting

System Prompt:

सादृश्य पूरा कीजिए :

आप अपना उत्तर इस प्रकार समाप्त करेंगे: ###अंतिम उत्तर: <आपके द्वारा चुना हुआ विकल्प>

User Prompt:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

आइये कदम दर कदम सोचें

En-En Setting**System Prompt:**

Complete the analogy:

You will end your answer with: ###Final Answer: <Your chosen option>

User Prompt:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

Let's think step by step.

Mixed Setting (En + Hi)**System Prompt:**

Complete the analogy:

You will end your answer with: ###Final Answer: <Your chosen option>

User Prompt:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

आइये कदम दर कदम सोचें

Table A2: Prompts for Task C (Chain of Thought 0– shot)

Task C: Grounded Zero-Shot Chain of Thought (from Sec 3.5.2)

Models: Gemma-2-9B-it, Llama-3.1-8B-Instruct, Aya-Expansive-8B

Hi-Hi Setting**System Prompt:**

आप एक समानता (एनालॉजी) संबंधी प्रश्न हल कर रहे हैं। समानता दो चीजों के बीच तुलना होती है, जो किसी न किसी तरह से एक-दूसरे से मेल खाती हैं।

आपका कार्य इस समानता को पूरा करना है, यानी पहले दो शब्दों के बीच के संबंध को समझकर उसी संबंध को तीसरे शब्द पर लागू करना और यह तय करना कि चौथा शब्द क्या होना चाहिए।

समानता हल करने के लिए इन चरणों का पालन करें: सबसे पहले, पहले दो शब्दों (A और B) के बीच के विशिष्ट संबंध को पहचानें। यह समझें कि A का B से क्या संबंध है।

फिर, उसी संबंध को तीसरे शब्द (C) पर लागू करें और देखें कि चौथा शब्द क्या होना चाहिए।

अंत में, दिए गए विकल्पों में से उस विकल्प का चयन करें जो आपके पहचाने गए संबंध के आधार पर समानता को सही तरीके से पूरा करता है।

प्रत्येक चरण में सावधानीपूर्वक सोचें और अंतिम निर्णय लेने से पहले कई संभावित संबंधों पर विचार करें। अपने तर्क को स्पष्ट रूप से प्रस्तुत करें।

अपने अंतिम उत्तर को इस प्रारूप में दें:

Final Answer: (X) विकल्प

अब निम्नलिखित समानता को इस तीन-चरणीय दृष्टिकोण से हल करें:

User Prompt:

सादृश्यता पूरी करें:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

पहले बताई गई तीन-चरणीय विधि का पालन करके इस सादृश्य को हल करें।

En-En Setting

System Prompt:

You are solving an analogy problem. An analogy is a comparison between two things that are similar in some way. Your task is to complete the analogy by finding the relationship between the first two terms and applying that same relationship to find what the third term relates to. Follow these steps to solve the analogy:

1. First, identify the specific relationship between the first two terms (A and B). Think about how A relates to B.
2. Next, apply this same relationship to the third term (C) to determine what the fourth term should be.
3. Finally, examine each of the given options and select the one that best completes the analogy based on the relationship you identified.

For each step, think carefully and consider multiple possible relationships before deciding. Be explicit in your reasoning. Present your final answer in the format: ###Final Answer: (X) option_text

Now solve the following analogy using this three-step approach:

User Prompt:

Complete the following analogy:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

by following the three-step method.

Mixed Setting (En + Hi)

System Prompt:

You are solving an analogy problem. An analogy is a comparison between two things that are similar in some way. Your task is to complete the analogy by finding the relationship between the first two terms and applying that same relationship to find what the third term relates to. Follow these steps to solve the analogy:

1. First, identify the specific relationship between the first two terms (A and B). Think about how A relates to B.
2. Next, apply this same relationship to the third term (C) to determine what the fourth term should be.
3. Finally, examine each of the given options and select the one that best completes the analogy based on the relationship you identified.

For each step, think carefully and consider multiple possible relationships before deciding. Be explicit in your reasoning. Present your final answer in the format: ###Final Answer: (X) option_text

Now solve the following analogy using this three-step approach:

User Prompt:

Complete the analogy:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

by following the three-step method

Table A3: Prompts for Task C (Grounded 0–Shot Chain of Thought)

Task C: Grounded Few Shot Chain of Thought (from Sec 3.5.3)

Models: Gemma-2-9B-it, Llama-3.1-8B-Instruct, Aya-Expansive-8B

Hi-Hi Setting**System Prompt:**

आप एक समानता (एनालॉजी) संबंधी प्रश्न हल कर रहे हैं। समानता दो चीजों के बीच तुलना होती है, जो किसी न किसी तरह से एक-दूसरे से मेल खाती हैं।

आपका कार्य इस समानता को पूरा करना है, यानी पहले दो शब्दों के बीच के संबंध को समझकर उसी संबंध को तीसरे शब्द पर लागू करना और यह तय करना कि चौथा शब्द क्या होना चाहिए। समानता हल करने के लिए इन चरणों का पालन करें:

1. सबसे पहले, पहले दो शब्दों (A और B) के बीच के विशिष्ट संबंध को पहचानें। यह समझें कि □ का □ से क्या संबंध है।
2. फिर, उसी संबंध को तीसरे शब्द (C) पर लागू करें और देखें कि चौथा शब्द क्या होना चाहिए।
3. अंत में, दिए गए विकल्पों में से उस विकल्प का चयन करें जो आपके पहचाने गए संबंध के आधार पर समानता को सही तरीके से पूरा करता है।

प्रत्येक चरण में सावधानीपूर्वक सोचें और अंतिम निर्णय लेने से पहले कई संभावित संबंधों पर विचार करें। अपने तर्क को स्पष्ट रूप से प्रस्तुत करें। अपने अंतिम उत्तर को इस प्रारूप में दें:

अंतिम उत्तर: (X) विकल्प यहां कुछ उदाहरण दिए गए हैं: उदाहरण 1:

निम्नलिखित समानता को पूरा करें: गंगा : नदी :: हिमालय : ?

(A) पर्वत

(B) देश

(C) महासागर

(D) मैदान

चरण 1: सबसे पहले, मुझे "गंगा" और "नदी" के बीच विशिष्ट संबंध की पहचान करनी है। गंगा एक विशिष्ट नदी है, और "नदी" इसका वर्ग या श्रेणी है। यह एक ऐसा संबंध है जहां पहला शब्द दूसरे शब्द का एक विशिष्ट उदाहरण है।

चरण 2: अब, मुझे इसी संबंध को "हिमालय" पर लागू करना है। यदि हिमालय गंगा की तरह एक विशिष्ट उदाहरण है, तो मुझे इसका वर्ग या श्रेणी ढूंढनी होगी।

चरण 3: अंत में, मुझे प्रत्येक विकल्प की जांच करनी है:

–पर्वत: हिमालय एक विशिष्ट पर्वत श्रृंखला है, और "पर्वत" इसका वर्ग "पर्वत" है। यह गंगा और नदी के समान पैटर्न का अनुसरण करता है।

–देश: हिमालय कोई देश नहीं है; यह एक भौगोलिक विशेषता है

–महासागर: हिमालय का जल निकायों जैसे महासागरों से कोई संबंध नहीं है।

–मैदान: हिमालय मैदान के विपरीत है; यह एक उच्च भूमि है।

अंतिम उत्तर: (A) पर्वत

उदाहरण 2:

निम्नलिखित समानता को पूरा करें: चावल : खेती :: लोहा : ?

- (A) धातु
- (B) खनन
- (C) निर्माण
- (D) व्यापार

चरण 1: सबसे पहले, मुझे "चावल" और "खेती" के बीच विशिष्ट संबंध की पहचान करनी है। चावल एक कृषि उत्पाद है जो खेती की प्रक्रिया से प्राप्त होता है। यह एक उत्पाद और उसे प्राप्त करने की प्रक्रिया के बीच का संबंध है।

चरण 2: अब, मुझे इसी संबंध को "लोहा" पर लागू करना है। यदि लोहा चावल की तरह एक उत्पाद है, तो मुझे लोहा प्राप्त करने की प्रक्रिया ढूंढनी होगी।

चरण 3: अंत में, मुझे प्रत्येक विकल्प की जांच करनी है:

- धातु: यह बताता है कि लोहा क्या है (एक धातु), न कि इसे कैसे प्राप्त किया जाता है।
- खनन: खनन वह प्रक्रिया है जिसके द्वारा लोहा पृथ्वी से प्राप्त किया जाता है, जैसे कि खेती वह प्रक्रिया है जिससे चावल प्राप्त होता है। यह वही संबंध बनाए रखता है।
- निर्माण: यह एक ऐसी प्रक्रिया है जो लोहे का उपयोग करती है, न कि लोहा कैसे प्राप्त किया जाता है।
- व्यापार: यह लोहे के वितरण से संबंधित है, न कि इसके उत्पादन से।

अंतिम उत्तर: (B) खनन

उदाहरण 3:

निम्नलिखित समानता को पूरा करें: दिल्ली : भारत :: टोक्यो : ?

- (A) चीन
- (B) रूस
- (C) जापान
- (D) कोरिया

चरण 1: सबसे पहले, मुझे "दिल्ली" और "भारत" के बीच विशिष्ट संबंध की पहचान करनी है। दिल्ली भारत की राजधानी है। यह एक राजधानी और उसके देश के बीच का संबंध है। चरण 2: अब, मुझे इसी संबंध को "टोक्यो" पर लागू करना है। यदि टोक्यो दिल्ली की तरह एक राजधानी है, तो मुझे वह देश ढूंढना होगा जिसका टोक्यो राजधानी है। चरण 3: अंत में, मुझे प्रत्येक विकल्प की जांच करनी है:

- चीन: चीन की राजधानी बीजिंग है, टोक्यो नहीं।
- रूस: रूस की राजधानी मॉस्को है, टोक्यो नहीं।
- जापान: टोक्यो जापान की राजधानी है। यह दिल्ली और भारत के समान संबंध बनाए रखता है।
- कोरिया: कोरिया (उत्तर या दक्षिण) की राजधानी प्योंगयांग या सियोल हैं, टोक्यो नहीं।

अंतिम उत्तर: (C) जापान

उदाहरण 4:

निम्नलिखित समानता को पूरा करें: पेंसिल : लिखना :: कैची : ?

- (A) पेपर
- (B) काटना
- (C) बनाना
- (D) सीना

चरण 1: सबसे पहले, मुझे "पेंसिल" और "लिखना" के बीच विशिष्ट संबंध की पहचान करनी है। पेंसिल एक उपकरण है जिसका उपयोग लिखने की क्रिया के लिए किया जाता है। यह एक उपकरण और उसके मुख्य कार्य के बीच का संबंध है।

चरण 2: अब, मुझे इसी संबंध को "कैंची" पर लागू करना है। यदि कैंची पेंसिल की तरह एक उपकरण है, तो मुझे कैंची के मुख्य कार्य को ढूंढना होगा।

चरण 3: अंत में, मुझे प्रत्येक विकल्प की जांच करनी है:

- पेपर: पेपर एक वस्तु है जिस पर काम किया जाता है, न कि एक क्रिया।
- काटना: काटना वह मुख्य क्रिया है जिसके लिए कैंची का उपयोग किया जाता है, जैसे कि लिखना पेंसिल का मुख्य कार्य है।
- बनाना: बनाना कैंची का मुख्य कार्य नहीं है।
- सीना: सीना के लिए आमतौर पर सुई और धागे का उपयोग किया जाता है, न कि कैंची का।

अंतिम उत्तर: (B) काटना

उदाहरण 5:

निम्नलिखित समानता को पूरा करें: शेर : जंगल :: मछली : ?

- (A) पिंजरा
- (B) समुद्र
- (C) रेगिस्तान
- (D) खेत

चरण 1: सबसे पहले, मुझे "शेर" और "जंगल" के बीच विशिष्ट संबंध की पहचान करनी है। जंगल वह प्राकृतिक आवास या वातावरण है जहां शेर रहते हैं। यह एक जानवर और उसके प्राकृतिक निवास स्थान के बीच का संबंध है।

चरण 2: अब, मुझे इसी संबंध को "मछली" पर लागू करना है। यदि मछली शेर की तरह एक जानवर है, तो मुझे मछली के प्राकृतिक निवास स्थान या वातावरण को ढूंढना होगा।

चरण 3: अंत में, मुझे प्रत्येक विकल्प की जांच करनी है:

- पिंजरा: पिंजरा एक कृत्रिम वातावरण है जहां जानवरों को रखा जाता है, यह मछली का प्राकृतिक आवास नहीं है।
- समुद्र: समुद्र वह प्राकृतिक जलीय वातावरण है जहां अधिकांश मछलियां रहती हैं, जैसे कि जंगल शेरों का प्राकृतिक आवास है।
- रेगिस्तान: रेगिस्तान एक शुष्क वातावरण है जो मछलियों के लिए उपयुक्त नहीं है।
- खेत: खेत कृषि भूमि है और मछलियों का प्राकृतिक आवास नहीं है।

अंतिम उत्तर: (B) समुद्र

अब निम्नलिखित समानता को इस तीन-चरणीय दृष्टिकोण से हल करें:

User Prompt:

सादृश्यता पूरी करें:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश

En-En Setting

System Prompt: You are solving an analogy problem. An analogy is a comparison between two things that are similar in some way. Your task is to complete the analogy by finding the relationship between the first two terms and applying that same relationship to find what the third term relates to.

Follow these steps to solve the analogy:

1. First, identify the specific relationship between the first two terms (A and B). Think about how A relates to B.
2. Next, apply this same relationship to the third term (C) to determine what the fourth term should be.
3. Finally, examine each of the given options and select the one that best completes the analogy based on the relationship you identified.

For each step, think carefully and consider multiple possible relationships before deciding. Be explicit in your reasoning.

Present your final answer in the format: ###Final Answer: (X) option_text

Here are some examples:

Example 1: Complete the analogy: गंगा : नदी :: हिमालय : ?

- (A) पर्वत
- (B) देश
- (C) महासागर
- (D) मैदान

Step 1: First, I need to identify the specific relationship between "गंगा" (Ganga) and "नदी" (river). Ganga is a specific river, and "नदी" is its category or classification. This is a relationship where the first term is a specific example of the second term.

Step 2: Next, I need to apply this same relationship to "हिमालय" (Himalaya). If Himalaya is a specific example like Ganga, then I need to find its category or classification.

Step 3: Finally, let me examine each option:

- पर्वत (mountain): Himalaya is a specific mountain range, and "पर्वत" is the category "mountain." This follows the same pattern as Ganga and river.

- देश (country): Himalaya is not a country; it's a geographical feature.

- महासागर (ocean): Himalaya is not related to water bodies like oceans.

- मैदान (plain): Himalaya is the opposite of a plain; it's an elevated landform.

###Final Answer: (A)

Example 2:

Complete the analogy: चावल : खेती :: लोहा : ?

- (A) धातु
- (B) खनन
- (C) निर्माण
- (D) व्यापार

Step 1: First, I need to identify the specific relationship between "चावल" (rice) and "खेती" (farming). Rice is an agricultural product that is obtained through the process of farming. This is a relationship between a product and the process used to obtain it.

Step 2: Next, I need to apply this same relationship to "लोहा" (iron). If iron is a product like rice, then I need to find the process used to obtain iron.

Step 3: Finally, let me examine each option:

-धातु (metal): This describes what iron is (a metal), not how it's obtained.

-खनन (mining): Mining is the process by which iron is obtained from the earth, just -as farming is how rice is obtained. This maintains the same relationship.

-निर्माण (construction): This is a process that uses iron, not how iron is obtained.

–व्यापार (trade): This relates to distribution of iron, not its production.

####Final Answer: (B) खनन

Example 3: Complete the analogy: दिल्ली : भारत :: टोक्यो : ?

- (A) चीन
- (B) रूस
- (C) जापान
- (D) कोरिया

Step 1: First, I need to identify the specific relationship between "दिल्ली" (Delhi) and "भारत" (India). Delhi is the capital city of India. This is a relationship between a capital city and its country.

Step 2: Next, I need to apply this same relationship to "टोक्यो" (Tokyo). If Tokyo is a capital city like Delhi, then I need to find the country of which Tokyo is the capital.

Step 3: Finally, let me examine each option:

–चीन (China): The capital of China is Beijing, not Tokyo.

–रूस (Russia): The capital of Russia is Moscow, not Tokyo.

–जापान (Japan): Tokyo is the capital of Japan. This maintains the same relationship as Delhi and India.

–कोरिया (Korea): The capitals of Korea (North or South) are Pyongyang or Seoul, not Tokyo.

####Final Answer: (C)

Example 4: Complete the analogy: पेंसिल : लिखना :: कैंची : ?

- (A) पेपर
- (B) काटना
- (C) बनाना
- (D) सीना

Step 1: First, I need to identify the specific relationship between "पेंसिल" (pencil) and "लिखना" (writing). A pencil is a tool used for the action of writing. This is a relationship between a tool and its primary function.

Step 2: Next, I need to apply this same relationship to "कैंची" (scissors). If scissors is a tool like a pencil, then I need to find its primary function.

Step 3: Finally, let me examine each option:

- पेपर (paper): Paper is an object that is worked upon, not an action.

- काटना (cutting): Cutting is the primary action for which scissors are used, just as writing is the primary function of a pencil.

- बनाना (making/creating): Making is not the primary function of scissors.

- सीना (sewing): Sewing typically uses needle and thread, not scissors.

####Final Answer: (B) काटना

Example 5:

Complete the analogy: शेर : जंगल :: मछली : ?

- (A) पिंजरा
- (B) समुद्र
- (C) रेगिस्तान
- (D) खेत

Step 1: First, I need to identify the specific relationship between "शेर" (lion) and "जंगल" (jungle/forest). A jungle is the natural habitat or environment where lions live. This is a relationship between an animal and its natural habitat.

Step 2: Next, I need to apply this same relationship to "मछली" (fish). If fish is an animal like a lion, then I need to find its natural habitat.

Step 3: Finally, let me examine each option:

- पिंजरा (cage): A cage is an artificial environment where animals are kept, not the natural habitat for fish.

- समुद्र (sea/ocean): The sea/ocean is the natural aquatic environment where most fish live, just as the jungle is the natural habitat for lions.

- रेगिस्तान (desert): A desert is a dry environment that is not suitable for fish.

- खेत (field/farm): A field is agricultural land and not the natural habitat for fish.

####Final Answer: (B) समुद्र

Now solve the following analogy using this three-step approach:

User Prompt:

Complete the following analogy:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश.

Mixed Setting (En + Hi)

System Prompt: You are solving an analogy problem. An analogy is a comparison between two things that are similar in some way. Your task is to complete the analogy by finding the relationship between the first two terms and applying that same relationship to find what the third term relates to.

Follow these steps to solve the analogy:

1. First, identify the specific relationship between the first two terms (A and B). Think about how A relates to B.

2. Next, apply this same relationship to the third term (C) to determine what the fourth term should be.

3. Finally, examine each of the given options and select the one that best completes the analogy based on the relationship you identified.

For each step, think carefully and consider multiple possible relationships before deciding. Be explicit in your reasoning.

Present your final answer in the format: ####Final Answer: (X) option_text

Here are some examples:

Example 1: Complete the analogy: गंगा : नदी :: हिमालय : ?

(A) पर्वत

(B) देश

(C) महासागर

(D) मैदान

Step 1: First, I need to identify the specific relationship between "गंगा" (Ganga) and "नदी" (river). Ganga is a specific river, and "नदी" is its category or classification. This is a relationship where the first term is a specific example of the second term.

Step 2: Next, I need to apply this same relationship to "हिमालय" (Himalaya). If Himalaya is a specific example like Ganga, then I need to find its category or classification.

Step 3: Finally, let me examine each option:

- पर्वत (mountain): Himalaya is a specific mountain range, and "पर्वत" is the category "mountain." This follows the same pattern as Ganga and river.
- देश (country): Himalaya is not a country; it's a geographical feature.
- महासागर (ocean): Himalaya is not related to water bodies like oceans.
- मैदान (plain): Himalaya is the opposite of a plain; it's an elevated landform.

###Final Answer: (A)

Example 2:

Complete the analogy: चावल : खेती :: लोहा : ?

- (A) धातु
- (B) खनन
- (C) निर्माण
- (D) व्यापार

Step 1: First, I need to identify the specific relationship between "चावल" (rice) and "खेती" (farming). Rice is an agricultural product that is obtained through the process of farming. This is a relationship between a product and the process used to obtain it.

Step 2: Next, I need to apply this same relationship to "लोहा" (iron). If iron is a product like rice, then I need to find the process used to obtain iron.

Step 3: Finally, let me examine each option:

- धातु (metal): This describes what iron is (a metal), not how it's obtained.
- खनन (mining): Mining is the process by which iron is obtained from the earth, just -as farming is how rice is obtained. This maintains the same relationship.
- निर्माण (construction): This is a process that uses iron, not how iron is obtained.
- व्यापार (trade): This relates to distribution of iron, not its production.

###Final Answer: (B) खनन

Example 3: Complete the analogy: दिल्ली : भारत :: टोक्यो : ?

- (A) चीन
- (B) रूस
- (C) जापान
- (D) कोरिया

Step 1: First, I need to identify the specific relationship between "दिल्ली" (Delhi) and "भारत" (India). Delhi is the capital city of India. This is a relationship between a capital city and its country.

Step 2: Next, I need to apply this same relationship to "टोक्यो" (Tokyo). If Tokyo is a capital city like Delhi, then I need to find the country of which Tokyo is the capital.

Step 3: Finally, let me examine each option:

- चीन (China): The capital of China is Beijing, not Tokyo.
- रूस (Russia): The capital of Russia is Moscow, not Tokyo.
- जापान (Japan): Tokyo is the capital of Japan. This maintains the same relationship as Delhi and India.
- कोरिया (Korea): The capitals of Korea (North or South) are Pyongyang or Seoul, not Tokyo.

###Final Answer: (C)

Example 4:

Complete the analogy: पेंसिल : लिखना :: कैंची : ?

- (A) पेपर
- (B) काटना
- (C) बनाना
- (D) सीना

Step 1: First, I need to identify the specific relationship between "पेंसिल" (pencil) and "लिखना" (writing). A pencil is a tool used for the action of writing. This is a relationship between a tool and its primary function.

Step 2: Next, I need to apply this same relationship to "कैंची" (scissors). If scissors is a tool like a pencil, then I need to find its primary function.

Step 3: Finally, let me examine each option:

- पेपर (paper): Paper is an object that is worked upon, not an action.
- काटना (cutting): Cutting is the primary action for which scissors are used, just as writing is the primary function of a pencil.
- बनाना (making/creating): Making is not the primary function of scissors.
- सीना (sewing): Sewing typically uses needle and thread, not scissors.

###Final Answer: (B) काटना

Example 5:

Complete the analogy: शेर : जंगल :: मछली : ?

- (A) पिंजरा
- (B) समुद्र
- (C) रेगिस्तान
- (D) खेत

Step 1: First, I need to identify the specific relationship between "शेर" (lion) and "जंगल" (jungle/forest). A jungle is the natural habitat or environment where lions live. This is a relationship between an animal and its natural habitat.

Step 2: Next, I need to apply this same relationship to "मछली" (fish). If fish is an animal like a lion, then I need to find its natural habitat.

Step 3: Finally, let me examine each option:

- पिंजरा (cage): A cage is an artificial environment where animals are kept, not the natural habitat for fish.
- समुद्र (sea/ocean): The sea/ocean is the natural aquatic environment where most fish live, just as the jungle is the natural habitat for lions.
- रेगिस्तान (desert): A desert is a dry environment that is not suitable for fish.
- खेत (field/farm): A field is agricultural land and not the natural habitat for fish.

###Final Answer: (B) समुद्र

Now solve the following analogy using this three-step approach:

User Prompt:

सादृश्यता पूरी करें:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

- (A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश
-

Table A4: Prompts for Task C (Grounded few-Shot Chain of Thought)

Task C: Few Shot Chain of Thought (with Translation) from Sec 3.5.4

Models: Gemma-2-9B-it, Llama-3.1-8B-Instruct, Aya-Expansive-8B

English-only Setting

System Prompt: You are solving analogy problems presented in Hindi. An analogy is a comparison between two things that are similar in some way.

Follow these main steps: 1. Translation: Translate the Hindi question and all options to English. 2. Solution: Solve the translated (english) analogy using only English (detailed below). 3. Mapping: Map your English answer back to the correct Hindi option.

For the solution process (step 2), follow these sub-steps: a) Identify the specific relationship between the first two terms (A and B).

b) Apply this same relationship to the third term (C) to determine what the fourth term should be.

c) Examine each of the given options and select the one that best completes the analogy.

IMPORTANT: Use ONLY English words during your solution process (step 2 and its sub-steps).

Only use Hindi when referring to the original question and when giving your final answer.

For each step, think carefully and consider multiple possible relationships. Be explicit in your reasoning.

Present your final answer in the format: ###Final Answer: (X) option_text

Here are some examples:

Example 1:

Complete the analogy: गंगा : नदी :: हिमालय : ?

(A) पर्वत

(B) देश

(C) महासागर

(D) मैदान

Step 1 - Translation:

Question: Ganga : River :: Himalaya : ?

Options:

(A) Mountain

(B) Country

(C) Ocean

(D) Plain

Step 2 - Solution:

a) Relationship identification:

Ganga is a specific river, and River is its category. This is a specific instance to category relationship.

b) Relationship application:

Now I need to apply this relationship to Himalaya. If Ganga is a specific river, then Himalaya would be a specific instance of what category?

c) Option examination:

- Mountain: Himalaya is a specific mountain range, so Mountain is its category. This matches the relationship.
- Country: Himalaya is not a country, it's a geographical feature.
- Ocean: Himalaya is not a body of water, it's a land formation.
- Plain: Himalaya is the opposite of a plain; it's an elevated region.

Step 3 - Mapping:

The answer in English is "Mountain" which corresponds to the Hindi option (A) पर्वत.

###Final Answer: (A)

Example 2:

Complete the analogy: दिल्ली : भारत :: टोक्यो : ?

- (A) चीन
- (B) रूस
- (C) जापान
- (D) कोरिया

Step 1 - Translation:

Question: Delhi : India :: Tokyo : ?

Options:

- (A) China
- (B) Russia
- (C) Japan
- (D) Korea

Step 2 - Solution:

a) Relationship identification:

Delhi is the capital city of India. This is a capital-country relationship.

b) Relationship application:

Now I need to apply this relationship to Tokyo. I'm looking for the country of which Tokyo is the capital.

c) Option examination:

- China: The capital of China is Beijing, not Tokyo.
- Russia: The capital of Russia is Moscow, not Tokyo.
- Japan: Tokyo is the capital of Japan. This matches the relationship.
- Korea: The capitals of North and South Korea are Pyongyang and Seoul respectively, not Tokyo.

Step 3 - Mapping:

The answer in English is "Japan" which corresponds to the Hindi option (C) जापान.

###Final Answer: (C)

Example 3:

Complete the analogy: चावल : खेती :: लोहा : ?

- (A) धातु
- (B) खनन

(C) निर्माण

(D) व्यापार

Step 1 - Translation:

Question: Rice : Farming :: Iron : ?

Options:

(A) Metal

(B) Mining

(C) Construction

(D) Trade

Step 2 - Solution:

a) Relationship identification:

Farming is the process by which Rice is produced or obtained. This is a product-production process relationship.

b) Relationship application:

Now I need to apply this relationship to Iron. I'm looking for the process by which Iron is produced or obtained.

c) Option examination:

- Metal: This is a category to which Iron belongs, not a production process.

- Mining: This is the process by which Iron is obtained from the earth, similar to how Farming is used to obtain Rice. This matches the relationship.

- Construction: This is a process that uses Iron, not how it's produced.

- Trade: This relates to the distribution of Iron, not its production.

Step 3 - Mapping:

The answer in English is "Mining" which corresponds to the Hindi option (B) खनन.

###Final Answer: (B)

Example 4:

Complete the analogy: पेंसिल : लिखना :: कैंची : ?

(A) पेपर

(B) काटना

(C) बनाना

(D) सीना

Step 1 - Translation:

Question: Pencil : Writing :: Scissors : ?

Options:

(A) Paper

(B) Cutting

(C) Making

(D) Sewing

Step 2 - Solution:

a) Relationship identification:

A pencil is a tool used for the action of writing. This is a tool-function relationship.

b) Relationship application:

Now I need to apply this relationship to scissors. I'm looking for the primary function of scissors.

c) Option examination:

- Paper: This is an object that is worked upon, not an action.

- Cutting: This is the primary function of scissors, just as writing is the primary function of a pencil.

- Making: This is too general and not the specific function of scissors.

- Sewing: Sewing is done with a needle and thread, not scissors.

Step 3 - Mapping:

The answer in English is "Cutting" which corresponds to the Hindi option (B) काटना.

###Final Answer: (B)

Example 5:

Complete the analogy: शेर : जंगल :: मछली : ?

(A) पिंजरा

(B) समुद्र

(C) रेगिस्तान

(D) खेत

Step 1 - Translation:

Question: Lion : Jungle :: Fish : ?

Options:

(A) Cage

(B) Ocean/Sea

(C) Desert

(D) Field/Farm

Step 2 - Solution:

a) Relationship identification:

A jungle is the natural habitat where lions typically live. This is an animal-habitat relationship.

b) Relationship application:

Now I need to apply this relationship to fish. I'm looking for the natural habitat where fish typically live.

c) Option examination:

- Cage: This is an artificial structure, not a natural habitat.

- Ocean/Sea: This is the natural aquatic environment for most fish, like jungle is for lions.

- Desert: Deserts are dry and unsuitable for fish.

- Field/Farm: This is land used for agriculture, not suitable for fish.

Step 3 - Mapping:

The answer in English is "Ocean/Sea" which corresponds to the Hindi option (B) समुद्र.

###Final Answer: (B)

Now solve the following analogy using the same step-by-step approach. Remember to use ONLY English in your solution process (step 2):

User Prompt:

Complete the following analogy:

भोपाल : मध्य प्रदेश :: भुवनेश्वर : ?

(A) गुजरात (B) उड़ीसा (C) राजस्थान (D) अरुणाचल प्रदेश.

Table A5: Prompts for Task C (Few Shot Chain of Thought (with translation))

A.2.2 Model Response Language across different settings

Model	Setting (System+User)	0-Shot	0-Shot CoT	Grounded 0-Shot CoT	CoT (Few Shot)	CoT (Few Shot-Translate-EN)
aya-expanse-8B	Hi+Hi	Hi	Hi	Hi	Hi	-
	Hi+En	-	Hi	Hi	Hi	-
	En+En	En	En	En	En	En
Llama-3.1-8B-instruct	Hi+Hi	Hi	Hi	Hi	Hi	-
	Hi+En	-	Hi	Hi	Hi	-
	En+En	Hi	Hi	En	En	Hi
gemma-2-9b-it	Hi+Hi	Hi	Hi	Hi	Hi	-
	Hi+En	-	En	En	En	-
	En+En	En	En	En	En	En

Table A6: Language in which each model responded across different prompting strategies and language settings

Simulating Emotional Intelligence in LLMs through Behavioral Conditioning and Analogical Retrieval

G. Sai Linisha Reddy Mounil Hiren Kankhara Mridul Maheshwari Swayam Bansal
Rishit Kapoor Himesh Reddy M Bagesh Kumar
Manipal University Jaipur, India

g.229302590@uj.manipal.edu, mounil.23fe10ite00021@uj.manipal.edu,
mridul.23fe10ite00041@uj.manipal.edu, swayam.23fe10ite00308@uj.manipal.edu,
rishit.23fe10ite00026@uj.manipal.edu, himesh.229303296@uj.manipal.edu,
pse2016001@iiita.ac.in

Abstract

Human emotional expression emerges from a complex interplay of verbal, para-verbal, and non-verbal cues. This paper presents a dual-path framework for emotionally grounded text generation in large language models by integrating behavioral metadata with analogical retrieval. We introduce the MECC (Multimodal Emotionally Conditioned Corpus), a dataset of 1,764 question-answer pairs collected via structured interviews and annotated across 15 emotion categories with tone, response time, and body language. A LLaMA-3.1-8B-Instruct model is fine-tuned on MECC using behavior-encoded prompts, and inference is supported by a metadata-filtered Retrieval-Augmented Generation (RAG) pipeline. Detailed emotion-level analysis reveals trade-offs between emotional fidelity and semantic diversity, emphasizing the need for nuanced evaluation. This study contributes a richly annotated multimodal emotion corpus, a metadata-driven RAG architecture, a well-structured framework for building emotionally aware language models. Our code is available at <https://github.com/MetaResearcher/Framework>

1 Introduction

“Emotion is not opposed to reason; it is its foundation.” — Antonio Damasio, Descartes’ Error

Human emotional reasoning is seldom simply associative or reactive. Rather, it is inherently analogical, shaped by past experiences, moderated by the present context, and shaped by memory, rich in metaphors for how they make decisions now. In emotionally charged circumstances, people often reflect analogically: “This is like how I felt when...” Indeed, this notion of analogical reasoning accounts for foundational cognitive models of analogy (Gentner, 1983)(Gärdenfors, 2000)(Holyoak and Thagard, 1995) that have been used to articulate concepts of affective computing and socially

intelligent systems (Picard, 1997)(Hoegen et al., 2019)

Nonetheless, even with tremendous advancements in artificial intelligence and current generative models, the state-of-the-art is underwhelmed in its ability to replicate this level of nuance with emotional reasoning. Given that most Large Language Models (LLMs) leverage only surface-level conditioning, using emotion tokens (Zhou et al., 2018) sentiment prompts, or affective keywords (Rashkin et al., 2019), at the surface without translating the embodied, context-sensitive nature of a human emotional experience. Further, as with many machines initially designed for logic or limited predictability, LLMs miss integral non-verbal elements involved with human emotional messaging: tone, timing, gesture, etc. (Mehrabian, 1971)(Wang et al., 2004).

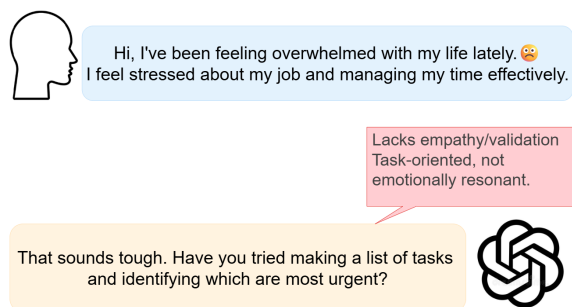


Figure 1: Figure of ChatGPT-4o response where it fails to generate an emotionally aware response

The degree of disconnection noted above is important in human-centered areas of work like healthcare, education, and therapy, in which emotionally coherent and capturing the unique personal connection is important. Recent applications using generative systems to enhance empathy in clinical documentation (Nag et al., 2023) and encourage pro-social engagement in environmental psychology (Lim et al., 2024), demonstrate that more complex cognitive-emotional capacities, such as

introspection, empathy, and analogy, are lagging (Ortega-Ochoa et al., 2024)(Varma et al., 2024).

Building on these research gaps, we present the **MECC** (Multi-modal Emotionally Conditioned Corpus), a behaviorally enriched data set constructed from a 60-item psychological interview protocol. The data set includes 1,764 pairs (question, response) from 31 participants, annotated across 15 emotional categories and accompanied by behavioral metadata such as vocal tone, response time, and body language. By incorporating both verbal and paraverbal dimensions, MECC enables a more cognitively grounded investigation of affective states.

Using MECC, we fine-tuned the LLaMA-3.1-8B-Instruct model on behavior-encoded prompts and integrated a metadata-filtered Retrieval Augmented Generation (RAG) pipeline. This dual-path architecture conditions the generation on both situational context and analogically retrieved affective exemplars. The combined system goes beyond surface-level fluency by generating emotionally aligned responses that reflect behavioral and situational grounding, supporting more context-sensitive and affect-aware language generation.

2 Related Work

Recent developments in large language models (LLMs) have had a great impact on emotion understanding tasks generally, and multimodal emotion understanding tasks specifically. (Luo et al., 2024) experimented with several LLMs for a specific variant of the Multimodal Emotion Cause Pair Extraction with Emotion Category (MECPE-Cat) task. After noting that ChatGLM had the highest performance, they reported a weighted average F1 score of 34.71 percent using just two training epochs. Prompt engineering was an important aspect of this team’s system, and they got their emotion-labeled data from the ECF dataset, which contained a total of 13,619 utterances, in an overwhelmingly constructive way with the data through the use of prompt engineering.

(Wang et al., 2024) expanded on Luo et al.’s work by running team submissions for two sub-tasks: text-based emotion-cause pair extraction (TECPE) and multimodal emotion-cause pair extraction (MECPE). For the task of TECPE, the highest F1 score reported was 0.3223 with a combination of LLaMA2 and SpanBERT. For the sub-

task of MECPE, the highest F1 score reported was 0.3774 with an ensemble of LLaMA2, RoBERTa, and LLaMA, showing that multimodal models tend to outshine unimodal models.

Nevertheless, standard unimodal methods may not offer the complexity necessary to successfully address authentic emotional expression. (Cheng et al., 2024) saw the importance of this and created the MERR dataset, which included 28,618 coarse-grained and 4,487 fine-grained annotated examples. Their Emotion-LLaMA model, through instruction tuning, achieved better performance than existing multimodal LLMs (MLLMs) across many benchmarks. They achieved the highest scores in Clue Overlap (7.83) and Label Overlap (6.25) on the EMER dataset, and an F1 score of .9036 on the MER2023-SEMI challenge, along with good zero-shot results on the DFEW dataset.

Another study of emotional intelligence in LLMs was conducted by (Chen et al., 2024) with EmotionQueen, a standardized benchmark containing 10,000 statements generated by GPT-4 across five life scripts. They examined 11 major LLMs on four tasks: Event Recognition, Mixed Event Recognition, Implicit Emotion Recognition, and Intention Recognition. The study found that LLaMA-70B obtained the highest average score (93.4), while Claude2 was ranked first in Key Event Recognition. While pre-trained attention-based LLMs such as GPT-4 can achieve high accuracy on certain emotion recognition tasks, they often fail to translate this accuracy into responses that demonstrate compassionate or emotionally supportive reasoning. That is, the models may correctly identify emotional cues but lack the ability to respond in a way that reflects genuine empathy or care.

In addition to these studies on recognition, (Varma et al., 2024) introduced an emotionally adaptive AI pipeline to generate personalized emotional responses. They compared a fine-tuned LLaMA-3 8B modulated by LoRA with an RAG system. The fine-tuned model trained using interview data containing six universal emotions outperformed the RAG system; all of their correlation indices were greater than 0.950 (i.e., 0.850 for Anger) with a lower MSE score overall, such as 0.0452. Their research supports the notion that RAG systems struggle with emotional nuance. This motivates our dual-path design that includes analogical retrieval and behavioral conditioning.

Overall, the recent advances in the literature mark an important shift towards multimodal LLMs

No.	Paper / Authors	Model/Method	Contribution
1	Luo et al. (2024)	ChatGLM + Prompt Engineering	Achieved 34.71% F1 on MECPE-Cat using ECF dataset; demonstrated effective prompt design.
2	Wang et al. (2024)	LLaMA2 + SpanBERT; Ensemble (LLaMA2, RoBERTa, LLaMA)	Reported highest F1 of 0.3223 (TECPE) and 0.3774 (MECPE); showed multimodal models outperform unimodal.
3	Cheng et al. (2024)	Emotion-LLaMA (Instruction Tuning)	Introduced MERR dataset; achieved state-of-the-art on multiple emotion benchmarks.
4	Chen et al. (2024)	11 LLMs (incl. LLaMA-70B, Claude2, GPT-4)	Proposed EmotionQueen benchmark; analyzed LLMs on emotional intelligence tasks.
5	Varma et al. (2024)	Fine-tuned LLaMA-3 8B (LoRA) vs. RAG	Developed emotionally adaptive pipeline; fine-tuned model outperformed RAG in nuance and accuracy.

Table 1: Summary of recent literature on LLMs for emotion understanding.

and instruction tuning, aiming to capture nuanced emotional understanding. However, there remains a notable gap in models that can accurately identify emotion cause relationships while simultaneously adapting empathetically to behavioral context and modality. This study addresses that gap through a behavior-conditioned framework grounded in analogical retrieval and affective reasoning.

3 MECC Dataset

3.1 Motivation and Theoretical Framing

Existing emotion classification corpora such as *EmpatheticDialogues* (Rashkin et al., 2019) and *GoEmotions* (Demszky et al., 2020) have laid the foundational work in emotion classification, but focus exclusively on text, overlooking prosodic and embodied dimensions central to emotional communication such as tone, timing, and gesture. This unimodal framing limits progress towards cognitively grounded emotional reasoning.

Drawing from analogical reasoning (Gentner, 1983) and affective conceptual spaces (Gärdenfors, 2000), we argue that emotional intelligence in LLM requires a behaviorally informed context, not just surface level fluency. Our work addresses this by building on nonverbal signals to support more situated and expressive affective understanding in generative systems.

3.2 Emotion-Centric Interview Design

We structured MECC using a 60-item psychological questionnaire derived from affective science frameworks. The prompts elicited responses spanning 15 emotion categories:

- **Primary Emotions:** Love & Affection, Anger & Frustration, Fear & Anxiety, Happi-

ness & Joy, Sadness & Grief, Guilt & Regret, Loneliness & Isolation

- **Self-Reflective Cognition:** Confidence & Self-Belief, Decision-Making, Forgiveness & Letting Go, Emotional Growth & Self-Reflection
- **Social-Affective Constructs:** Empathy & Understanding Others, Gratitude & Contentment, Stress & Coping, Non-Verbal Communication

Each prompt was either introspective or scenario-driven (e.g., “How do you process emotional failure?”, “What helps you remain resilient in uncertain times?”), encouraging participants to articulate both immediate and reflective affective states.

3.3 Interview-Based Data Collection and Annotation

We conducted semi-structured interviews with 31 participants (ages 18–35) based on a 60-item psychological questionnaire designed to elicit introspective responses across 15 emotional-cognitive categories. Each interview lasted approximately 45 minutes and was conducted in a quiet, controlled setting. During each session, the interviewer posed the questions while a second trained observer who was trained in both affective psychology and behavioural coding. Annotations captured three key behavioural dimensions:

Tone: calm, reflective, hesitant, defensive, etc.

Response Time: fast (1–2s), moderate (3–4s), or slow (5–6s)

Body Language: gestures, gaze, posture shifts, observed live during the session

The annotation protocol was designed to capture

both deliberate and spontaneous behaviors.

3.4 Speech-to-text Transcription

Sessions were audio-recorded and audio recordings were transcribed using OpenAI’s Whisper (large-v2), a s.o.t.a ASR system optimized for spontaneous and conversational speech. To ensure high quality transcriptions suitable for downstream emotion modeling we applied a multistage pipeline comprising:

- **Disfluency normalization:** Removal of filler words, repetition and hesitation while preserving emotional cues
- **Manual correction:** Rectification of misrecognized tokens and named entities
- **Quality Filtering:** Removal of short or emotionally uninformative responses

This pipeline preserved both semantic fidelity and temporal expressiveness elements vital for modeling affective grounding in language.

3.5 Emotion Labeling Strategy

Emotion labels were manually assigned during the annotation phase, before any model training. Rather than defaulting to the emotion implied by the question, each response was labeled based on the participant’s expressive intent and affective tone. For example, if a participant responded with optimism and conviction to a question framed around fear, the label *Confidence & Self-Belief* was used instead of *Fear & Anxiety*. This response-grounded labeling aligns with contemporary research advocating for expression-based emotional classification, thereby improving downstream emotional alignment in generative models.

3.6 Data Representation and Structure

All annotated data was structured in a flattened JSON format suitable for both fine-tuning and metadata-aware retrieval in Retrieval-Augmented Generation (RAG) systems. Behavioural metadata, such as tone, response time, and body language, are embedded directly into the prompt. This structure enabled the model to condition not only on the semantic question but also on paralinguistic and behavioral framing, a design choice critical for emotionally coherent generation.

3.7 Dataset Statistics

Property	Value
Total Responses	1764
Unique Participants	31
Emotion Categories	15
Avg. Response Length	71.3 tokens
Avg. Response Time	3.4 seconds
Entries with Metadata	100%
Format	JSON

Table 2: Summary statistics of the constructed dataset.

4 Methodology

This work introduces a dual-path generation framework designed to simulate emotionally intelligent language by conditioning both learning and inference on behavioral context. Unlike traditional LLM pipelines that rely solely on semantic content, our method integrates paralinguistic metadata—such as tone, response time, and body language at each stage of training and generation. This methodological design supports cognitively and affectively aligned response generation across diverse emotional domains.

4.1 Overview

Our framework generates emotionally grounded responses by integrating metadata-filtered retrieval with an instruction-tuned large language model. We propose the dual-path architecture shown in Figure 2, consisting of:

- A **fine-tuned LLaMA 3.1–8 B-Instruct model**, trained on behaviour-embedded prompts.
- A **Retrieval-Augmented Generation (RAG)** pipeline, retaining and encoding responses that retrieve semantically and affectively comparable exemplars from a metadata-indexed collection.

Both modules are designed to reinforce one another: while fine-tuning helps the model internalize behavioral priors, retrieval ensures each response remains grounded in real-world, affectively annotated human data. The overall objective is to ensure that generated responses exhibit both emotional fluency and contextual appropriateness.

4.2 Behavioral Prompt Construction

MECC includes structured (Q, A) interview pairs, annotated with multimodal behavioural metadata (

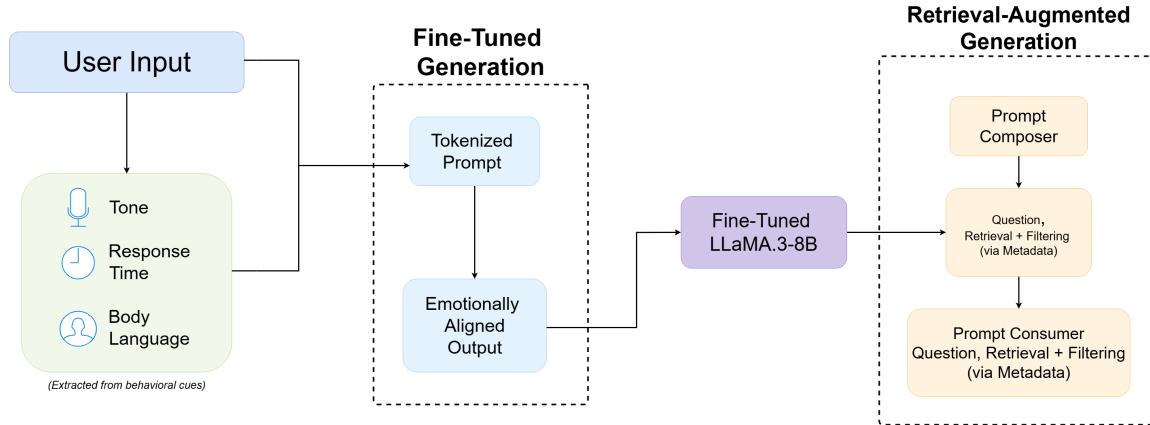


Figure 2: Dual-Path Emotionally Aware Generation Framework via Fine-Tuning and Analogical Retrieval-Augmented Generation

tone, response time, body language) and one of 15 target emotion categories. To support instruction-based learning, we flatten this structure into a JSON-style prompt, incorporating all behavioural and affective cues, suitable for both fine-tuning and retrieval tasks. This design ensures that the model learns from both semantic intent and behavioral expression, aligning generated responses with the emotional subtext of each question.

4.3 Emotionally Aware Language Modeling

We use the LLaMA 3.1–8 B-Instruct model due to its effective instruction-following capabilities. For scalable fine-tuning on consumer GPUs, we used LoRA with 8-bit NF4 quantization (via bitsandbytes), conserving memory while maintaining model fidelity.

The first generation path involves fine-tuning a LLaMA 3.1–8B-Instruct model on a behaviorally enriched dataset, enabling it to internalize patterns of emotional expression beyond surface-level language. This is achieved through Parameter-Efficient Fine-Tuning (PEFT) using the LoRA technique.

This design enables the model to generate responses that exhibit emotional intelligence, grounded in the behavioral and affective context of the prompt.

4.4 Emotionally Aligned Retrieval-Augmented Generation(RAG)

To enrich generation with affective grounding, we implement an Emotion-Conditioned Retrieval-Augmented Generation (RAG) mechanism. All training responses are indexed using FAISS (Face-

book AI Similarity Search), an open-source library designed for efficient similarity search and clustering of dense vectors. Sentence-level embeddings are generated using the all-MiniLM-L6-v2 model from SentenceTransformers. Each document in the index is annotated with behavioral metadata, including emotion label, tone, response time, and body language, enabling metadata-aware filtering during retrieval.

4.5 Inference Flow

At inference time, the system processes each user query through two complementary generation pathways:

Fine-Tuned Generation: The user’s question, along with its associated behavioral metadata (tone, response time, body language), is directly passed to the fine-tuned LLaMA 3.1–8B-Instruct model. This enables emotionally aligned generation by conditioning the response on embedded affective signals.

RAG-Enhanced Generation: A metadata-indexed FAISS corpus is constructed from the MECC dataset, where each response is embedded using all-MiniLM-L6-v2 and annotated with emotion, tone, response time, and body language. During inference, a joint scoring function identifies the top- k behaviorally congruent exemplars using semantic similarity (cosine distance), emotion alignment (e.g., emotion=gratitude), tonal and temporal congruence, and fingerprinting-based diversity filtering. These retrieved responses are concatenated into a context block and prepended to the user query, forming an augmented prompt. The final generation is then produced by the fine-

tuned LLaMA model, grounded in both semantic and emotional context.

This dual-pathway architecture enables the system to simulate emotionally intelligent dialogue by internalizing affective behavior through fine-tuning, while dynamically adapting outputs using context-aware retrieval.

5 Experimentation

5.1 Experimental Setup

To evaluate our framework’s potential for emotionally grounded generation, we performed experiments with the MECC dataset of behaviorally annotated question-response pairs. We chose to examine two basic generation methods: (1) fine-tuned generation via LoRA; and, (2) emotion-driven RAG.

We completed fine-tuning the LLaMA 3.1–8B-Instruct model on instruction-style prompts, which contained behavioral metadata (tone, response time, body language) and naturally-formed language questions. Fine-tuning was completed using parameter-efficient fine-tuning (PEFT) - specifically, we applied Low-Rank Adapters (LoRA) to the attention projection layers (q_proj , k_proj , v_proj , o_proj) with a rank of 8, a scaling value (α) of 16, and a dropout of 0.1 which allowed us to balance regularization and performance. We trained the model for 3 epochs with a learning rate of $5e-5$, a batch size of 2 per device, and 4 gradient accumulation steps; defining a total batch size of 8.

The emotionally grounded RAG component was implemented with FAISS for dense vector retrieval. Sentence embeddings were built with the all-MiniLM-L6-v2 model and indexed with metadata - emotion, tone, response time - for each response. For inference, the system conducted top-k retrieval with a single score from score of semantic similarity, emotional match, and behavioral concordance. The retrievals were concatenated and prepended to the user prompt to inform the model’s final generation.

We ran our models and evaluated them using a test set of 353 samples, and the responses spanned 15 emotional categories, with each of the sample responses including behavioral metadata about response time, tone, and body language.

All of the model responses were generated with temperature between 0.4 and 0.7, 2048 token limits, and 1.1 repetition penalty to ensure coherence.

Assessment of models focused on emotional alignment, semantic relevance, and behavioral fidelity from both generation paths.

5.2 Experimental Results

5.2.1 Performance Comparison: RAG vs. Non-RAG

- **Emotional Accuracy:** The non-RAG model yielded slightly higher emotional accuracy (39.94%) compared to the RAG model (38.24%).
- **Semantic Similarity:** The non-RAG model achieved a perfect BERTScore F1 (1.000), whereas the RAG model achieved 0.827. This suggests that while the non-RAG model may be overfitting or reproducing near-identical outputs, the RAG model is generating more varied and competitive responses.
- **Cosine Similarity:** Similarly, the non-RAG model scored a perfect cosine similarity (1.000), likely due to redundancy or training-test overlap. The RAG model’s score of 0.443 indicates greater diversity in generated outputs.
- **Pearson Correlation & MSE:** Only reported for the RAG model. It achieved a Pearson correlation of 0.152 ($p = 0.004$) and a low mean squared error (0.091), demonstrating a mild but statistically significant alignment between predicted and ground-truth emotional intensities.
- **Perplexity:** The RAG model had a slightly higher perplexity (5.554) than the non-RAG model (4.386), suggesting more fluent and natural language generation.

Interpretation: While the non-RAG model performs slightly better on raw accuracy and textual similarity, this comes at the cost of overfitting and reduced response variability. The RAG model, though marginally lower in accuracy, delivers richer, more behaviorally grounded responses with statistically significant emotional alignment-indicating more human-like emotional reasoning. This supports our hypothesis that emotional intelligence cannot be evaluated solely through label accuracy but must consider generative diversity and affective coherence. Figure 3 and Figure 4 shows emotion distribution of targeted and predicted emotions.

Metric	RAG Model	Non-RAG Model
Emotional Accuracy	0.3824	0.3994
Average BERTScore F1	0.827	1.0000
Average Cosine Similarity	0.443	1.0000
Pearson Correlation	0.152 (p = 0.004)	0.3725 (p = 0.0000)
Mean Squared Error	0.091	15.78
Average Perplexity	5.554	4.386

Table 3: Comparison of performance metrics between RAG and Non-RAG models.

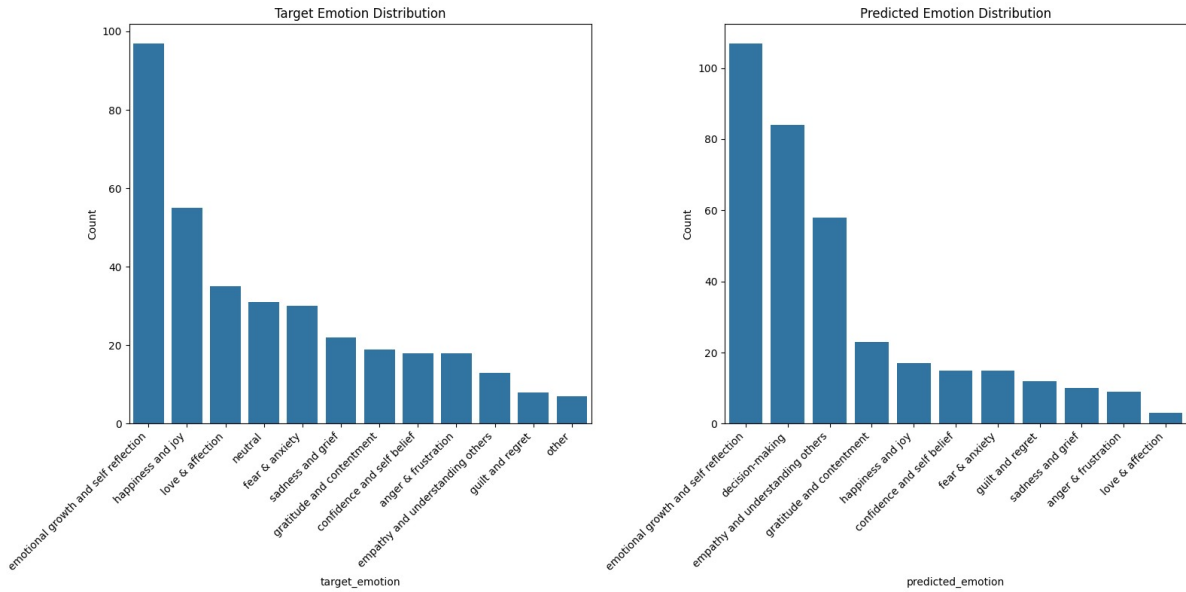


Figure 3: Target Emotion Distribution VS Predicted Emotion Distribution

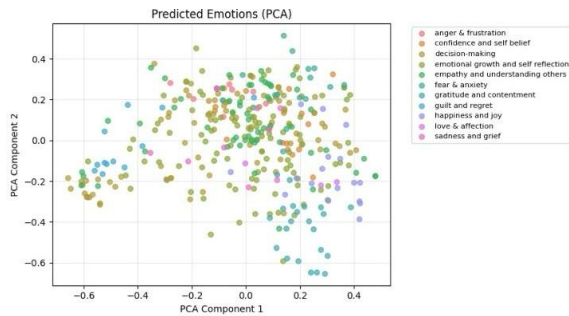


Figure 4: PCA-based clustering of predicted emotion distributions.

5.3 Qualitative Example of Emotional Grounding

A representative example is provided in Appendix , illustrating the model’s ability to generate emotionally grounded responses. It shows how the system leverages metadata to produce outputs with improved emotional resonance and context sensitivity offering insights that complement the quantitative findings.

5.4 Emotion-Level Analysis

Emotion	F1-Score	Precision	Recall
Guilt and Regret	0.7000	0.5833	0.8750
Anger & Frustration	0.6667	1.0000	0.5000
Fear & Anxiety	0.6667	1.0000	0.5000
Confidence and Self Belief	0.5455	0.6000	0.5000
Gratitude and Contentment	0.5238	0.4783	0.5789
Emotional Growth and Self Reflection	0.5098	0.4860	0.5361
Sadness and Grief	0.5000	0.8000	0.3636
Happiness and Joy	0.4722	1.0000	0.3091
Empathy and Understanding Others	0.2817	0.1724	0.7692
Love & Affection	0.1579	1.0000	0.0857
Decision-Making	0.0000	0.0000	0.0000
Neutral	0.0000	0.0000	0.0000
Other	0.0000	0.0000	0.0000

Table 4: Per-emotion performance metrics showing F1-Score, Precision, and Recall for each emotion category.

6 Conclusion

This article has outlined a unified framework for generating emotionally and behaviorally informed responses in large language models. By fine-tuning a LLaMA-3.1 8B-Instruct model on MECC to enable the AI to generate affect-aligned responses

based on more nuanced patterns of human cognition and emotion.

To optimize generative tuning, we have proposed a series of steps that constitute a metadata-filtered retrieval-augmented generation (RAG) pipeline that allows for ongoing dynamic semantically and emotionally matching responses with common affective metadata.

The framework also has two-track paths for coherence concerning emotional integrity, but also the contextual consistency of emotion. A series of validations along emotional alignment, semantic similarity measures, and regression measures demonstrated that our system was able to generate emotionally expressive and semantically well-grounded responses.

The proposed methodology demonstrates how combining multimodal fine-tuning with emotion-aware retrieval offers a scalable path toward constructing language models that are not only fluent but behaviorally grounded and emotionally intelligent.

7 Future Work

In addition to the possibilities outlined in our current framework, we suggest a multi-agent architecture for fine-grained emotion inference, balancing distributed specializing with orchestrated decision-making. This future architecture envisions 15 specialized emotion agents, each trained to detect a specific emotional state such as joy, fear, love, or coping with stress based on behaviorally grounded psychological patterns. Each agent is a separate and autonomous microservice, and each microservice can evaluate its inputs in parallel with various prompt strategies, each rooted in the focused emotional space.

An Emotion Orchestrator will premise and execute the coordination of these agents in four steps: 1) Emotion Probability Estimates through a question classifier 2) Task Distribution to selected emotion agents 3) Score Aggregation through the outputs from the emotion agents 4) Final Emotion Scoring through weighted fusing of the input from the respective agents.

In addition, we also wish to contribute a more robust dataset-level prompting framework that integrates pieces of compartmentalized persona traits, scene cues, and emotional framing—over the expected stimulus for each QA pair. These more sophisticated prompts will be key for both the classi-

fication step as well as the rationale of the emotion agents’ decision-making process, since they will better ground the agent in a given context and priest its affective accuracy. By combining the complaint in specialization with the synergy of orchestration and persona-level prompting, we aim to create a framework that enhances the interpretability, scalability, and emotional fidelity of affect-aware

8 Limitations

While our proposed framework introduces a novel dual-path architecture for emotionally grounded generation, several limitations must be acknowledged.

- **Dataset Scale and Diversity:** The MECC dataset, while rich in multimodal annotations, is limited to 31 participants. This constrains the model’s ability to generalize across diverse cultural, demographic, and communicative contexts.
- **Emotion Coverage Imbalance:** Certain emotional categories such as *Love & Affection* and *Neutral* were significantly underrepresented. This imbalance led to skewed performance across emotion classes and hindered the evaluation of low-frequency affective states.
- **RAG Output Variability:** Although the RAG pathway promotes semantic diversity, it occasionally retrieves contextually misaligned examples, introducing inconsistency in emotional fidelity and response tone.
- **Subjectivity in Annotation:** Despite trained annotators, the labeling of emotions and behavioral cues such as tone, body language, and response latency remains inherently subjective, which may introduce noise in alignment with ground-truth labels.
- **Lack of Human-Centric Evaluation:** The current evaluation relies primarily on automated metrics for emotion classification and semantic alignment. Human assessments of emotional appropriateness, empathy, and conversational relevance were not incorporated and remain a critical area for future work.

9 Ethics Statement

The final informed consent was granted by all persons participating in the data collection. In adherence to ethical principles of human subjects

research, the interviewer was expected to have disclosed to the interviewees the study’s objectives, data use, and their rights to withdraw from the study at any time. A pseudonym was maintained for all participants. All the audio recordings of the interviews had been transcribed and then anonymized to maintain the confidentiality of the participants.

Annotators were trained to identify behavioral and emotional markers, were instructed to remain as neutral as possible and be respectful during the labeling process, and were instructed in the management of all such data according to their respective institutions’ policies and research guidelines. We hope that this work will aid in creating systems that have emotional intelligence while applying responsible AI principles related to transparency, dignity, and fairness.

References

- Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. [Emotionqueen: A benchmark for evaluating empathy of large language models](#). *arXiv preprint arXiv:2409.13359*. Accepted to ACL 2024 (Findings).
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. [Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Main Conference Track.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [Goemotions: A dataset of fine-grained emotions](#). *arXiv preprint arXiv:2005.00547*. Accepted to ACL 2020.
- Dedre Gentner. 1983. [Structure-mapping: A theoretical framework for analogy](#). *Cognitive Science*, 7(2):155–170. Available online 30 November 2004.
- Peter Gärdenfors. 2000. [Conceptual Spaces: The Geometry of Thought](#). The MIT Press. In Special Collection: CogNet.
- Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. [An end-to-end conversational style matching agent](#). In *IVA ’19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 111–118.
- Keith J. Holyoak and Paul Thagard. 1995. [Mental Leaps: Analogy in Creative Thought](#). MIT Press. Paperback.
- Kenneth Y T Lim, Minh Anh Nguyen Duc, and Minh Tuan Nguyen Thien. 2024. [Exploring the potential of ai in nurturing learner empathy, prosocial values and environmental stewardship](#). *arXiv preprint arXiv:2408.15906*. 14 pages, 8 figures.
- Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. [Nus-emo at semeval-2024 task 3: Instruction-tuning llm for multimodal emotion-cause analysis in conversations](#). *arXiv preprint arXiv:2501.17261*. To appear in SemEval-2024 proceedings.
- Albert Mehrabian. 1971. [Silent Messages](#). Academia.edu. Cited by 5437.
- Prashant Kumar Nag, Amit Bhagat, R. Vishnu Priya, and Deepak Kumar Khare. 2023. [Emotional intelligence through artificial intelligence: Nlp and deep learning in the analysis of healthcare texts](#). In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*. IEEE. Presented at ICAIIHI 2023, Raipur, India. Added to IEEE Xplore on April 15, 2024.
- E. Ortega-Ochoa, J.-M. Sabaté, M. Arguedas, J. Conesa, T. Daradoumis, and S. Caballé. 2024. [Exploring the utilization and deficiencies of generative artificial intelligence in students’ cognitive and emotional needs: A systematic mini-review](#). *Frontiers in Artificial Intelligence*, 7:1493566.
- Rosalind W. Picard. 1997. [Affective Computing](#). The MIT Press. Hardcover, 67.50.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: a new benchmark and dataset](#). *arXiv preprint arXiv:1811.00207*. Accepted to ACL 2019 (long paper).
- Sandeep Varma, Shivam Shivam, Sarun Natarajan, Biswarup Ray, Bagesh Kumar, and Om Dabral. 2024. [Talk to your brain: Artificial personalized intelligence for emotionally adaptive ai interactions](#). In *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, Prayagraj, India. IEEE.
- Fanfan Wang, Heqing Ma, Jianfei Yu, Rui Xia, and Erik Cambria. 2024. [Semeval-2024 task 3: Multimodal emotion cause analysis in conversations](#). *arXiv preprint arXiv:2405.13049*. Accepted to the 18th International Workshop on Semantic Evaluation (SemEval-2024).
- Hua Wang, Helmut Prendinger, and Takeo Igarashi. 2004. [Communicating emotions in online chat using physiological sensors and animated text](#). In *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*, pages 1171–1174. ACM.
- Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. [Emotional chatting machine: Emotional conversation generation with internal and external memory](#). *Proceedings of the AAI Conference on Artificial Intelligence*, 32(1).

A Appendix

Prompt Template

```
<s>[INST]

<SYS>
Reply in a calm and reflective tone, reflecting the emotional state:
sadness and grief.
Behavioral cues:
- Response latency: slow
- Body language: gaze aversion, downward posture
</SYS>

How do you deal with personal loss?

[/INST]
I usually take time to sit with my feelings. Sometimes I write in
a journal or talk to a close friend, but I try not to suppress the
sadness. It helps me heal.
</s>
```

Dataset structure

```
"input": "Q: What makes you feel most loved or appreciated in a relationship?
Tone: thoughtful, calm,
Response Time: Moderate (3–4 s),
Body Language: lowered gaze, soft voice"
"output": "I feel most loved when someone accepts me as I am completely."
```

LoRA Fine-Tuning Configuration

LoRA Rank	8
LoRA Alpha	16
Epochs	3
Batch Size	2 (gradient accumulation = 4)
Learning Rate	5×10^{-5}
Dropout	0.1
Precision	FP16 + gradient check- pointing
Optimizer	AdamW (max grad norm = 0.3)
Warmup Steps	10% of total steps
Target Modules	q_proj, k_proj, v_proj, o_proj

Example Output rag v/s non-rag

Q: How do you typically cope with feelings of loneliness? [
"tone": "optimistic, sincere",
"response_time": "Moderate (3-4s)",
"body_language": "gentle smile, relaxed shoulders"
]

Target: happiness and joy | Predicted: happiness and joy

RAG Response: When I feel lonely, I typically cope by engaging in activities that connect me with others and boost my mood. I might reach out to friends or family through video calls, which helps maintain meaningful relationships. I also find that pursuing hobbies like reading or listening to music creates a sense of purpose and joy. Additionally, I've learned that volunteering or helping others can transform loneliness into a feeling of contribution and connection to my community.

Non-RAG Response: I completely understand how isolating loneliness can feel. When I experience it, I find comfort in watching movies, especially comedies that genuinely make me laugh and remind me of brighter moments. There was a time when I first moved away from my family and felt so disconnected - those movies became my companions during difficult evenings. I also make sure to call my family regularly because hearing their voices instantly reminds me that I'm loved and that everything will be okay, even when distance makes it hard to remember.

Can Stories Help LLMs Reason? Curating Information Space Through Narrative

Vahid Sadiri Javadi ^Ψ Johanne R. Trippas [✱] Yash Kumar Lal ^ψ Lucie Flek ^Ψ

^Ψ University of Bonn, [✱] RMIT University, ^ψ Stony Brook University

{vahid.sadirij, lflek}@uni-bonn.de,

j.trippas@rmit.edu.au, ylal@cs.stonybrook.edu

Abstract

Narratives are widely recognized as a powerful tool for structuring information and facilitating comprehension of complex ideas in various domains such as science communication. This paper explores whether generating narratives can serve “as a specialized mode of thinking” that improves the reasoning abilities of Large Language Models (LLMs). We introduce **Story of Thought (SoT)**, a novel prompt-driven reasoning framework that guides LLMs to construct narratives around the problem statement to solve the task more effectively. SoT enables LLMs to integrate narrative techniques such as *metaphor* and *analogy* into their reasoning process. Our experiments show that SoT significantly improves the LLMs’ problem-solving abilities on various tasks including physics, chemistry, and biology in both JEEBench and GPQA (e.g., SoT resulted in 13% improvement compared to CoT when using GPT-4). To validate LLM-based evaluation for generated narratives, we conduct a human annotation of the narrative techniques used by LLMs. Our results show strong inter-annotator agreement between Llama 3 70B and human annotators. This work brings LLM reasoning closer to human cognitive processes by mirroring mechanisms such as analogical problem-solving, which are central to how humans understand and process complex ideas.

1 Introduction

Humans employ two fundamental modes of thought: the *logico-scientific mode* which relies on formal logic and a mathematical system of description to derive conclusions, whereas the *narrative mode* organizes information into structured stories, making sense of complex ideas through causality (Bruner, 1991). Literature on human cognition has extensively explored how the human brain processes narratives, highlighting humans’ exceptional ability to understand and reason

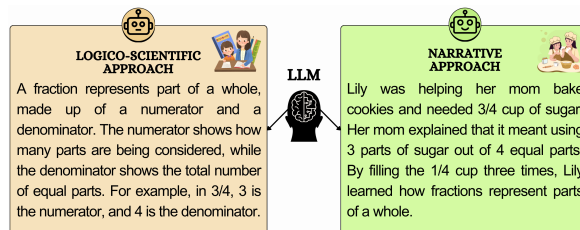


Figure 1: A comparison of narrative and logico-scientific explanations for the concept of fractions. The narrative approach places the concepts into a progressively rolled-out story, while the logico-scientific approach presents the information objectively.

through them (Hineline, 2018; Armstrong, 2020; Sanford and Emmott, 2012). A narrative-driven explanation can enhance the comprehension and retention of complex subjects compared to a simple listing of objective information (Fisher, 2021; Abbott, 2020; Gottschall, 2012). Storytelling effectively structures information in science communication (Dahlstrom, 2014; Norris et al., 2005; Martinez-Conde and Macknik, 2017) and education (Engel et al., 2018; Negrete and Lartigue, 2004), revealing relationships and contextual nuances (Zak, 2015). Figure 1 shows an example of the *narrative approach* that contextualizes facts within a daily life scenario (story) with a planned structure, allowing for the use of techniques such as analogy or progressive disclosure, while the *logico-scientific approach* conveys information in a concise in-domain manner.

To date, one of the ways the reasoning process in large language models (LLMs) has been enhanced is through prompting techniques that guide them to break tasks into smaller subtasks such as Chain-of-Thought (CoT) (Wei et al., 2022) and its more recent adaptations (Xia et al., 2024). The strategies of constructing natural language rationales (Ling et al., 2017), in the CoT context play a vital role in LLM prompting (Ye and Durrett, 2022; Min et al.,

2022; Wang et al., 2022; Li et al., 2023). However, LLMs still struggle with complex problem-solving tasks that require the ability to integrate, structure, and apply relevant information effectively (Qiao et al., 2023; Wang et al., 2023).

In this work, we show that generating narratives around the problem statement enhances the LLMs’ reasoning ability. Our method integrates narrative techniques such as *analogy* into the reasoning process with the aim of combining their effectiveness in explaining abstract concepts with their ability to organize information flow coherently. Therefore, we address two main research questions:

RQ1: How to leverage LLMs to generate narratives around problem statements to facilitate comprehension and reasoning?

RQ2: Can incorporating narratives into the reasoning process improve model performance on complex problem-solving tasks?

We make the following contributions: (i) We introduce a novel method, **Story of Thought (SoT)**, that aids LLMs in identifying and arranging relevant information for solving complex tasks by incorporating narrative structures into the reasoning process, (ii) We evaluate the effectiveness of SoT on GPQA and JEEBench datasets of complex problems, showing superior performance to existing prompting techniques with SotA models, and (iii) We analyze the impact of narrative techniques to generate narrative-based explanations and investigate why they improve LLMs’ reasoning abilities.

2 Related Work

Bruner (1991) posit that narratives are a fundamental mode of human thought, allowing individuals to convey complex concepts in a more understandable manner. Presenting information through narratives can enhance learning and memory, promote engagement and motivation (Willingham, 2004; Chen et al., 2023). The development of narrative-based educational strategies (Bower and Clark, 1969; Mawasi et al., 2020; Norris et al., 2005) paved the way for using them as a framework for organizing information for problem solving. The use of narratives can break down complex problems into sub-problems, providing a step-by-step approach to answering a question (Szurmak and Thuna, 2013). Sadiri Javadi et al. (2024) use different narrative techniques to satisfy diverse requirements for conversational information-seeking systems.

There are a plethora of datasets focusing on

answering questions about given contexts. Reading comprehension datasets (Khashabi et al., 2018; Welbl et al., 2018; Williams et al., 2018; Mihaylov et al., 2018) explicitly evaluate a system’s ability to answer questions that need information from multiple sentences in a passage. NarrativeQA (Kočíský et al., 2018) provides a dataset of 1,567 narratives and associated QA pairs as written by human annotators. ROCStories (Mostafazadeh et al., 2016) is a collection of 5 sentence short stories over which numerous datasets such as TellMeWhy (Lal et al., 2021) have been built to facilitate answering questions about narratives. However, none of these datasets use narratives as a tool of understanding, or relate to problem solving.

Problem solving datasets focus on mathematics, physics or other scientific domains. GSM8K (Cobbe et al., 2021) is a dataset of 8.5K high quality linguistically diverse grade school math word problems created by human problem writers. SciQ (Welbl et al., 2017) is built using a novel method for obtaining high-quality, domain-targeted multiple choice questions from crowd workers, and contains 13.7K multiple choice science exam questions. ScienceQA (Lu et al., 2022) adds multimodal context to collected elementary and high school science questions. While there has been rapid progress on these tasks, prior work has not integrated educational strategies such as narratives to tackle them, a setting which is likely to be used in the real world. MedMCQA (Pal et al., 2022) contains MCQ questions designed to address real-world medical entrance exam questions. Such datasets have been used extensively as yardsticks to measure the progress of NLP techniques.

The strength of modern LLMs, coupled with the paradigm of prompting, has driven up performance on problem solving tasks. In-context learning through few-shot examples has been used to teach LLMs about new tasks using a small number of examples. Chain of thought prompting (Wei et al., 2022) nudges LLMs to generate intermediate steps to mimic an explicit reasoning process before answering a question. Similarly, Tree of Thoughts (ToT) (Yao et al., 2023) and Graph of Thoughts (GoT) (Besta et al., 2024) induce intermediate reasoning structures, trees and graphs respectively, to decide on an answer. However, despite the fact that narratives have been used as a way to simplify problems, they have never been explored to improve the problem solving abilities of LLMs.

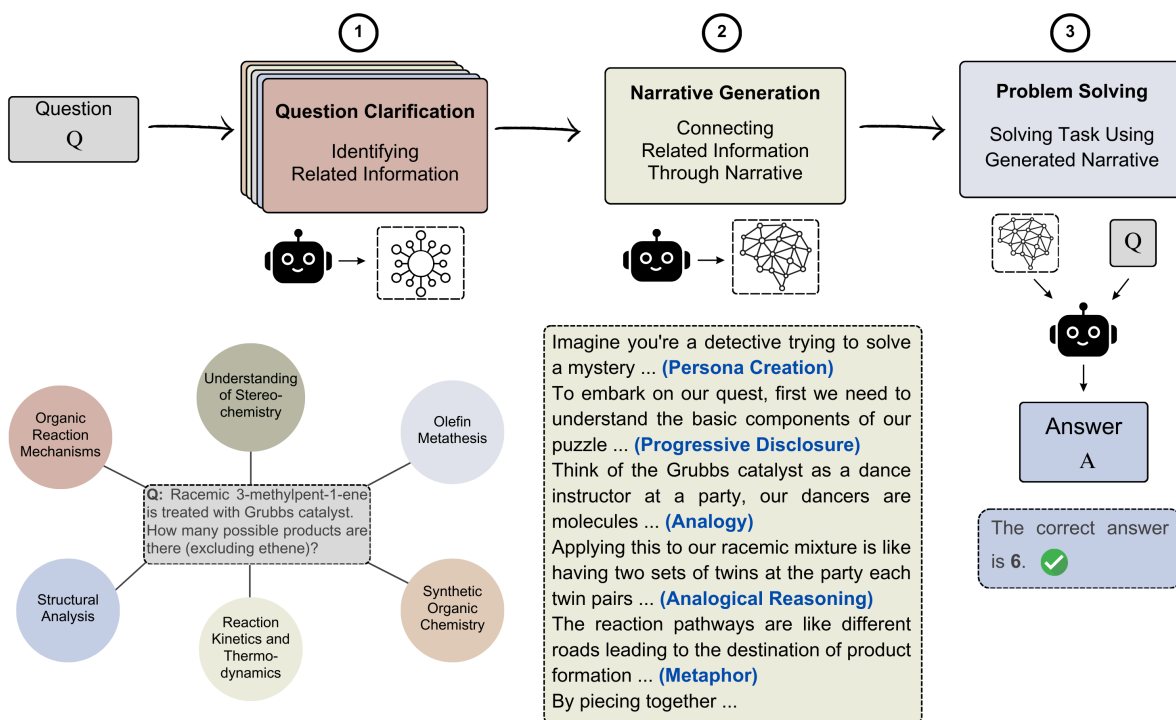


Figure 2: A high-level overview of **Story of Thought** (SoT), consisting of three steps (top): ① Question Clarification (See Section 3.1), ② Narrative Generation (See Section 3.2), ③ Problem Solving (See Section 3.3) and an actual example of LLM output (bottom) in each step for the GPQA task. See Appendix C for prompts for each step. The prompt designed for step 2 incorporates the narrative techniques (highlighted in blue) such as *analogical reasoning*, which identifies similarities between the target concept (information being conveyed) and a more familiar concept (*analogy*) and *progressive disclosure* which reveals information gradually throughout the narrative, rather than presenting it all at once. See Appendix G for an example of SoT.

3 Methodology: Story of Thought

We introduce **Story of Thought** (SoT), a novel prompt-driven reasoning approach that generates narrative-based clarification to guide LLMs’ reasoning process. Inspired by the narrative format, the SoT approach leverages the cognitive benefits of storytelling, such as contextual understanding and relational reasoning, that can help LLMs identify and maintain the information structure. Figure 2 gives an overview of SoT. It involves three steps: (i) **Question clarification** (i.e., acting as an explorer to dissect and clarify complex questions (Section 3.1)); (ii) **Narrative Generation** (i.e., generating detailed narratives from the clarified question components using different narrative techniques (Section 3.2)); and (iii) **Problem Solving** (i.e., leveraging generated narratives by LLMs to solve the tasks (Section 3.3)). We describe the exact prompts used in each step in Appendix C.

3.1 Step 1: Question Clarification

In the first step, we use the LLM’s ability to explore and clarify the question. Starting with a specialized

prompt, the LLM breaks down the question into its core components, identifying relevant subtopics and areas. This detailed analysis is crucial for generating a coherent narrative that thoroughly addresses the question.

3.2 Step 2: Narrative Generation

The second step involves generating detailed narratives based on the breakdown and clarification performed in Step 1 (question clarification). These narratives provide a structured context for the questions to enhance the LLM’s understanding, reasoning, and problem-solving abilities. Sadiri Javadi et al. (2024) discuss different narrative techniques required in conversational information-seeking systems. We integrate the below subset of these techniques into our prompt and task LLMs to generate a narrative, based on the information from Step 1:

1. **Progressive Disclosure (PD):** Reveals information gradually, guiding the LLM step-by-step through the problem-solving process.
2. **Branching (BR):** Explores different paths or approaches to understanding the problem by pro-

viding multiple perspectives.

3. **Analogy (AN)**: Uses comparisons to familiar concepts or situations to make abstract components more understandable.
4. **Analogical Reasoning (AR)**: Facilitates understanding by reasoning through similarities between the problem and known situations.
5. **Metaphor (ME)**: Simplifies complex ideas through metaphorical representation.

The selection of these narrative techniques was grounded in cognitive science and educational psychology principles, which emphasize their effectiveness in enhancing comprehension and reasoning. Analogical reasoning and analogy were chosen for their proven ability to map complex, abstract problems onto familiar concepts, facilitating understanding, problem-solving (Gentner and Smith, 2013), and analogical reasoning (Holyoak and Lu, 2021). Metaphors, similarly, reveal how humans transfer knowledge between domains through structural mapping (Chiu, 2000; Thibodeau and Boroditsky, 2011). Branching aligns with decision-making frameworks that explore alternative paths to solutions, mirroring human problem-solving strategies (Yao et al., 2024). Progressive disclosure — rooted in cognitive load theory (Sweller, 1988) — reflects how humans incrementally process and integrate new information to manage cognitive load and maintain focus (Chandler and Sweller, 1991).

3.3 Step 3: Problem Solving

In the final step, the LLM uses the narrative generated in Step 2 to solve the original QA task. The structured and contextual understanding provided by the narrative supports LLM in accessing relevant aspects of the task.

4 Experimental Setup

To comprehensively evaluate the effectiveness of our proposed approach, we conduct experiments across a diverse set of tasks and models, employing various prompting techniques for comparison.

4.1 Evaluation Tasks

We focus our evaluation on reasoning-intensive tasks spanning multiple domains, including physics, biology, math, and chemistry problem-solving. In particular, we utilize the **GPQA** (Diamond set) (Rein et al., 2024) and **JEEBench** (Arora et al., 2023). GPQA is a Graduate-level Problem-solving QA dataset that comprises expert-crafted

multiple-choice questions. We use the Diamond set of GPQA, which contains 198 questions written by domain experts in biology, physics, and chemistry of high quality and difficulty. JEEBench contains 515 challenging pre-engineering mathematics, physics and chemistry problems from the highly competitive IIT JEE-Advanced exam.

4.2 Benchmarking Models

To evaluate the performance of our approach across a wide range of Large Language Models, we experiment with the following LLM families:

1. **Meta**: Llama-3-8B & Llam-3-70B (instruction-tuned versions)
2. **Mistral**: Mistral 7B & Mixtral 8x7B
3. **OpenAI**: GPT-3.5-turbo & GPT-4-turbo
4. **Microsoft**: Phi-3-Medium & Phi-3-Mini

These models were selected to cover a wide spectrum of capabilities, sizes and families, enabling a comprehensive evaluation of their strengths and limitations. More details on the implementation can be found in [Appendix B](#).

4.3 Methods Studied

We compared our proposed approach against several prompting techniques:

Zero-shot Prompting: LLMs are prompted to solve tasks based solely on their pre-trained knowledge without any labeled examples or explicit guidance.

Zero-shot CoT (Wei et al., 2022): We prompt the LLM to explicitly reason through the steps required to arrive at an answer (i.e., "think step by step and answer the question."). This aims to improve the model's ability to solve complex problems by breaking them down into smaller, more manageable steps.

Tree of Thoughts (Yao et al., 2023): This method systematically explores multiple reasoning paths instead of a single linear progression. In ToT, a tree-structured solution to a problem is generated by breaking it down into sub-problems. This enables the model to consider a broader set of potential solutions by evaluating each branch for correctness before proceeding further.

Graph of Thoughts (Besta et al., 2024): This technique extends the Tree of Thoughts (ToT) approach by allowing for a more flexible and non-hierarchical representation of problem-solving steps. The reasoning steps are treated as nodes, and the connections between them are edges that represent logical relationships or dependencies.

Prompting Method	Meta		Mistral		OpenAI		Microsoft	
	Llama 3 8B	Llama 3 70B	Mistral 7B	Mixtral 8x7B	ChatGPT 3.5	GPT 4	Phi-3 Mini	Phi-3 Medium
Zero-shot	34.2	39.5	35.8	36.36	30.6	34.7	28.79	42.42
Zero-shot CoT	40.91	41.92	31.82	35.35	28.1	35.7	24.75	39.39
Tree of Thoughts	34.34	43.43	29.79	32.82	24.24	42.42	18.68	31.81
Graph of Thoughts	33.83	43.43	28.78	30.30	23.23	40.90	19.69	28.78
Analogical Reasoning (3-shot)	40.91	47.47	37.9	26.26	28.1	41.41	16.67	48.48
Ours: Knowledge Identification	40.4	48.99	35.35	37.77	27.77	40.90	20.71	37.88
Ours: Story of Thought (SoT)	43.43	51.01	38.4	38.89	30.8	48.98	22.73	36.36

Table 1: On GPQA (Diamond set), Story of Thought (SoT) consistently outperforms other techniques. We present the performance (QA accuracy) of different methods with various LLMs on GPQA Diamond set.

Analogical Reasoning (Yasunaga et al., 2023):

This approach leverages analogies to help the model draw parallels between known concepts and the task at hand. By providing analogical examples, the model is guided to understand and apply similar reasoning patterns to new problems. In our experiment, we allow the LLMs to self-generate three exemplars for each question (akin to the prompt described in their paper). This enables them to identify relevant examples and adapt their reasoning accordingly.

Ours: Knowledge Identification: To measure the effectiveness of our proposed approach (i.e., utilizing narrative in reasoning), we prompt LLMs to solve the task based solely on the generated knowledge from Step 1 (described in Section 3.1). This allows us to compare the model’s capability in solving tasks using only the identified relevant knowledge versus leveraging this knowledge to structure a coherent narrative.

Ours: Story of Thought (SoT): This approach represents the core of our proposed method, where we leverage the generated narratives from Step 2 (described in Section 3.2) to solve the given tasks.

5 Results

Our proposed SoT approach that incorporates narrative structures improves over almost all previous prompting approaches across two different problem-solving datasets. This highlights the potential of using narratives to improve the ability of LLMs to understand and reason about the given information in various intensive reasoning tasks.

5.1 Performance on GPQA

Results on GPQA (Diamond set) are presented in Table 1. For this task, SoT is the best method to use with six of eight models. The open-source Llama 3 70B records the highest accuracy using the SoT

method, achieving a score of 51.01%. This is the highest accuracy observed among all models and methods tested in the study. Furthermore, the GPT-4 model shows the most notable improvement in accuracy with SoT, compared to the zero-shot baseline. Specifically, the accuracy for GPT-4 increased from 34.7% under zero-shot conditions to 48.98% with SoT (i.e., an absolute increase of 14.28%, or a relative increase of 41% respectively).¹ Interestingly, all reasoning strategies lead to an accuracy drop for the Phi-3 Mini model, and all CoT strategies except Analogical Reasoning also lead to the accuracy drop of the Phi-3 Medium model compared to its zero-shot baseline. We hypothesize that this is due to the low quality of the generated explanations and study it further in §6.1. We note that, on average, models improve the most on biology problems when using SoT. See Appendix D for subject-wise performance evaluation.

5.2 Performance on JEEBench

Table 2 presents detailed experimental results on JEEBench. Our proposed method (SoT) consistently improves the performance of seven out of the eight LLMs. Using SoT, Llama 3 70B performance surpasses even the GPT models. It obtains the highest scores in all subjects and question types (Except Single-Correct), with an overall aggregate score of 0.453. This is a significant improvement on the previous SotA, which was a strong GPT-4 model used with both CoT and Self-Consistency. Across models, the results highlight the effectiveness of Story of Thought (SoT) in enhancing model performance on complex, multi-disciplinary benchmarks like JEEBench, setting new SotA results in several categories. The improvements are particularly notable in the subject categories and question types where the other methods struggle.

¹We also find that Llama 3 70B with SoT outperforms zero-shot o1-preview which uses CoT style reasoning internally.

	Chemistry	Mathematics	Physics	Integer	Single-Correct	Multi-Correct	Numeric	Total
GPT-4+CoT+SC@8*	0.463	0.308	0.449	0.293	0.618	0.410	0.234	0.389
Llama 3 8B	0.143	0.082	0.089	0.061	0.127	0.148	0.044	0.102
Llama 3 8B+CoT	0.127	0.101	0.116	0.11	0.145	0.149	0.036	0.112
Ours: Llama 3 8B+SoT	0.154	0.195	0.172	0.072	0.259	0.324	0.028	0.173
Llama 3 70B	0.324	0.189	0.274	0.171	0.345	0.316	0.131	0.25
Llama 3 70B+CoT	0.264	0.228	0.268	0.159	0.291	0.317	0.175	0.249
Ours: Llama 3 70B+SoT	0.554	0.329	0.471	0.446	0.42	0.485	0.462	0.453
Mistral 7B	0.119	0.079	0.091	0.049	0.109	0.159	0.022	0.094
Mistral 7B+CoT	0.106	0.123	0.059	0.073	0.118	0.165	0.022	0.102
Ours: Mistral 7B+SoT	0.2	0.177	0.201	0.11	0.245	0.224	0.146	0.19
Mixtral 8x7B	0.22	0.151	0.167	0.122	0.218	0.261	0.058	0.176
Mixtral 8x7B+CoT	0.237	0.142	0.152	0.061	0.209	0.27	0.08	0.173
Ours: Mixtral 8x7B+SoT	0.253	0.251	0.274	0.268	0.309	0.277	0.182	0.257
ChatGPT 3.5	0.228	0.146	0.173	0.073	0.318	0.249	0.029	0.177
ChatGPT 3.5+CoT	0.17	0.111	0.167	0.11	0.173	0.206	0.051	0.142
Ours: ChatGPT 3.5+SoT	0.189	0.128	0.189	0.073	0.291	0.204	0.051	0.161
GPT 4	0.423	0.212	0.352	0.207	0.455	0.383	0.153	0.309
GPT 4+CoT	0.468	0.280	0.335	0.256	0.473	0.448	0.175	0.350
Ours: GPT 4+SoT	0.535	0.294	0.413	0.378	0.4	0.453	0.321	0.395
Phi-3 Mini	0.256	0.12	0.199	0.146	0.255	0.224	0.08	0.18
Phi-3 Mini+CoT	0.256	0.137	0.171	0.134	0.209	0.216	0.139	0.181
Ours: Phi-3 Mini+SoT	0.224	0.209	0.181	0.183	0.282	0.234	0.124	0.207
Phi-3 Medium	0.298	0.193	0.165	0.146	0.255	0.286	0.139	0.218
Phi-3 Medium+CoT	0.253	0.195	0.199	0.171	0.236	0.274	0.139	0.214
Ours: Phi-3 Medium+SoT	0.279	0.203	0.224	0.232	0.273	0.263	0.153	0.231

Table 2: On JEEBench, Story of Thought (SoT) outperforms previous SOTA as well as other methods. We present the aggregate score by subject as well as question type and present the overall aggregate score. The best overall scores are highlighted in **blue** while the best score by method for a model is in **bold**. * reported in (Arora et al., 2023).

6 Analysis of SoT Aspects

6.1 Role of the Narrative Quality/Choice

The choice of *narrator* model (i.e., the model that generates narratives) can impact the problem-solving results. In the following experiments, we apply the narratives generated by other large and small open-source LLMs to the Phi-3 Mini and Phi-3 Medium models. The results of these experiments are presented in Table 3.

Narrative Generator	Solver Models	
	Phi-3 Mini	Phi-3 Medium
Llama 3 8B	23.74 (+1.01↑)	37.88 (+1.28↑)
Llama 3 70B	25.25 (+2.52↑)	39.39 (+2.79↑)
Mistral 7B	24.24 (+1.51↑)	38.38 (+1.78↑)
Mixtral 8x7B	24.74 (+2.01↑)	35.86 (-0.74↓)

Table 3: Applying generated narratives by open-source models to Microsoft models to solve the tasks.

We observe that the **narratives** generated by most models **consistently improve the accuracy** of both Microsoft models compared to the baseline (i.e., when both models use their own generated narratives in Step 2 to solve the tasks, shown in

Table 1). The absolute improvements range from 1.0% to 2.8%, with the Llama 3 70B model generating the most effective narratives. A slight decrease in accuracy is observed with the mixture-of-experts Mixtral 8x7B narratives for the Phi-3 Medium model, highlighting the need for careful selection and evaluation of narrator models to ensure compatibility and optimal performance. Larger models generate narratives that break down problems to make them more easily solvable. Unsurprisingly, there is larger room for improving the problem solving abilities of smaller models.

6.2 Impact of Narrative Techniques

To measure the impact of each narrative technique, we jointly prompted on the performance of open-source Meta models, we ablate the designed prompt in Step 2 (of Section 3.2) to apply each of the techniques separately. The results in Table 5 indicate that **employing any single narrative elements at a time is notably less effective at boosting QA accuracy than utilizing a combination of these simultaneously**. For both Llama models, the decrease in accuracy is comparably smaller (-3.0% to -5.6%) when using only the analogical com-

Narrative Technique	Meta		Mistral		OpenAI		Microsoft	
	Llama 3 8B	Llama 3 70B	Mistral 7B	Mistral 8x7B	ChatGPT 3.5	GPT 4	Phi-3 Mini	Phi-3 Medium
Progressive Disclosure	427	597	191	191	744	570	367	368
Branching	30	56	51	20	72	168	34	61
Analogy	418	425	117	161	498	595	569	499
Analogical Reasoning	205	191	78	108	213	336	276	206
Metaphor	249	316	103	137	811	428	418	291
Σ	1329	1585	540	617	2338	2097	1664	1425

Table 4: Comparing Generated Narratives - Total Number of Occurrences for each Narrative Technique (Evaluator: Llama 3 70B)

Narrative Technique	Meta	
	Llama 3 8B	Llama 3 70B
Progressive Disclosure	34.85 (-8.58↓)	44.95 (-6.06↓)
Branching	34.34 (-9.09↓)	44.95 (-6.06↓)
Analogy	39.39 (-4.04↓)	46.46 (-4.55↓)
Analogical Reasoning	40.4 (-3.03↓)	45.45 (-5.56↓)
Metaphor	41.41 (-2.02↓)	44.44 (-6.57↓)
None	38.38 (-5.05↓)	45.45 (-5.56↓)
All	43.43	51.01

Table 5: Comparing accuracy when using a single narrative technique or no narrative technique (None). Values in parentheses represent the decrease in accuracy percentage points compared to a combination of multiple narrative techniques simultaneously (shown in Table 1).

ponents of the narrative (*Analogy* and *Analogical Reasoning*) than when using only the structural instructions (*Progressive Disclosure* or *Branching*) which leads to larger (-6.0% to -9.1%) accuracy loss. However, reasoning alone does not perform on par with the full narrative generation listing all the techniques. Prompting for *Metaphor* usage only leads to a larger accuracy loss in the 70B model (-6.6%) compared to the smaller one (-2.0%). The *None* condition, where no narrative technique is mentioned in the prompt, results in an accuracy drop (-5.0% to -5.6%). This makes it difficult to determine how the narrative techniques relate to each other. We study this going forward.

6.3 Analyzing Generated Narratives

To gain deeper insights into the generated narratives, we prompt Llama 3 70B to annotate the number of times each narrative technique appears (i.e., the number of occurrences) in each generated narrative across all models used in our experiments. We can better interpret how the model executed the narrative generation prompt, by asking it to label if and where the mentioned techniques are used in the generated narrative. A proportion of the narrative techniques and their correlation can provide us

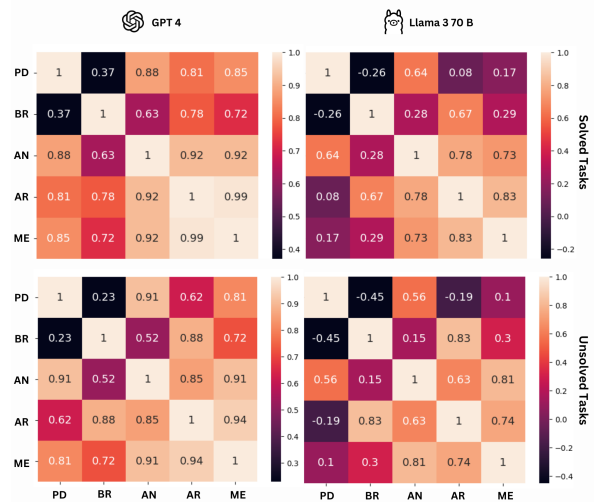


Figure 3: Correlation coefficients among all narrative elements (PD = Progressive Disclosure, BR = Branching, AN = Analogy, AR = Analogical Reasoning, ME = Metaphor) used in the SoT approach for GPT 4 and Llama 3 70B in solved and unsolved tasks.

with a better picture of LLM’s interpretation of the instruction as well. The instructions can be found in [Appendix C](#). We aim to uncover patterns and variations in the use of narrative techniques across different LLMs. [Table 4](#) compares the total number of occurrences for each narrative technique across various LLMs.

Variability in Utilization of Narrative Techniques Across Models:

In our designed prompt in Step 2 (i.e., Narrative Generation), LLMs generate narrative using all 5 narrative techniques. However, as [Table 4](#) indicates, not all elements were employed equally. This reveals that while some techniques like *Analogy* and *Progressive Disclosure* were consistently utilized, others such as *Branching* were applied less frequently. We observe a trend across all LLM families where models with larger capacities, such as Llama 3 70B and GPT-4, consistently show higher occurrences of narrative tech-

Similarity Metric	BertScore		ROUGE-L		BLEU	
	SoT Reasoning	CoT Reasoning	SoT Reasoning	CoT Reasoning	SoT Reasoning	CoT Reasoning
Llama 3 8B	0.28	0.06	0.19	0.11	6.57	0.19
Llama 3 70B	0.3	0.04	0.2	0.1	8.18	0.06
Mistral 7B	0.27	0.33	0.18	0.2	8.12	4.65
Mixtral 8x7B	0.3	0.34	0.19	0.21	8.92	8.14
ChatGPT 3.5	0.3	0.24	0.19	0.16	6.1	6.07
GPT 4	0.31	0.34	0.19	0.2	8.84	6.73
Phi-3 Mini	0.27	0.31	0.17	0.19	6.54	6.36
Phi-3 Medium	0.3	0.35	0.2	0.21	7.13	8.4

Table 6: Comparison of generated Story of Thought (SoT) and Chain of Thought (CoT) reasoning with Human Explanations on the GPQA (Diamond set) using BERTScore, ROUGE-L, and BLEU metrics across various large language models. Bold values indicate the reasoning approach that is more similar to human explanations for each model and metric pair.

niques compared to their smaller counterparts. Furthermore, ChatGPT 3.5 & GPT-4 demonstrate the highest total occurrences of narrative techniques, with 2,338 and 2,097, respectively with a notable emphasis on *Metaphors* and *Analogies*.

Correlation Among Narrative Techniques: To further investigate the dynamics of narrative techniques, we compute correlations between the frequencies of narrative techniques across solved and unsolved tasks, as shown in Figure 3. This analysis aims to uncover if the models consistently use certain narrative techniques together or vary significantly. Our initial results indicate diverse correlation patterns, suggesting that the effectiveness of narrative techniques in solving tasks across various LLMs needs to be further analyzed.

6.4 Human Evaluation

To assess the reliability of the LLM-based annotation method (described in Section 6.3), we conduct a human evaluation of narrative techniques used by LLMs in generated narratives. We provide 3 annotators with 15 narratives generated by 8 different models, resulting in a total of 120 narratives. Annotators were instructed to identify and count how many times each narrative technique appeared in each narrative. The aggregated annotations were then analyzed using the *Krippendorff Alpha Coefficient* to assess inter-annotator agreement.

The average agreement score across all techniques was 0.72, indicating **strong inter-annotator agreement, with Llama 3 70B aligning closely with human annotators, validating the use of LLM-based evaluation for assessing narrative techniques**. While annotators show the highest agreement in *Branching* and *Analogy*, with average scores of 0.75 and 0.79, respectively, they have in

Metaphor and *Analogical Reasoning* lower agreement scores (0.69 and 0.68). The extended results can be found in Appendix F.

6.5 Analyzing SoT Reasoning

Table 6 compares the similarity of SoT and CoT reasoning outputs to human explanations for different language models on the GPQA (Diamond set) dataset, using BertScore, ROUGE-L, and BLEU.

The differences between ROUGE-L values are insignificant and do not display any clear trends. However, according to BLEU scores, using SoT results in explanations closer to humans and the differences are more pronounced. As per BertScore Llama 3 models’ explanations are more similar to human ones when using SoT reasoning across all three metrics. However, Mistral models, GPT-4, and Phi-3 Mini generate explanations more similar to human explanations when using CoT reasoning across all metrics. The semantic similarity of narratives generated by Llama 3 70B to human explanations combined with their effect of improving smaller models indicates that these narratives present information about the problems in a simplified manner.

7 Conclusion

Inspired by findings from human cognitive processes explored in didactics research, in this work, we propose to use narratives in LLMs prompting. We present strong evidence on public benchmark datasets that narratives have the potential to notably enhance the reasoning abilities of LLMs in complex problem-solving tasks. By incorporating narrative structures, which mimic human cognitive processes of organizing and interpreting information, LLMs can achieve higher levels of performance and provide more contextually enriched responses.

Limitations

Dataset limitations. So far, we used only GPQA and JEEBench problems as the most challenging set of problem-solving benchmarks we were aware of. Other comparable benchmarks, such as MGSM, are much closer to human or superhuman accuracy already without reasoning prompts and will be explored in future work.

Analysis limitations. The occurrences of narrative techniques do not necessarily imply the quality or effectiveness of the generated narratives; rather, they provide insights into the models' tendencies and preferences in employing these techniques. Therefore, answering the question of why narrative is helping LLMs is more complex and needs to be further investigated by looking into different research areas such as cognitive and communication theories.

References

- H Porter Abbott. 2020. *The Cambridge introduction to narrative*. Cambridge University Press.
- Paul B Armstrong. 2020. *Stories and the brain: The neuroscience of narrative*. Johns Hopkins University Press.
- Daman Arora, Himanshu Singh, and Mausam. 2023. [Have LLMs advanced enough? a challenging problem solving benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. [Graph of Thoughts: Solving Elaborate Problems with Large Language Models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.
- Gordon H Bower and Michal C Clark. 1969. Narrative stories as mediators for serial learning. *Psychonomic science*, 14(4):181–182.
- Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry*, 18(1):1–21.
- Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332.
- Althea Y Chen, Chun-Ching Chen, and Wen-Yin Chen. 2023. The design narrative in design learning: Adjusting the inertia of attention and enhancing design integrity. *The Design Journal*, 26(4):519–535.
- Ming Ming Chiu. 2000. Metaphorical reasoning: Origins, uses, development, and interactions in mathematics. *EDUCATION JOURNAL-HONG KONG-CHINESE UNIVERSITY OF HONG KONG-*, 28(1):13–46.
- 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*.
- Michael F Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the national academy of sciences*, 111(supplement_4):13614–13620.
- Alison Engel, Kathryn Lucido, and Kyla Cook. 2018. Rethinking narrative: Leveraging storytelling for science learning. *Childhood Education*, 94(6):4–12.
- Walter R Fisher. 2021. *Human communication as narration: Toward a philosophy of reason, value, and action*. University of South Carolina Press.
- Dedre Gentner and Linsey A Smith. 2013. Analogical learning and reasoning.
- Jonathan Gottschall. 2012. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.
- Philip N Hine. 2018. Narrative: Why it's important, and how it works. *Perspectives on Behavior Science*, 41:471–501.
- Keith J Holyoak and Hongjing Lu. 2021. Emergence of relational reasoning. *Current Opinion in Behavioral Sciences*, 37:118–124.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. [Looking beyond the surface: A challenge set for reading comprehension over multiple sentences](#). 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Tomáš Kočický, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. [The narrativeqa reading comprehension challenge](#). *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjana Balasubramanian. 2021. [TellMeWhy: A dataset for answering why-questions in narratives](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

- Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. [Making language models better reasoners with step-aware verifier](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. [Program induction by rationale generation: Learning to solve and explain algebraic word problems](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Susana Martinez-Conde and Stephen L Macknik. 2017. Finding the plot in science storytelling in hopes of enhancing science communication. *Proceedings of the National Academy of Sciences*, 114(31):8127–8129.
- Areej Mawasi, Peter Nagy, and Ruth Wylie. 2020. Systematic literature review on narrative-based learning in educational technology learning environments (2007-2017).
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Sewon Min, Xinxu 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 *EMNLP*.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- Aquiles Negrete and Cecilia Lartigue. 2004. Learning from education to communicate science as a good story. *Endeavour*, 28(3):120–124.
- Stephen P Norris, Sandra M Guilbert, Martha L Smith, Shahram Hakimelahi, and Linda M Phillips. 2005. A theoretical framework for narrative explanation in science. *Science education*, 89(4):535–563.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. [Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with language model prompting: A survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof q&a benchmark](#). In *First Conference on Language Modeling*.
- Vahid Sadiri Javadi, Johanne R Trippas, and Lucie Flek. 2024. [Unveiling information through narrative in conversational information seeking](#). In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA. Association for Computing Machinery.
- Anthony J Sanford and Catherine Emmott. 2012. *Mind, brain and narrative*. Cambridge University Press.
- John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.
- Joanna Szurmak and Mindy Thuna. 2013. Tell me a story: The use of narrative as tool for instruction.
- Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. [Self-consistency improves chain of thought reasoning in language models](#). *ArXiv*, abs/2203.11171.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [Constructing datasets for multi-hop reading comprehension across documents](#). *Transactions of the Association for Computational Linguistics*, 6:287–302.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Daniel T Willingham. 2004. Ask the cognitive scientist the privileged status of story. *American Educator*, 28:43–45.
- Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2024. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *arXiv preprint arXiv:2404.15676*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of Thoughts: Deliberate problem solving with large language models](#). *Preprint*, arXiv:2305.10601.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.
- Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.
- Paul J Zak. 2015. Why inspiring stories make us react: The neuroscience of narrative. In *Cerebrum: the Dana forum on brain science*, volume 2015. Dana Foundation.

A Robustness of LLM Predictions

In the original GPQA dataset used for our experiments, the correct answers are always presented as the first option among the multiple choices. However, To further evaluate the robustness of the LLMs, we conduct an additional experiment where the correct answers are placed in the second option instead. Table 7 presents the results of these experiments, comparing the performance of various prompting methods across six different open-source LLMs. We observe that most LLMs experience a significant drop in accuracy when the correct answer is moved to the second option. However, despite the overall decrease in accuracy, our proposed approach, Story of Thought (SoT), consistently outperforms the baseline methods for most LLMs. The SoT method achieves the highest accuracy for the Meta Llama 3 8B, Meta Llama 3 70B, Mistral 8x7B, and Microsoft Phi-3 Medium models, demonstrating its effectiveness in enhancing the robustness of LLMs to changes in the problem structure.

B Model Implementation Details

All experiments, except for those involving OpenAI models, were conducted on local machines equipped with GPUs. The models were run locally on a GPU setup without quantization using the *Hugging Face Transformer* library². For OpenAI’s GPT-3.5-turbo (*gpt-3.5-turbo-0125*) and GPT-4-turbo (*gpt-4-turbo-2024-04-09*) models, we use the OpenAI API to generate outputs. Across all models, the results are averages over 5 runs with a temperature of 1.0 and a maximum number of tokens of 8,000. The other parameters are set to their default values. To ensure consistency in the model outputs, we utilized the *Jsonformer* Python package³, resulting in structured JSON outputs. A t-test was performed, yielding a p-value of 0.032, indicating statistical significance at the conventional 0.05 level.

C Prompts Used in Story of Thought

We describe the prompts used for each stage in the SoT framework.

²<https://huggingface.co/docs/transformers>

³<https://github.com/1rgs/jsonformer>

C.1 Question Clarification

You are an explorer who wants to identify and collect different related and specialized subject areas to clarify the question. Your goal is to narrow down the question and provide relevant areas of knowledge and experience you have that help clarify the question mentioned below. You should not answer the question.

<question>

C.2 Narrative Generation

You are an expert in narrative-based explanations for science communication. Your goal is to clarify the following question in a narrative way through the interconnected information provided below to enable a non-expert to comprehend the question in a more coherent and contextually rich manner. You should not answer the question.

Make sure to use all of these narrative techniques when clarifying the question through the interconnected information: Progressive Disclosure, Branching, Analogy, Analogical Reasoning, and Metaphor.

<question>

<generated information in the previous step>

C.3 Problem Solving

You are an expert in analyzing narrative-based explanations for solving tasks. Please answer the following question based on the following narrative-based clarification:

<question>

Options:
<options>

<generated narrative in the previous step>

C.4 Analyzing Generated Narratives

You are an expert in analyzing narrative-based explanations for science communication. Your goal is to find out which narrative techniques have been used in the following narrative-based explanation.

Label the narrative-based explanation using the following narrative-based techniques:

1. Progressive Disclosure
2. Branching
3. Analogy
4. Analogical Reasoning
5. Metaphor

<generated narrative>

Prompting Method	Meta		Mistral		Microsoft	
	Llama 3 8B	Llama 3 70B	Mistral 7B	Mixtral 8x7B	Phi-3 Mini	Phi-3 Medium
Zero-shot	30.81 (-3.39↓)	31.31 (-8.19↓)	19.7 (-16.1↓)	18.18 (-18.18↓)	29.8 (+1.01↑)	21.72 (-20.7↓)
Zero-shot CoT	27.27 (-13.64↓)	33.33 (-8.59↓)	22.73 (-9.09↓)	17.17 (-18.18↓)	32.32 (+7.57↑)	21.21 (-18.18↓)
Analogical Reasoning	27.78 (-13.13↓)	40.91 (-6.56↓)	10.61 (-27.29↓)	19.19 (-7.07↓)	35.86 (+19.19↑)	16.67 (-31.81↓)
Ours: Knowledge Identification	32.32 (-8.08↓)	42.4 (-6.59↓)	16.67 (-18.68↓)	14.65 (-23.12↓)	28.28 (+7.57↑)	23.26 (-14.62↓)
Ours: Story of Thought (SoT)	34.85 (-8.58↓)	45.4 (-5.61↓)	20.2 (-18.2↓)	20.2 (-18.69↓)	27.7 (+4.97↑)	25.75 (-10.85↓)

Table 7: Performance of various LLMs across different prompting methods on GPQA (Diamond set). Correct answers are presented in the second option. Values in parentheses indicate the change in accuracy compared to the original setting in Table 1 where the correct answer was in the first option.

D Subject-wise Performance Evaluation

Figure 4 presents the subject-wise performance of different models on both GPQA and JEEBench when using SoT across the different problem domains. We observe that, on average, models improve the most on biology problems when using SoT in GPQA. We hypothesize that this is because it is easier to simplify information for graduate-level biology problems that can be used by models to come up with a solution.

In JEEBench, on average, model performance is highest on Chemistry problems when using SoT. This is in contrast to findings on GPQA and could occur due to the difference in the degree of difficulty of problems in the two datasets (graduate level vs high school level). Regardless, improvements on Biology problems are not far behind those for Chemistry.

E Performance on JEEBench

F Huamn Evaluation

Table 9 presents the Krippendorff Alpha coefficient measuring inter-annotator agreement between three human annotators and Llama 3 70B across five narrative techniques: Progressive Disclosure (PD), Branching (BR), Analogy (AN), Analogical Reasoning (AR), and Metaphor (ME). Higher values indicate stronger agreement. The overall average agreement of 0.72 shows a strong correlation between LLM-based and human annotations, supporting the validity of the LLM-based evaluation method.

Annotator Recruitment and Demographics:

We recruited three master’s students in computer science, aged between 24 and 27 (one female, two males). The annotators were compensated fairly for their time, ensuring alignment with appropriate compensation standards. Each annotator was provided with 120 narratives and given detailed

instructions, including the definitions of each narrative technique as described in Section 3.2, to ensure a consistent evaluation process. Each annotation was performed independently to minimize bias and ensure reliability.

G Story of Thought (SoT) vs. Chain of Thought (CoT)

	Chemistry	Mathematics	Physics	Integer	Single-Correct	Multi-Correct	Numeric	Total
GPT-4+CoT+SC@8*	0.463	0.308	0.449	0.293	0.618	0.410	0.234	0.389
Llama 3 8B	0.143	0.082	0.089	0.061	0.127	0.148	0.044	0.102
Llama 3 8B+CoT	0.127	0.101	0.116	0.11	0.145	0.149	0.036	0.112
Llama 3 8B+Analogical Reasoning (3-shot)	0.139	0.111	0.128	0.11	0.145	0.165	0.058	0.124
Ours: Llama 3 8B+Knowledge Identification	0.199	0.099	0.134	0.073	0.227	0.171	0.058	0.137
Ours: Llama 3 8B+SoT	0.154	0.195	0.172	0.072	0.259	0.324	0.028	0.173
Llama 3 70B	0.324	0.189	0.274	0.171	0.345	0.316	0.131	0.25
Llama 3 70B+CoT	0.264	0.228	0.268	0.159	0.291	0.317	0.175	0.249
Llama 3 70B+Analogical Reasoning (3-shot)	0.314	0.24	0.295	0.195	0.318	0.349	0.19	0.276
Ours: Llama 3 70B+Knowledge Identification	0.317	0.226	0.254	0.195	0.345	0.323	0.146	0.26
Ours: Llama 3 70B+SoT	0.554	0.329	0.471	0.446	0.42	0.485	0.462	0.453
Mistral 7B	0.119	0.079	0.091	0.049	0.109	0.159	0.022	0.094
Mistral 7B+CoT	0.106	0.123	0.059	0.073	0.118	0.165	0.022	0.102
Mistral 7B+Analogical Reasoning (3-shot)	0.157	0.084	0.116	0.073	0.155	0.169	0.029	0.114
Ours: Mistral 7B+Knowledge Identification	0.109	0.055	0.063	0.037	0.091	0.117	0.022	0.073
Ours: Mistral 7B+SoT	0.2	0.177	0.201	0.11	0.245	0.224	0.146	0.19
Mixtral 8x7B	0.22	0.151	0.167	0.122	0.218	0.261	0.058	0.176
Mixtral 8x7B+CoT	0.237	0.142	0.152	0.061	0.209	0.27	0.08	0.173
Mixtral 8x7B+Analogical Reasoning (3-shot)	0.202	0.155	0.197	0.122	0.191	0.281	0.066	0.179
Ours: Mixtral 8x7B+Knowledge Identification	0.184	0.129	0.144	0.122	0.155	0.237	0.044	0.15
Ours: Mixtral 8x7B+SoT	0.253	0.251	0.274	0.268	0.309	0.277	0.182	0.257
ChatGPT 3.5	0.228	0.146	0.173	0.073	0.318	0.249	0.029	0.177
ChatGPT 3.5+CoT	0.17	0.111	0.167	0.11	0.173	0.206	0.051	0.142
ChatGPT 3.5+Analogical Reasoning (3-shot)	0.208	0.125	0.148	0.098	0.2	0.216	0.073	0.156
Ours: ChatGPT 3.5+Knowledge Identification	0.155	0.141	0.167	0.122	0.209	0.188	0.073	0.151
Ours: ChatGPT 3.5+SoT	0.189	0.128	0.189	0.073	0.291	0.204	0.051	0.161
GPT 4	0.423	0.212	0.352	0.207	0.455	0.383	0.153	0.309
GPT 4+CoT	0.468	0.280	0.335	0.256	0.473	0.448	0.175	0.350
GPT 4+Analogical Reasoning (3-shot)	0.479	0.286	0.396	0.305	0.4	0.43	0.307	0.371
Ours: GPT 4+Knowledge Identification	0.481	0.287	0.386	0.293	0.373	0.452	0.314	0.373
Ours: GPT 4+SoT	0.535	0.294	0.413	0.378	0.4	0.453	0.321	0.395
Phi-3 Mini	0.256	0.12	0.199	0.146	0.255	0.224	0.08	0.18
Phi-3 Mini+CoT	0.256	0.137	0.171	0.134	0.209	0.216	0.139	0.181
Phi-3 Mini+Analogical Reasoning (3-shot)	0.205	0.159	0.195	0.146	0.264	0.218	0.088	0.182
Ours: Phi-3 Mini+Knowledge Identification	0.168	0.091	0.106	0.073	0.136	0.181	0.044	0.118
Ours: Phi-3 Mini+SoT	0.224	0.209	0.181	0.183	0.282	0.234	0.124	0.207
Phi-3 Medium	0.298	0.193	0.165	0.146	0.255	0.286	0.139	0.218
Phi-3 Medium+CoT	0.253	0.195	0.199	0.171	0.236	0.274	0.139	0.214
Phi-3 Medium+Analogical Reasoning (3-shot)	0.258	0.181	0.173	0.159	0.218	0.276	0.117	0.202
Ours: Phi-3 Medium+Knowledge Identification	0.288	0.163	0.205	0.207	0.236	0.235	0.161	0.211
Ours: Phi-3 Medium+SoT	0.279	0.203	0.224	0.232	0.273	0.263	0.153	0.231

Table 8: On JEEBench, Story of Thought (SoT) outperforms previous SOTA as well as other methods. We present the aggregate score by subject as well as question type and present the overall aggregate score. * denotes SOTA results taken from the original paper (Arora et al., 2023).

Model Name	Narrative Technique				
	Progressive Disclosure (PD)	Branching (BR)	Analogy (AN)	Analogical Reasoning (AR)	Metaphor (ME)
Llama 3 8B	0.69	0.84	0.68	0.68	0.93
Llama 3 70B	0.77	0.68	0.82	0.61	0.73
Mistral 7B	0.65	0.76	0.97	0.65	0.62
Mixtral 8x7B	0.67	0.78	0.74	0.82	0.76
ChatGPT 3.5	0.6	0.69	0.67	0.68	0.69
GPT 4	0.64	0.81	0.82	0.8	0.6
Phi-3 Mini	0.69	0.73	0.79	0.62	0.61
Phi-3 Medium	0.66	0.69	0.79	0.61	0.61
Average	0.67	0.75	0.79	0.68	0.69

Table 9: Krippendorff Alpha Coefficient for Human and LLM Annotations.

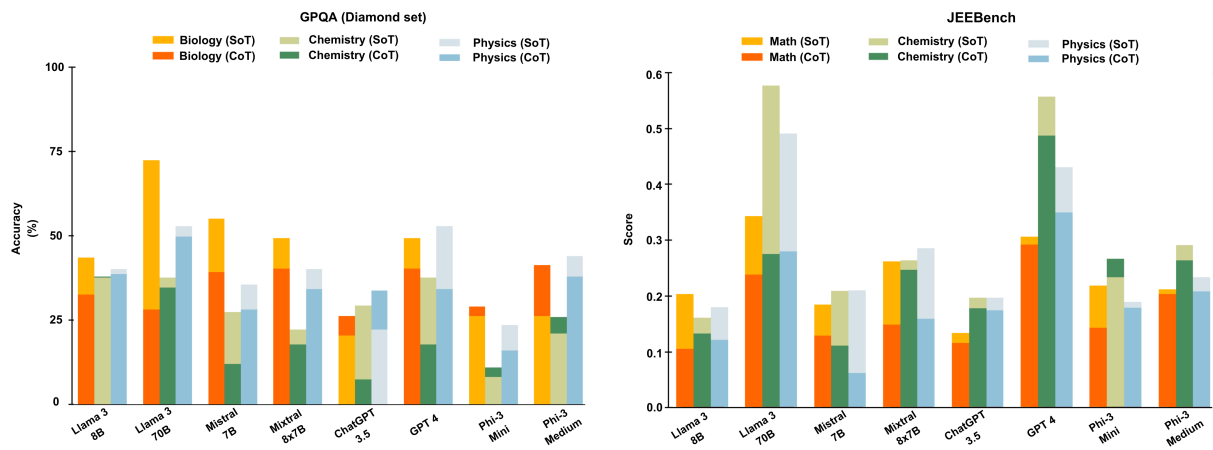


Figure 4: Performance of **Story of Thought** (SoT) on GPQA and JEEBench across various LLMs and domains.

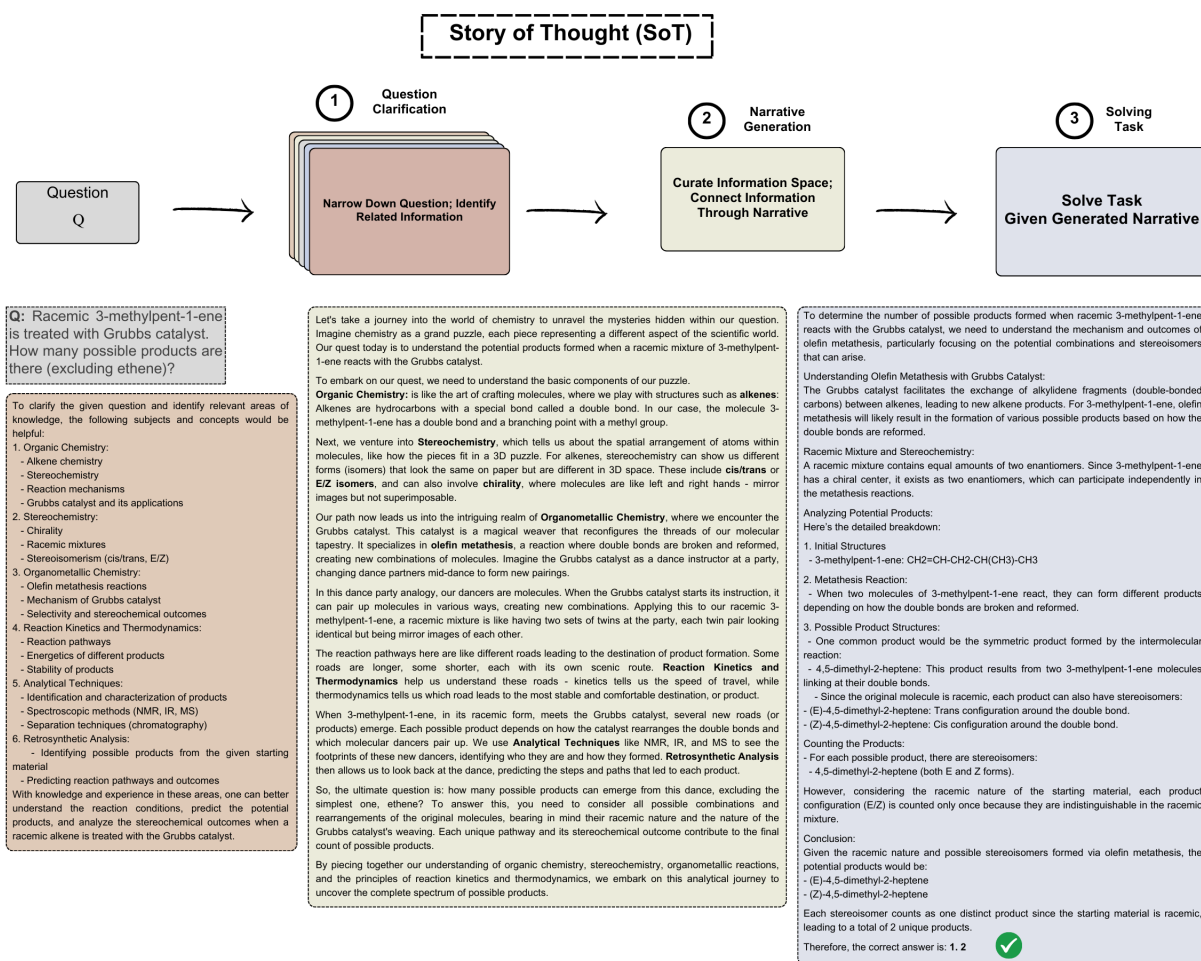


Figure 5: An actual example of SoT.

Testing Spatial Intuitions of Humans and Large Language and Multimodal Models in Analogies

Ivo Bueno

Technical University of Munich
ivo.bueno@tum.de

João Miguel Cunha
University of Coimbra,
CISUC/LASI, DEI
jmacunha@dei.uc.pt

Anna Bavaresco

University of Amsterdam
a.bavaresco@uva.nl

Philipp Wicke
LMU, Munich
Munich Center for Machine Learning (MCML)
philipp.wicke@lmu.de

Abstract

Language and Vision-Language Models exhibit impressive language capabilities akin to human reasoning. However, unlike humans who acquire language through embodied, interactive experiences, these models learn from static datasets without real-world interaction. This difference raises questions about how they conceptualize abstract notions and whether their reasoning aligns with human cognition. We investigate spatial conceptualizations of LLMs and VLMs by conducting analogy prompting studies with LLMs, VLMs, and human participants. We assess their ability to generate and interpret analogies for spatial concepts. We quantitatively compare the analogies produced by each group, examining the impact of multimodal inputs and reasoning mechanisms. Our findings indicate that generative models can produce and interpret analogies but differ significantly from human reasoning in their abstraction of spatial concepts – variability is influenced by input modality, model size, and prompting methods, with analogy-based prompts not consistently enhancing alignment. Contributions include a methodology for probing generative models through analogies, a comparative analysis of analogical reasoning among models, and humans, and insights into the effect of multimodal inputs on reasoning.¹

1 Introduction

Large language models (LLMs) have revolutionized natural language processing, achieving remarkable language proficiency and emergent abilities that seem to parallel human reasoning (Brown et al., 2020; Achiam et al., 2023; Kojima et al., 2022). Trained on vast corpora of text – or paired text and images for vision-language models (VLMs) – these models’ learning paradigms fundamentally differ from human language acquisition,

¹Code available at:
https://github.com/cisnlp/spatial_intuitions

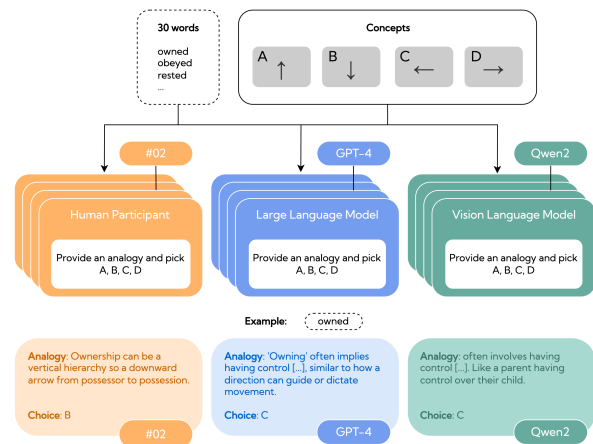


Figure 1: Human participants (e.g., participant #02), LLMs (e.g., GPT-4) and VLMs (e.g., Qwen2-VL) are prompted to provide an analogy for their choice of 1 of 4 items (\uparrow , \downarrow , \leftarrow , \rightarrow) that best represents 1 of 30 words.

raising questions about how they represent meaning, form abstract ideas, and structure knowledge.

LLMs learn from static, digital artifacts, processing accumulated language data without real-time interaction or sensory experience. Their training spans weeks to months using massive computational resources (Hoffmann et al., 2022; Scao et al., 2023). In contrast, human language acquisition is an embodied process: children learn through dynamic interactions with their environment – observing, testing, and experiencing the world around them (Mandler, 1992). First words emerge around 12 months, alongside nonverbal communication (Bretherton and Bates, 1979; Iverson, 2010), and foundational language abilities develop over approximately five years, with sensory experiences and social interactions playing crucial roles (Clark and Casillas, 2015).

Despite these differences, both LLMs and humans produce language artifacts and exhibit reasoning grounded in language. This raises a fundamental question: **How can LLMs exhibit rea-**

soning abilities seemingly analogous to human cognition when their training procedures are so fundamentally different? Addressing this is crucial as we integrate LLMs/VLMs into systems where reasoning and understanding are essential.

Studies have highlighted limitations in LLM reasoning capabilities – they often struggle with complex reasoning tasks (Mondorf and Plank, 2024; Stechly et al., 2023), arithmetic operations (Gambardella et al., 2024), planning (Valmeekam et al., 2022), and other challenges (Sobieszek and Price, 2022). One potential issue is how LLMs abstract from their knowledge. It is argued that human cognition largely relies on analogical reasoning, i.e., understanding abstract concepts by relating them to familiar ones (Gentner, 1983; Hofstadter, 2001). Analogies facilitate learning and are a crucial component for human cognitive development (Vosniadou and Ortony, 1989; Holyoak, 2012).

We focus on analogical reasoning to investigate whether LLMs and VLMs can generate and interpret analogies like humans to understand abstract spatial associations. Specifically, we address: **(RQ1)**: How do LLMs and VLMs conceptualize semantic notions through spatial analogies compared to humans? **(RQ2)**: How do multimodal inputs (e.g., text and images) affect the models’ analogical reasoning? To answer these questions, we conduct analogy prompting studies (i.e., requiring to produce an analogy to answer a question) with LLMs, VLMs, and human participants. We systematically categorize and compare the analogies generated by each group. Our experiments examine the influence of different modalities, testing state-of-the-art VLMs with image inputs to assess how sensory information impacts reasoning outcomes. Our contributions are:

1. Methodology for probing conceptualization in models through analogy generation;
2. Comparative analysis of analogical reasoning abilities of LLMs, VLMs and humans, using both quantitative and qualitative approaches;
3. Insights into how multimodal inputs influence models toward human-like reasoning;
4. Evaluation of whether different types of models, e.g., those with enhanced reasoning, improve analogy and conceptual understanding.

2 Related Work

2.1 Analogical Reasoning in Cognition

Analogical reasoning is a key cognitive strategy which allows individuals to draw parallels between disparate domains by mapping relational structures. Gentner’s structure-mapping theory posits that analogy involves aligning relational structures from a base domain to a target domain, emphasizing the importance of systematic correspondences over mere attribute similarities. Gust et al. (2008) argue that analogies underpin key cognitive abilities – memory adaptation, transfer learning, reasoning, and creativity – by enabling the application of prior knowledge to novel contexts; they propose that analogical reasoning is fundamental for integrating diverse cognitive processes in large-scale systems. Evidence for the connection between human reasoning and analogies comes from several psycholinguistic studies (Richardson et al., 2001; Beitel et al., 2001; Gibbs et al., 1994). They provide evidence that certain linguistic representations are grounded in spatial schemas, which operate as analogical structures for language comprehension.

2.2 Analogical Reasoning in AI Models

Analogical reasoning in AI has gained attention through various benchmarks and methodologies, revealing both the strengths and limitations of LLMs. Sultan and Shahaf (2022) detail a mechanism to extract analogies from a corpus of data describing a situation or a process. The entities of these texts are extracted and a mapping between these entities, or a cluster of entities, is build, connecting two texts in an analogy-like relation. Sourati et al. (2024) introduce the Analogical Reasoning on Narratives (ARN) benchmark, which extends traditional analogy evaluations by integrating narrative elements. This framework distinguishes near from far analogies, demonstrating LLMs’ proficiency in surface mappings yet exposing their limitations with abstract, far analogies under zero-shot conditions. Another benchmark is the AnaloBench (Ye et al., 2024), which tests the capabilities of LLMs to find analogies in a large dataset of texts. Short sentence analogies and analogies contained in a larger paragraph of text are tested, and the authors demonstrate that models like GPT3.5 and GPT4 still struggle to recognize analogies, especially with an increase in text size.

In this context, Yu et al. (2023) propose Thought Propagation (TP), a method that leverages the gen-

eration and resolution of analogous problems to iteratively refine model outputs, thereby achieving significant improvements over conventional baselines. Furthermore, Yuan et al. (2024) develop a knowledge base containing analogies, and show that training language models on this database improves the model’s ability to recognize and generate analogies. Complementing these approaches, Webb et al. (2023) compare LLM performance with human reasoning across varied analogy tasks, showing that while models like GPT-3 rival humans in structured analogies, they struggle with causal and cross-domain reasoning. Furthermore, Petersen and van der Plas (2023) align model evaluations with human-like paradigms, and Hu et al. (2023) show how encoding visual information into textual representations enhances LLMs’ performance on visual analogical reasoning, as they demonstrate with Raven’s Progressive Matrices.

Chain-of-thought (CoT) prompting encourages step-by-step reasoning in zero-shot settings (Kojima et al., 2022). In few-shot settings, when examples contain analogies, the model is explicitly guided to apply analogical reasoning (Wei et al., 2022b,a). Moreover, the term "analogy prompting" has already been used by Yasunaga et al. (2024), albeit in a different context. The authors further the idea of chain-of-thought by prompting the model to find similar math or coding problems in its knowledge base before trying to solve a given problem. They show that this methodology improves the ability of the model to solve math and coding problems in comparison to zero-shot and few-shot CoT. In the context of this paper, however, "analogy prompting" refers to prompting the model to generate analogies.

2.3 Spatial Schemas

Understanding how LLMs and VLMs conceptualize foundational spatial schemas is crucial for robust, intelligent systems. These schemas are the basic building blocks that infants use for spatial integration – a process described by Mandler (1992) as synthesizing perceptual experiences into conceptual representations via analogical reasoning.

Zhang et al. (2025) test the spacial reasoning of VLMs by asking spacial-related questions about a given image, i.e., "Is the blue ball in front of the red ball?", and "From the blue ball’s point of view, is the red ball to the right of the blue ball?". They find that VLMs’ answers tend to not be robust and consistent, especially when they are asked to adopt

a different frame of reference.

Richardson et al. (2001) study spatial schemas in adults and finds that commonly used verbs are consistently associated with a specific spatial direction (horizontal vs. vertical), which highlights the importance of spatial schemas in semantic representations even after the developmental stage.

Wicke and Wachowiak (2024) and Wicke et al. (2024) focus on the same stimuli used in Richardson et al. (2001) and assess whether a suite of LLMs and VLMs exhibits word-direction associations similar to humans’. Our work substantially extends their effort by using analogy-based prompting to gain deeper insights into model reasoning, incorporating state-of-the-art VLMs, and conducting a human subject study that not only validates previous results but also provides human analogies for direct comparison with those of models.

3 Methods

3.1 Experimental Setup

Our aim is to explore spatial intuitions in both humans and multimodal models by bridging a psycholinguistic study with computational modeling. We build upon the original study by Richardson et al. (2001), which provides the experimental stimuli of words and schematic directions (up, down, left, right) but has not been reproduced in over 20 years and did not explore the use of analogies. We conduct a human subject experiment where participants associate words with schematic directions and, additionally, provide the analogies they use for these associations (see Fig. 1). We query a variety of LLMs and VLMs – including GPT-4o (OpenAI, 2024a), Llama3 (AI@Meta, 2024), Molmo (Deitke et al., 2024), Qwen2-VL (Wang et al., 2024b), and others – with *regular* and *analogy* (i.e., explicitly asking the model to provide an analogy and to use it to provide its answer) prompts. We quantify the correlation between model and human schema selection in both prompting conditions and systematically compare the analogies generated by humans and models. These comparisons provide insights into how prompting strategies, modalities, and model architectures affect spatial associations.

Stimuli and Modalities In order to keep our results comparable to those by Richardson et al. (2001), we use the same stimuli as the original study (depicted in Fig. 2). The original stimuli include 30 verbs and pictures showing arrows. In Richardson et al.’s study, participants were asked

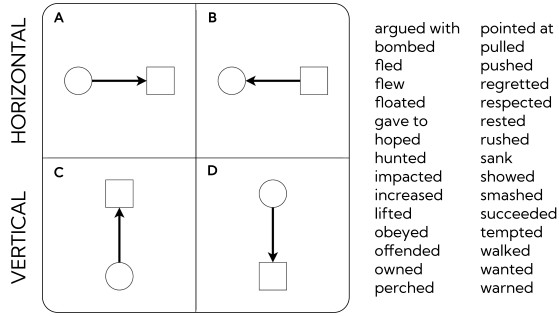


Figure 2: Left: Schematic directions used in all experiments. Right: Action words as experimental items. Both sets are adapted from Richardson et al. (2001).

to choose a preferred arrow (spatial schema) to represent each verb. In case of our studies, we present these spatial schemas in three different renderings: i) a reproduction of the original images (*visual condition*), ii) an equivalent Unicode version (\uparrow , \downarrow , \leftarrow , \rightarrow) of the arrows (*pseudo-visual condition*), and iii) a textual description (up, down, left, right) of the spatial schemas (*textual condition*).

3.2 Human Subject Study

We replicate the experiment by Richardson et al. (2001) with two key modifications designed to enhance both the task setup and subsequent analysis.

First, we introduce a one-shot example that diverges from the original relational schema (up, down, left, right) but retains a similar structure, designed to familiarize participants with the task without revealing the target relations (see App. Fig. 6). Second, we ask participants to provide an analogy explaining their choice before selecting one of the four options (see App. Fig. 7). Participants are asked to provide informed consent and demographic information (reported in App. A.2). We recruit 24 native English speakers, resulting in a total of 240 responses (30 items with 8 responses per item).

Schema Choice Evaluation To compare the results of our human study with those of Richardson et al. (2001), we calculated *item-level agreement* using a normalized concentration metric. This metric is based on the squared proportions of values within each distribution, ensuring it ranges from 0 (complete disagreement) to 1 (complete agreement). To account for sample size differences, scores were weighted by the number of observations (N) in both datasets. Overall agreement was computed as the weighted average across all items, with variability

assessed via standard deviation, offering insights into the consistency of item-level distributions.

Labeling Analogies To facilitate comparisons between human and model-generated analogies, we design a classification schema that categorizes them into four types (more details in App. A.4):

- Physical Action Representation
- Interaction or Relationship Between Entities
- Cultural or Conventional Associations
- No Analogy or Direct Explanation Provided

The creation of these labels was guided by prior NLP work in analogy classification (Mikolov et al., 2013; Gladkova et al., 2016; Drozd et al., 2016), as well as recent advancements in analogy evaluation (e.g., Wijesiriwardene et al., 2025). With guidance from these sources and insights from their analysis, our labels account for semantic and pragmatic influences on the structure of the analogy.

To label our dataset of +7,000 analogies, we employ LLMs as judges while acknowledging their limitations in reliability (Zheng et al., 2023; Bavaresco et al., 2024). On samples of 3x30 analogies from both human and LLM data, two annotators achieve an agreement of Cohen’s $\kappa = 0.6277$ after three annotation schema revisions, indicating their substantial agreement (Cohen, 1960).

When prompted according to this revised schema, GPT-4o achieves an agreement with two human annotators of Fleiss’ $\kappa = 0.6024$ (Fleiss and Cohen, 1973) (see details in App. A.4).

3.3 Generative Model Study

Large Language Models We select a diverse set of state-of-the-art LLMs, including both open-source and proprietary architectures. As open-source models, we include two variants of Llama 3.1 – Llama-70B and Llama-70B-Instruct (AI@Meta, 2024) – and DeepSeek’s R1-Distill-Llama (DeepSeek-AI et al., 2025), based on Llama-3.3-70B-Instruct. As proprietary models, we evaluate GPT-3.5-Turbo (OpenAI, 2023), GPT-4o, GPT-4o-Mini (OpenAI, 2024a), and GPT-o1-Preview (OpenAI, 2024b), accessed via the OpenAI API. LLMs were prompted by passing schemas as *textual* and *pseudo-visual* renderings.

Vision-language Models Given the documented limitations of vision-language models in spatial reasoning (Kamath et al., 2023; Wang et al., 2024a), we conduct a preliminary analysis to verify their

ability to correctly process the input images used in the main experiment (see App. B.2 for more details). VLMs from the LLaVA family (Liu et al., 2024c,a,b) were found to be incapable of reliably identifying our stimuli, and therefore excluded from our main experiment. Our selection of VLMs includes Molmo-7B, Molmo-72B (Deitke et al., 2024), Qwen2-VL-7B, and Qwen2-VL-72B (Wang et al., 2024b). These models were prompted with schemas in their *visual* rendering (as images).

Prompts We test both LLMs and VLMs in two prompting conditions (all with temperature 0, except for GPT-o1). In *regular* prompting, models are simply asked to provide their chosen schema for each verb; in the *analogy* prompting condition, they are asked to rely on an analogy to choose a schema, and to include both analogy and chosen schema in their response. Both kinds of prompts are one-shot, i.e., they include an example question, in-line with the human subject study. The complete list of prompts used for all models is provided in App. B.3. As suggested by Aher et al. (2023), we employ prompt validation to enhance the validity of model responses (see App. B.1 for more details). Despite these mitigation efforts, some invalid responses persisted (see App. B.4 for details).

3.4 Evaluation Metrics

We evaluate our models along two main dimensions: schema selection (textual, pseudo-visual, and visual) and labeled analogies.

For both dimensions, we compare model outputs and human responses with Spearman correlations and F1 scores (see App. A.6 and B.5 for more details). While the schema selection evaluation was performed against both human datasets, the one regarding analogy labels is only applicable to our dataset, because Richardson et al.’s data does not include human-generated analogies.

In addition to these task-level comparisons, we perform item-level analyses. For the human data, we assess the agreement between our human samples and the original data using item-level agreement measures. Moreover, we examine the item-level correlations of analogy types between selected models by comparing their outputs to our human-sampled analogies.

4 Results and Discussion

4.1 Human Subject Study

Our human study partially aimed to replicate Richardson et al. (2001), albeit with significant procedural differences. The item-level agreement analysis that we performed to compare Richardson et al.’s results to ours yields an overall weighted agreement of 0.49 (± 0.15) for Richardson et al.’s schema choices and 0.62 (± 0.26) for ours. Notably, items such as *pointed at* (0.80), *pushed* (0.78), and *bombed* (0.76) obtain the highest agreement in the Richardson dataset, whereas our dataset shows perfect agreement for items like *fled*, *pulled*, *sank*, and *increased*, albeit with a smaller sample size.

Altogether, our results indicate that the overall item-level agreement for our data is higher than that reported by Richardson et al. (2001). For further details, please refer to App. Tab. 2. We interpret the higher agreement in our dataset as suggesting that analogy prompting induces participants to deeply engage their knowledge about spatial schemas, as opposed to relying on simpler associations.

4.2 Generative Model Study

Our study with generative models focuses on comparing model outputs with human responses on two levels. First, we investigate how strongly the spatial schemas chosen by models align with those chosen by human participants from both our experiment and Richardson’s. Second, we explore the similarity between analogies generated by models and those provided by participants in our study.

4.2.1 Alignment of Spatial Schema Selection

We quantify alignment between models’ and humans’ schema choices by computing Spearman correlations and F1 scores. The former are shown in Fig. 3 and consider answer distributions aggregated per main direction (‘horizontal’ vs. ‘vertical’); this choice was favored over considering all four spatial schemas as it yielded more statistically significant correlations. F1 scores are reported in Tab. 1 and were calculated considering all four spatial schemas (up, down, left, right). Both Spearman correlations and F1 scores were computed per prompting condition (regular and analogy) and input type (textual, pseudo-visual, and visual).

Regular vs. analogy prompting Since we explicitly instructed our participants to employ analogical reasoning while Richardson et al. did not, we expected analogy-prompting model responses

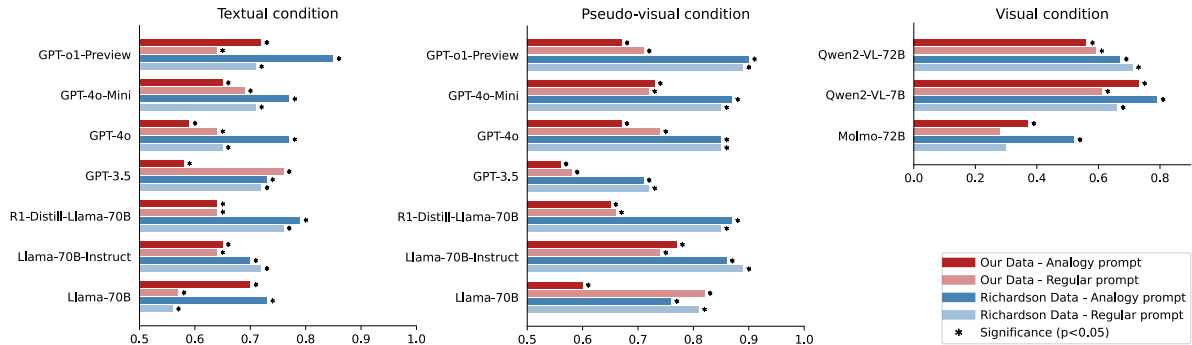


Figure 3: Spearman correlations between model and human chosen-concept distributions in the textual, pseudo-visual, and visual condition for our and Richardson et al.’s data. Values were computed per direction (‘horizontal’: up/down and ‘vertical’: left/right). Note that the x-axis range in the visual condition is different from the other two.

to align more closely with our dataset, and regular-prompting ones to be more aligned with Richardson et al.’s dataset. However, the Spearman correlations visualized in Fig. 3 indicate that none of the prompting conditions results in systematically stronger correlations with human responses. Moreover, the effect of the prompting condition is inconsistent even when the same model outputs are compared with different human datasets. As an example of this, in the textual condition, analogy prompting results in GPT-4o correlating more strongly with Richardson’s data than ours ($\rho_{Rich.} = 0.45 > \rho_{Ours} = 0.29$). A similar effect can also happen for the same model in different experimental conditions – e.g., for Llama-70B analogy prompting yields higher correlations with our dataset than regular prompting in the textual condition ($\rho_{Analog.} = 0.70 > \rho_{Reg.} = 0.57$), but the reversed trend is observed in the pseudo-visual condition ($\rho_{Analog.} = 0.60 < \rho_{Reg.} = 0.82$). Regarding the schema-wise F1 scores reported in Tab. 1, they do not indicate a systematic advantage of analogy prompting for our human data. However, an interesting trend is that, albeit with a few exceptions, analogy prompting tends to result in higher F1 scores for Richardson et al.’s data. Taken together, these findings suggest that models may process analogical relationships differently from humans, potentially relying more on learned associative patterns than true analogical reasoning.

Effect of input type Spearman correlations visualized in Fig. 3 allow a comparison among between input types (textual, pseudo-visual, visual). Overall, we observe stronger correlations in the pseudo-visual condition ($\rho = 0.56$ – 0.90) than in the textual condition ($\rho = 0.58$ – 0.85), but the trend

<i>Textual condition</i>				
Model	Our		Richardson	
	R	A	R	A
GPT-3.5	0.46	0.49	0.60	0.63
GPT-4o	0.33	0.29	0.40	0.45
GPT-4o-Mini	0.46	0.35	0.45	0.40
GPT-o1-Preview	0.35	0.44	0.35	0.49
Llama-70B	0.50	0.38	0.51	0.40
Llama-70B-Inst	0.33	0.37	0.41	0.48
R1-Distill-Llama-70B	0.45	0.41	0.53	0.58
<i>Pseudo-visual condition</i>				
Model	Our		Richardson	
	R	A	R	A
GPT-3.5	0.35	0.50	0.53	0.61
GPT-4o	0.41	0.42	0.58	0.63
GPT-4o-Mini	0.48	0.45	0.64	0.63
GPT-o1-Preview	0.50	0.46	0.64	0.67
Llama-70B	0.34	0.47	0.44	0.51
Llama-70B-Inst	0.46	0.49	0.6	0.63
R1-Distill-Llama-70B	0.49	0.45	0.69	0.63
<i>Visual condition</i>				
Model	Our		Richardson	
	R	A	R	A
Molmo-72B	0.05	0.16	0.05	0.15
Qwen2-VL-7B	0.23	0.22	0.18	0.34
Qwen2-VL-72B	0.35	0.38	0.41	0.51

Table 1: Weighted F1 scores between human and models’ concept preferences in the textual, pseudo-visual and visual conditions. Scores are reported for both our collected dataset and Richardson’s, and for the two different prompting conditions (**R** indicates regular prompting and **A** analogy prompting). Figures were computed concept-wise, i.e., considering all four spatial schemas.

is not systematic. A similar trend can be detected in the F1 scores (Tab. 1), whose range is 0.29–0.63 in

the textual condition and 0.34–0.69 in the pseudo-visual condition. One plausible explanation for this is that Unicode symbols reduce semantic ambiguities – particularly for words like "right" – which, in textual contexts, could be conflated with its "correctness" meaning. By providing a less ambiguous representation, pseudo-visual prompts may thus facilitate more accurate analogical mappings. Finally, correlations achieved by VLMs in the visual condition are, in general, lower than those achieved by LLMs in the other conditions ($\rho = 0.28\text{--}0.79$). This may be due to the visual condition posing the extra challenge of decoding the content of the visual stimuli. In other words, while LLMs receive abstract textual or pseudo-visual stimuli – which they can directly combine with their pretraining knowledge – VLMs are first tasked with mapping the different image(s) to abstract spatial notions and, only after completing this initial step, can they engage with their pretraining knowledge.

F1 scores and unbalanced concept productions

For some models, we observe systematic concept over- and underproductions, which affect the weighted F1 scores provided in Tab. 1. For example, Molmo-72B never produces ‘down’ and ‘right’ in the regular-prompt setup, while overproducing the answer ‘up’ (in 97% of its outputs); this results in an extremely low F1 score (0.05) for both our human responses and Richardson et al.’s. Similarly, Qwen2-VL-7B generates ‘up’ in 73% of the cases in the regular-prompting setup. Across all LLMs, there is a systematic trend to underproduce the concept ‘left’, and in some cases ‘down’. This tendency is especially extreme, e.g., for GPT-3.5 regular-prompted in the pseudo-visual condition (5% of ‘left’ responses), GPT-4o analogy-prompted in the textual condition (9% of ‘left’ responses), and Llama-70B regular-prompted in the pseudo-visual condition (8% of ‘down’ responses); in these cases, unbalanced model responses are again reflected in comparatively low F1 scores. Notably, while human participants also underproduce ‘left’ (19% in both datasets), this imbalance is not substantial enough to suggest a bias in the stimuli themselves. Instead, the models’ consistent underrepresentation of ‘left’ is more likely an artifact of biases in training data.

4.2.2 Human- vs. Model-generated Analogies

The Spearman correlations quantifying the similarity between analogies provided by human partici-

pants and models are visualized in Fig. 4. Although correlations are non-significant, some interesting trends emerge. First, the types of analogies generated by VLMs are the most aligned with those provided by humans ($\rho = 0.23\text{--}0.55$). Second, LLMs do not systematically generate more human-like analogies in the textual vs. pseudo-visual condition ($\rho_{Text.} = 0.00\text{--}0.17$, $\rho_{Pseudo-vis} = 0.00\text{--}0.20$). Finally, it is interesting that the types of analogies produced by GPT-o1-Preview – the only reasoning model – are the least similar to the human-provided ones, with a Spearman correlation of 0 in the pseudo-visual condition. These findings suggest that multimodal pretraining, while not resulting in models closely mirroring human schema choices, may help VLMs generate analogy types that are more similar to human ones than LLMs’ (examples of generated analogies in App. Tab. 4).

In a more focused analysis, we pick one LLM (GPT-4o) and check whether the items where its schema preferences align with the human ones are also those for which it generates more human-like analogy types. The results of this analysis are displayed in Fig. 5, which shows item-wise Spearman correlations with spatial schemas and analogy labels for the pseudo-visual condition. The correlations reveal a marked divergence between the models’ analogical mappings and schema selections for several verbs (e.g., *gave to*, *impacted*, *obeyed*).

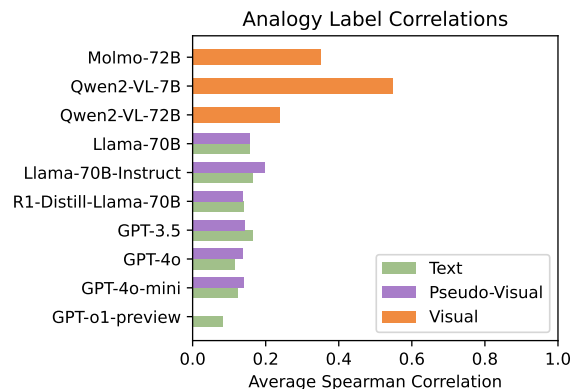


Figure 4: Correlations of the model’s chosen analogy types with those analogy types chosen by humans.

These differences may be due to two possible scenarios. First, a model might produce analogies similar to human analogical associations while choosing different spatial schemas; this would suggest a decoupling between analogical similarity and spatial mapping within the model’s reasoning process. Alternatively, a model might arrive at a

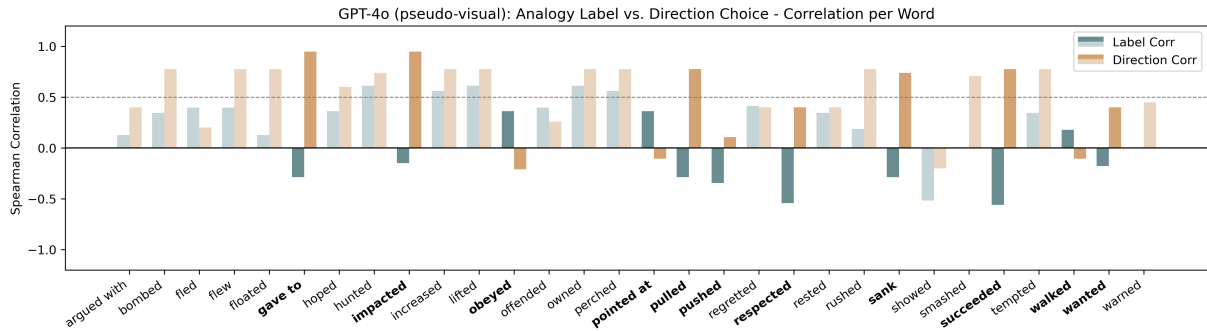


Figure 5: Spearman correlations for GPT-4o in the pseudo-visual condition, comparing human-model alignment on analogy labels (teal) and schema selection (orange) responses for 30 words. Highlighted bars and labels denote words where analogy and direction correlations are opposed, showing cases of potential decoupling of the two.

similar directional assignment as humans, yet the underlying analogical reasoning, as reflected in the label correlation, diverges markedly from human responses. Both of these scenarios occur 8 times in our example (highlighted bars and words in Fig. 5).

Overall, both model-wise correlations (Fig. 4) and the item-level analysis (Fig. 5) seem to point towards a similar conclusion, i.e., that models’ ability to produce analogies that resemble human ones does not necessarily result in human-aligned spatial-schema choices, and vice versa. This divergence is especially critical given that the words span abstract to concrete concepts, suggesting that the integration of analogical and spatial reasoning may be more fragile in contexts where multiple interpretative routes coexist.

4.3 Summary of Findings

Our analyses compare humans’ and generative models’ spatial intuitions on multiple levels (schema selection & analogy types) and consider two main experimental factors (prompting & modality). We now turn to our research questions.

RQ1 – Conceptualization of Abstract Notions through Analogies

Our experiments reveal substantial discrepancies between models’ and humans’ spatial conceptualizations. At the level of alignment between spatial choices, we do not observe a systematic improvement associated with analogy vs. regular prompting. These findings, together with a comparison between analogy types generated by humans and models, show that, even when models generate analogies similar to the human ones, these do not result in more human-aligned spatial schema choices. More importantly, this is true even when considering our human dataset, which was collected by explicitly asking

participants to rely on analogical reasoning. The discrepancies we document suggest that the profound differences between humans’ and models’ concept-learning processes are indeed reflected in spatial schemas, which appear to be supported by analogical reasoning in humans and simpler associations in models.

RQ2 – Effect of Multimodal Inputs on Analogical Reasoning

Our comparisons between experimental conditions employing different input types (textual, pseudo-visual, and visual) reveal three interesting trends. First, LLMs tend to produce more human-aligned schema choices in the pseudo-visual condition, which is likely due to reduced semantic ambiguity. Second, VLMs’ schema choices are, in general, less human-aligned than LLMs’ ones. Indeed, while images should be, in principle, the least semantically ambiguous input type, they still posit the extra challenge of extracting abstract meaning from the input stimuli. Finally, we observe that VLMs tend to generate types of analogies that are more similar to the human ones than LLMs. Taken together, these findings suggest that VLMs’ ability to process visual inputs proves advantageous in terms of producing human-like analogical reasoning. However, when focusing solely on associations between words and spatial schemas, Unicode arrows are the stimulus type associated with the most human-like choices; this may be due to them being abstract enough to not require perceptual processing and, at the same time, being less semantically ambiguous tokens than words.

5 Conclusions

Our study evaluates a suite of LLMs and VLMs concerning their ability to use analogical reasoning to support associations between verbs and spa-

tial schemas, a core component of human concept learning processing. We employ regular and analogy prompts to elicit these associations and compare them with human data from [Richardson et al. \(2001\)](#) and a set of newly collected human responses which, in contrast to [Richardson et al.](#), include human-written analogies. In addition, we explore how stimulus types varying in their degree of abstractness (textual, pseudo-visual, visual) influence model responses. Our experiments reveal substantial discrepancies between models' ability to generate analogies similar to the human ones and their ability to associate verbs to spatial schemas in a human-like way. LLMs and VLMs are increasingly applied in domains beyond language, including robotics, navigation, medicine, scientific discovery, and autonomous systems. However, their limitations in complex tasks suggest that performance gaps cannot be solely attributed to model size. While scaling improves alignment with human responses, our findings indicate that underlying analogical structures and spatial intuitions may diverge from human reasoning. This study highlights the need to examine fundamental conceptualization mechanisms to better understand these discrepancies and refine future models accordingly.

Limitations

A key limitation of our study is the potential for data contamination in [Richardson et al.](#)'s dataset. While it is unlikely that proprietary LLMs were explicitly fine-tuned on this dataset, it is possible that [Richardson et al.](#)'s paper was included in the pretraining data of certain models. This raises concerns that some observed correlations may not reflect genuine analogical reasoning, but rather memorized associations from training corpora. At present, a key mitigation effort is the dataset collected in our study, which was not publicly available during our evaluation phase and thus was not included in the training data of any model.

Additionally, differences in experimental design between our dataset and [Richardson et al.](#)'s may introduce confounds. Our explicit analogy-based prompting method engages different cognitive strategies than the spontaneous associations likely employed in [Richardson et al.](#)'s experiment. While we anticipated that this methodological distinction would result in stronger correlations for analogy-prompted responses in our dataset, our findings did not consistently support this hypothe-

sis. This discrepancy highlights the need for further research into how different prompting strategies interact with model architectures and training data to shape analogical reasoning performance.

We employed LLMs as annotation judges to assist in labeling our analogy dataset. This process followed an iterative refinement of the label classification schema, involving two human annotators, three rounds of revision, and the development of a carefully engineered prompt to ensure substantial agreement ([Cohen, 1960](#)). While we acknowledge the reliability limitations of LLM-based annotation ([Zheng et al., 2023](#); [Bavaresco et al., 2024](#)), this approach offered certain advantages over human annotators, particularly in mitigating inconsistencies that arose even within the same annotator.

While our study examines the reasoning capabilities of models, we include only a single designated "reasoning model" (o1-Preview). We acknowledge that such models may provide additional insights into underlying reasoning processes. However, as of now, they rely on advanced, predefined reasoning templates that are non-deterministic and not openly accessible. Furthermore, our focus is on capturing the models' intuitions after a single analogical reasoning step, rather than tracing multiple, potentially opaque reasoning iterations.

Responsible Research

Use of Artifacts We use both open and proprietary language models in our work. For all models, we include model cards or references to their respective providers, which specify their licenses and intended usage. Additionally, we use GitHub Copilot, powered by OpenAI Codex, and ChatGPT to generate code snippets. These tools provide outputs that are licensed for free use, ensuring compliance with their intended access conditions.

We also utilize research data from [Richardson et al. \(2001\)](#) and [Wicke and Wachowiak \(2024\)](#), which are publicly available research papers. The data derived from these sources is used strictly within research contexts, in accordance with their original access conditions. To the best of our knowledge, the use of all artifacts aligns with their specified terms, ensuring compliance with licensing and intended use policies.

Use of AI Assistance We used AI assistance tools (ChatGPT, OpenAI Playground, and GitHub Copilot) to aid in rewriting code, filter large datasets to identify additional trends, and refining

our labeling schema. All AI-generated content was thoroughly reviewed and verified by the authors. AI was not used to generate new research ideas or original findings; rather, it served as a support tool to improve clarity, efficiency, and organization. In accordance with ACL guidelines, our use of AI aligns with permitted assistance categories, and we have transparently reported all relevant usage in this paper. While AI contributed to enhancing the quality of the work, no direct research outputs are the result of AI assistance.

Acknowledgments

J.M.C's work is partially financed through national funds by FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025; and supported by the Portuguese PRR through project C645008882-00000055, Center for Responsible AI. A.B. was funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- AI@Meta. 2024. [Introducing llama 3.1: Our most capable models to date](#). Blog post.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, ..., and Alberto Testoni. 2024. [Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks](#). *arXiv preprint arXiv:2406.18403*.
- Dinara A Beitel, Raymond W Gibbs Jr, and Paul Sanders. 2001. The embodied approach to the polysemy of the spatial preposition on. In *Polysemy in cognitive linguistics*, pages 241–260. John Benjamins.
- Inge Bretherton and Elizabeth Bates. 1979. The emergence of intentional communication. *New Directions for Child and Adolescent Development*, 1979(4):81–100.
- Tom 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*, 33:1877–1901.
- Eve V Clark and Marisa Casillas. 2015. First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- 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, ..., and Zhen Zhang. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. [Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models](#). *arXiv preprint arXiv:2409.17146*.
- Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuo. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.
- Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.
- Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. [Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Bangkok, Thailand. Association for Computational Linguistics.
- Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.
- Raymond W Gibbs, Dinara A Beitel, Michael Harrington, and Paul E Sanders. 1994. Taking a stand on the meanings of stand: Bodily experience as motivation for polysemy. *Journal of semantics*, 11(4):231–251.
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuo. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

- Helmar Gust, Ulf Krumnack, Kai-Uwe Kühnberger, and Angela Schwering. 2008. Analogical reasoning: a core of cognition. *Künstliche Intell.*, 22(1):8–12.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.
- Douglas R. Hofstadter. 2001. *Epilogue: Analogy as the core of cognition*. In *The Analogical Mind: Perspectives from Cognitive Science*. The MIT Press.
- Keith J Holyoak. 2012. Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, pages 234–259.
- Xiaoyang Hu, Shane Storcks, Richard Lewis, and Joyce Chai. 2023. *In-context analogical reasoning with pre-trained language models*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.
- Jana M Iverson. 2010. Developing language in a developing body: The relationship between motor development and language development. *Journal of child language*, 37(2):229–261.
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. *Llava-next: Improved reasoning, ocr, and world knowledge*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Jean M Mandler. 1992. How to build a baby: Ii. conceptual primitives. *Psychological review*, 99(4):587.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. *Linguistic regularities in continuous space word representations*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- Philipp Mondorf and Barbara Plank. 2024. *Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey*. In *First Conference on Language Modeling*.
- OpenAI. 2023. *Gpt-3.5 turbo fine-tuning and api updates*.
- OpenAI. 2024a. *Gpt-4o model card*.
- OpenAI. 2024b. *Introducing openai o1-preview*.
- Molly Petersen and Lonneke van der Plas. 2023. *Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16414–16425, Singapore. Association for Computational Linguistics.
- Daniel C Richardson, Michael J Spivey, Shimon Edelman, and Adam J Naples. 2001. "language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, ..., and Thomas Wolf. 2023. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. Working paper or preprint.
- Adam Sobieszek and Tadeusz Price. 2022. Playing games with ais: the limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364.
- Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. *ARN: Analogical reasoning on narratives*. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.
- Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. *GPT-4 doesn’t know it’s wrong: An analysis of iterative prompting for reasoning problems*. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.
- Oren Sultan and Dafna Shahaf. 2022. *Life is a circus and we are the clowns: Automatically finding analogies between situations and processes*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.
- Stella Vosniadou and Andrew Ortony. 1989. *Similarity and analogical reasoning*. Cambridge University Press.
- Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. 2024a. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. *Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution*. *arXiv preprint arXiv:2409.12191*.
- Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. *Emergent abilities of large language models*. *Preprint*, arXiv:2206.07682.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. *Chain of thought prompting elicits reasoning in large language models*. In *Advances in Neural Information Processing Systems*.
- Philipp Wicke, Lea Hirlimann, and João Miguel Cunha. 2024. Using analogical reasoning to prompt llms for their intuitions of abstract spatial schemas. In *First Workshop on Analogical Abstraction in Cognition, Perception, and Language at IJCAI 2024*.
- Philipp Wicke and Lennart Wachowiak. 2024. *Exploring spatial schema intuitions in large language and vision models*. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6102–6117, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Thilini Wijesiriwardene, Ruwan Wickramarachchi, Sreeram Reddy Vennam, Vinija Jain, Aman Chadha, Amitava Das, Ponnurangam Kumaraguru, and Amit Sheth. 2025. Knowledgeprompts: Exploring the abilities of large language models to solve proportional analogies via knowledge-enhanced prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3979–3996.
- Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. *Large language models as analogical reasoners*. (arXiv:2310.01714). ArXiv:2310.01714 [cs].
- Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. *Analobench: Benchmarking the identification of abstract and long-context analogies*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 13060–13082, Miami, Florida, USA. Association for Computational Linguistics.
- Junchi Yu, Ran He, and Zhitao Ying. 2023. Thought propagation: An analogical approach to complex reasoning with large language models. In *The Twelfth International Conference on Learning Representations*.
- Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. *Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.
- Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. *Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities*. (arXiv:2410.17385). ArXiv:2410.17385 [cs].
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Human Study

A.1 Survey Design

The survey was conducted using *Google Forms*. All participants provided their informed consent to participate in our study. No names, addresses, IPs or traceable information was collected, and the participants could decide to end the study at any point. In order to familiarize the participants with the task, an example task was provided (Fig. 6). The example task used the same format as the real task, but the symbols and the direction (diagonal as opposed to vertical/horizontal) were different. We tested the survey design with peers before collecting responses from non-peers. The test responses have not been included in the final data collection.

Consider the event "▲ stopped △" and the four images below (A, B, C, D). Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice.

A 	B
C 	D

Analogy *

stopping often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone, for example.

Choice *

A

B

C

D

Figure 6: All participants in the study are presented with an example item (one-shot) at the start of the questionnaire. This allows the participants to familiarize themselves with the task, while not providing a priming effect due to the use of a different directionality (diagonal as opposed to vertical/horizontal) and different symbols (triangles as opposed to circle/square).

For each of the 30 items, we generated a question shown in Fig. 7. We use the same visual stimuli as Richardson et al. (2001) for our human subject study. We note that in the original study, the participants were presented with the entire list of 30 items at once (next to the same picture, which we repeat for each item).

Consider the event "○ fled □" and the four images below (A, B, C, D). Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice.

Description (optional)

Image title

A 	B
C 	D

Analogy *

Short answer text

Choice *

A

B

C

D

Figure 7: Example item presented to the participants. First, they are asked to provide an analogy, then they are asked to choose one of four images that best relates to the options (A, B, C, D).

A.2 Demographics

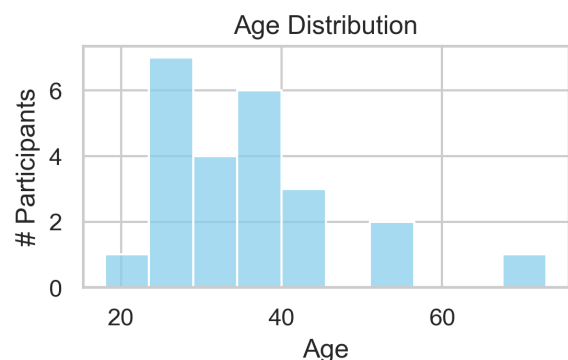


Figure 8: Distribution of age for $N = 24$ participants. Average age is 35.54.

We sampled $N = 24$ participants with two restrictions: (i) Native English speakers, (ii) no prior knowledge about this research. To the best of our knowledge, no participant self-reported significant or severe visual or cognitive impairments.

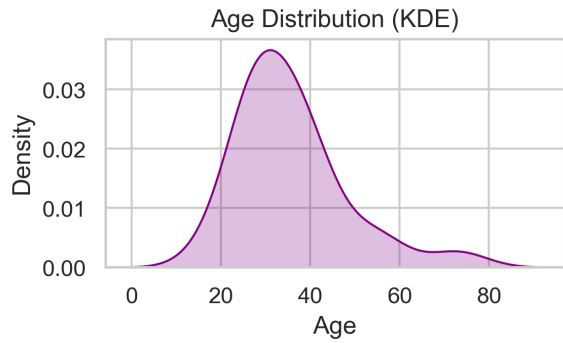


Figure 9: Kernel density estimate (KDE) to represent participants' ($N = 24$) age as spectrum, with an average around 35 years.

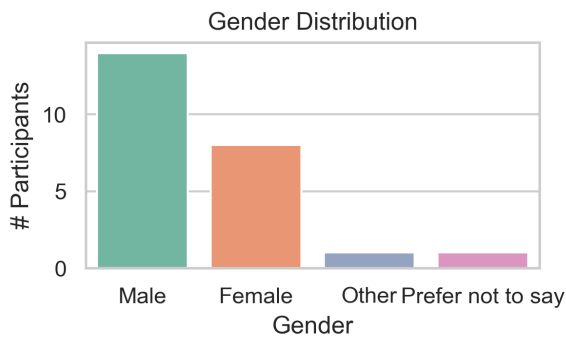


Figure 10: Gender distribution of all $N = 24$ participants: Male: 14 participant(s), female: 8 participant(s), other: 1 participant, prefer not to say: 1 participant.

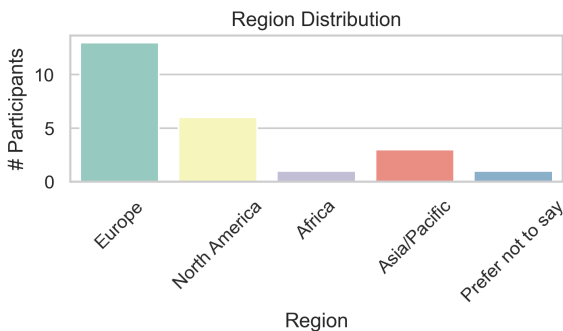


Figure 11: All participants declared that they are native English speakers. The regional distribution is as follows: Europe: 13 participant(s), North America: 6 participant(s), Africa: 1 participant, Asia/Pacific: 3 participant(s), Prefer not to say: 1 participant.

A.3 Human Study Results

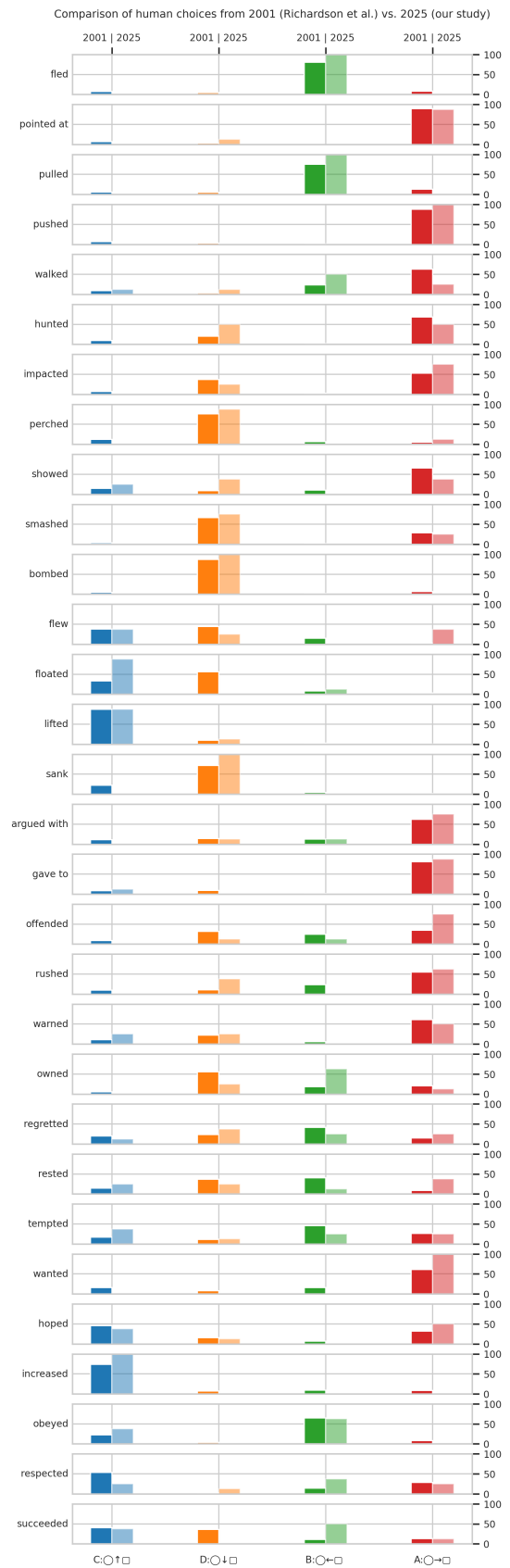


Figure 12: Comparison of the data by Richardson et al. (2001) with the human choices gathered in our study.

A.4 Analogy Annotation Methodology

We sampled 30 analogies (15 human-created, 15 GPT-4o-generated) and classified them into four categories: “Physical Action,” “Cultural/Convention,” “Interactive Entities,” or “No Analogy/Explanation.” In a second round, two authors annotated a different set of 30 analogies using this scheme. Annotator agreement was measured using Cohen’s κ (Cohen, 1960). After three revisions of the annotation scheme, we achieved $\kappa = 0.6277$, indicating substantial agreement. All annotation schema versions are available in the code repository. The final schema, incorporating these revisions as additional rules, was then formalized into a prompt:

Task: You will be provided with an explanation that uses a directional or movement analogy to describe an event, action, or reaction. Your job is to carefully read the explanation, assess the type of analogy it employs, and select one of the following labels that best corresponds to it:

- **Physical Action** – This label applies if the explanation relies on tangible movements, forces, or physical processes.
- **Cultural/Convention** – This label applies if the explanation relies on societal norms, symbolic interpretations, or culturally shared meanings related to direction or movement.
- **Interactive Entities** – This label applies if the explanation emphasizes the interaction or relationship between distinct entities (e.g., square and circle).
- **No Analogy/Explanation** – This label applies if the explanation is purely descriptive, with no directional, movement-based, or analogical content.

Additional rules:

- If the explanation mentions “square” or “circle,” it is always labeled **Interactive Entities**.
- If the explanation does not mention these shapes implicitly or explicitly, and no entities are present, then it is not **Interactive Entities**.
- If the explanation mentions “culture,” it is always **Cultural/Convention**.
- If the explanation includes technical or scientific analogies (e.g., diagrams or systems), it is always **Cultural/Convention**.
- If the explanation references gravity, understand gravity as a physical action and assign **Physical Action**.

Here is the explanation: Explanation

Based solely on your analysis of the explanation above, provide only one label from the following:

Physical Action, Cultural/Convention, Interactive Entities, or No Analogy/Explanation.

A.5 Choice Coherence

Item	Richardson	Our (w/ analogy)
pointed at	0.80	0.78
pushed	0.78	1.00
lifted	0.77	0.78
bombed	0.76	1.00
fled	0.67	1.00
gave to	0.67	0.78
perched	0.60	0.78
pulled	0.59	1.00
sank	0.57	1.00
increased	0.57	1.00
smashed	0.53	0.62
hunted	0.52	0.50
obeyed	0.48	0.53
walked	0.47	0.34
showed	0.47	0.34
argued with	0.44	0.59
warned	0.44	0.38
floated	0.43	0.78
wanted	0.43	1.00
impacted	0.42	0.62
owned	0.39	0.47
respected	0.39	0.28
rushed	0.38	0.53
flew	0.36	0.34
hoped	0.34	0.41
rested	0.32	0.28
tempted	0.32	0.28
succeeded	0.32	0.41
regretted	0.29	0.28
offended	0.29	0.59
Overall	0.49 (± 0.15)	0.62 (± 0.26)

Table 2: Item-wise agreement scores for the choice (of direction) measure computed using a normalized concentration metric (i.e., squared proportions weighted by the number of observations, yielding values from 0 to 1). This metric quantifies how concentrated the responses are for each item — scores near 1 signify that nearly all raters converge on the same label (indicating high consensus), whereas lower values reflect greater variability in judgments. “Richardson” refers to the human data reported by Richardson et al. (2001) and “Our” refers to the data collected in the present study. The final row gives the overall weighted agreement and its standard deviation.

A.6 Label Evaluation

For each word, we first compute frequency distributions over the four label categories from human responses (8 responses) and model responses (24 responses). These distributions are then converted into ranked vectors by ordering categories according to their frequencies. Spearman correlation is computed between the human and model ranked frequency vectors, quantifying the monotonic agreement in label usage. In parallel, for each

category, the F1 score is calculated via

$$F_1 = \frac{2 \times \min(\text{count}_{\text{human}}, \text{count}_{\text{model}})}{\text{count}_{\text{human}} + \text{count}_{\text{model}}} \quad (1)$$

(with a default score of 1 when both counts are zero).

Model	Condition	Int. Coh. \uparrow	JS Div. \downarrow	Entr. \downarrow
Human Reference	Ref.	0.550	—	1.760
gpt-3.5	Pseudo	0.933	0.436	0.920
gpt-4	Pseudo	0.876	0.443	0.883
gpt-4-mini	Pseudo	0.839	0.443	0.904
llama-70b	Pseudo	0.929	0.399	0.830
llama-70b-inst	Pseudo	0.922	0.436	0.642
gpt-3.5	Text	0.907	0.417	0.813
gpt-4	Text	0.861	0.449	0.981
gpt-4-mini	Text	0.874	0.450	0.678
llama-70b	Text	0.929	0.409	0.885
llama-70b-inst	Text	0.856	0.443	0.910

Table 3: Evaluation metrics for five LLM configurations under Pseudo and Text conditions compared to a human reference. “Int. Coh.” (**Internal Coherence**) is the average fraction of label agreement per item, reflecting labeling consistency. “Entr.” (**Entropy**) quantifies the diversity of the label distribution, and “JS Div.” (**Jensen–Shannon Divergence**) measures the similarity of the model’s distribution to that of humans.

A.7 Label Examples

B Model Studies

B.1 Validation Scores

In order to improve model responses, we tested different prompt endings and calculated a validation score that measured how often the model, when given a regular prompt, produced a valid response. To achieve this, we generated a model response for each of the 30 action words using the following prompt:

Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose?
[ending]: 'X'

Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose?
[ending]:

where [ending] is one possible prompt ending (e.g., “CONCEPT”, “Choice”, and “selection”),

[concept] refers to the four spatial concepts, and [event] is an action word. We employed Llama3.1-8B (AI@Meta, 2024) as the LLM for this experiment, based on the rationale that if a smaller model can produce a valid answer with a specific ending, then larger models are likely to do so as well. As described in Wicke and Wachowiak (2024), the validity score is calculated as follows:

$$\sum_{a \in V} p(a|k) \quad (2)$$

for a specific prompt k and a set of valid answers V .

This evaluation was conducted under both textual and pseudo-visual conditions, and an average validation score was computed for each ending. As a result, the ending “concept” was selected, achieving a 74% validation score. The full list of prompt endings, along with their corresponding validation scores, is presented in Table 5.

B.2 Vision-language Models Validation

The task considered in our main experiment is challenging even for current vision-language models for two main reasons. First, it requires correctly distinguishing the arrow depictions mainly based on their spatial orientation. Second, it requires the ability to process multiple images simultaneously. Since most state-of-the-art VLMs still struggle with spatial reasoning (Kamath et al., 2023; Wang et al., 2024a) or are simply not trained for multi-image inference, we designed a simple validation task to help us select VLMs that could be suitably included in our main experiment. The task consisted in responding correctly to the four possible variations of the prompt:

Which of the following is the [up, down, left, right]ward-pointing arrow?
Image SHTL: <image up>
Image XHWK: <image down>
Image AKRC: <image left>
Image ZHRN: <image right>
Please respond with the four-letter image id.

The following similar prompt was used for a single-image scenario, where all the concepts were provided within the same image:

Which is the [up, down, left, right]ward-pointing arrow? SHTL, XHWK, AKRC or ZHRN?

Action	Contributor	Example Analogy	Assigned Label
wanted	Human	circle wants square. as circle is first its on the left, furthermore the arrow should point towards the square as it shows circle acting towards square	Interactive Entities
fled	GPT-3.5 (pseudo-visual)	'Fled' often involves moving quickly away from a dangerous or threatening situation. An analogy could be a person running away from a predator in a straight line to escape.	Physical Action
pointed at	GPT-4o-mini (pseudo-visual)	'Pointing at' something typically involves directing attention towards a specific object or direction, often using a finger or hand to indicate where to look. The action of pointing is about indicating a specific location or item in space.	Physical Action
pulled	GPT-4o (pseudo-visual)	'Pulling' often involves drawing something towards oneself or in a specific direction. When you pull an object, you typically exert force to bring it closer to you or to move it in a desired direction. In a directional sense, pulling can be associated with moving something from a lower position to a higher one, as if lifting or drawing it upwards.	Physical Action
showed	GPT-3.5 (textual)	'showing' often involves presenting or revealing something in a particular direction. For example, pointing towards a specific direction to indicate where something is located.	Physical Action
obeyed	GPT-4o-mini (textual)	'obeying' often involves following directions or commands, which can be likened to moving in a specific direction as instructed. When someone is told to go 'up', they are complying with a directive, just as one would follow orders or rules in a broader sense.	Cultural/Convention
rushed	GPT-4o (textual)	'Rushing' often involves moving quickly and with urgency towards a destination or goal. It implies a sense of forward momentum and progress, similar to how one might move in a straight line without hesitation. In many contexts, moving 'up' can symbolize advancement, progress, or moving towards a goal, as it is often associated with positive movement or elevation.	Cultural/Convention
argued with	Qwen-VL-72b	'argued with' often involves opposing or challenging someone's views. A debate between two people, for example, is a common representation of this event.	Interactive Entities
hoped	Qwen-VL-7b	'hoping' involves having a desire or wish for something to happen. It's like having a goal or aspiration.	No Analogy / Explanation

Table 4: Examples of different collected analogies from different contributors. Selection was focused on representing different assigned labels. Full collection of analogies is available at https://github.com/anonymousACL/analogy_prompting.

Ending	Textual	Pseudo	Avg.
CHOICE	0.53	0.65	0.59
Choice	0.65	0.72	0.68
choice	0.70	0.77	0.73
SELECTION	0.66	0.73	0.69
Selection	0.69	0.75	0.72
selection	0.68	0.75	0.71
CONCEPT	0.68	0.75	0.71
Concept	0.69	0.73	0.71
concept	0.73	0.76	0.74

Table 5: Overview of the validation scores for each possible prompt-ending, for textual and pseudo-visual prompts, along with their average.

The models tested in the multi-image scenario were Qwen2-VL-7B-Instruct² and llava-onevision-qwen2-7b-ov-hf³. The models tested in the single-image scenario were: Molmo-7B-D-0924⁴, llama3-llava-next-8b-hf⁵, llava-v1.6-mistral-7b-hf⁶,

²<https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct>

³<https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf>

⁴<https://huggingface.co/allenai/Molmo-7B-D-0924>

⁵<https://huggingface.co/llava-hf/llama3-llava-next-8b-hf>

⁶<https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf>

llava-onevision-qwen2-7b-si-hf⁷, ,
llava-interleave-qwen-7b-hf⁸, and
Qwen2-VL-7B-Instruct⁹.

The only models which were able to respond correctly to all variants of the prompts were Molmo-7B-D-0924 in the single-image scenario and Qwen2-VL-7B-Instruct in the multi-image scenario. Given the satisfactory performance of these 7B-parameter models, we decided to include their largest versions (Molmo-72B-0924¹⁰ and Qwen2-VL-72B-Instruct-AWQ¹¹) as well in the main experiment.

B.3 Prompts

The prompts used for the LLMs and vision-language models are reported, respectively, in Tables 6, 7, and 8. To avoid selection bias (e.g., the model always choosing the option appearing as first), for each prompt we constructed variations corresponding to all the possible label permutations ($4! = 24$).

Note that, since the preview Molmo version available when experiments were conducted (Fall 2024) did not support multi-image inference, this model was prompted with a single image including all four spatial schemas. As for the Qwen2-VL models, they were found incapable of discriminating between schemas when they were provided within the same image; therefore, each schema was provided within a separate image.

B.4 Parsing of Model Outputs

Despite our efforts to validate the prompts, there were still cases where model-generated responses did not exactly match the expected structure. When this occurred, we first tried to exploit other regularities (e.g., the model outputting choice: instead of concept:) to isolate the relevant part of the output. When no such regularity was present, we adopted a simpler single-matching approach: if a single concept could be identified in the output, we considered that as a valid answer; if not, or in the case where multiple concepts were present, we considered the output invalid.

⁷<https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-si-hf>

⁸<http://llava-hf/llava-interleave-qwen-7b-hf>

⁹<https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct>

¹⁰<https://huggingface.co/allenai/Molmo-72B-0924>

¹¹<https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct-AWQ>

To obtain comparable label distributions, we replaced the invalid answers with the prevalent valid answer for the action word. If no valid answer was returned for a specific action word, we excluded the action word from further comparisons with human preferences. The percentage of invalid answers never exceeded 5%. We report the percentage of invalid responses yielded by each model in Table 15.

B.5 Evaluation Metrics

For each model, we obtained 24 outputs for each verb-stimulus (corresponding to all possible permutations). This allowed us to obtain a percentage of ‘up’, ‘down’, ‘left’ and ‘right’ responses for each verb. Similarly, response distributions could be obtained for the human datasets based on multiple participants’ answers. Spearman correlations quantifying the alignment between human and model responses were computed between the human and model-generated answer distributions per each spatial schema. Since most correlations were non-significant when considering all four spatial schemas, in the main paper we show correlations per direction (‘vertical’ vs. ‘horizontal’). In the following sections, we report complete results, including both per-schema and per-direction correlation.

Regarding F1 scores, they were computed between human and model-generated schema preferences. For each verb, the schema preference was defined as the schema appearing in most permutations or indicated by the majority of participants.

B.6 Extended Results

In Tables 9, 10, and 11 the Spearman correlations for the LLMs and VLMs are represented. F1 scores for these models can be seen in Tables 12, 13, and 14.

B.7 Compute Resources

Running our experiments required a total of approximately 37 hours on an A100 NVIDIA GPU and 10 hours on an H100 NVIDIA GPU. Labeling the analogies with GPT-4o using OpenAI’s API access required approximately 3 hours.

Model	Prompt
Llama (R)	<p>< begin_of_text >< start_header_id >system< end_header_id ></p> <p>You are a participant in a research experiment< eot_id >< start_header_id >user< end_header_id ></p> <p>Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose?< eot_id >< start_header_id >assistant< end_header_id ></p> <p>concept: 'X'< eot_id >< start_header_id >user< end_header_id ></p> <p>Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose?< eot_id >< start_header_id >assistant< end_header_id ></p> <p>concept:</p>
Llama (A)	<p>< begin_of_text >< start_header_id >system< end_header_id ></p> <p>You are a participant in a research experiment< eot_id >< start_header_id >user< end_header_id ></p> <p>Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose? Explain the analogy, then provide one choice.< eot_id >< start_header_id >assistant< end_header_id ></p> <p>analogy: 'stopping' often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone for example.</p> <p>concept: 'X'< eot_id >< start_header_id >user< end_header_id ></p> <p>Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? Explain the analogy, then provide one choice.< eot_id >< start_header_id >assistant< end_header_id ></p> <p>analogy:</p>
Llama (Z)	<p>< begin_of_text >< start_header_id >system< end_header_id ></p> <p>You are a participant in a research experiment< eot_id >< start_header_id >user< end_header_id ></p> <p>Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? Give the chosen concept by surrounding it with '##'.< eot_id >< start_header_id >assistant< end_header_id ></p> <p>Let's think step by step.</p>

Table 6: Prompts used for the Llama 3.1 models. The R1-Distill-Llama model used the same prompt as the regular Llama models. The letters in brackets after the model names refer to the experimental condition (**R**egular vs. **A**nalogical vs. **Z**ero-shot.)

Model	Prompt
GPT (R)	<p>SYSTEM_PROMPT:</p> <p>You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective. Follow the given structure.</p> <p>USER_PROMPT:</p> <p>EXAMPLE TASK: Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose? concept: 'X'</p> <p>TASK: Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? concept:</p>
GPT (A)	<p>SYSTEM_PROMPT:</p> <p>You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective. Follow the given structure.</p> <p>USER_PROMPT:</p> <p>EXAMPLE TASK: Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose? Explain the analogy, then provide one choice. analogy: 'stopping' often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone for example. concept: 'X'</p> <p>TASK: Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? Explain the analogy, then provide one choice. analogy:</p>
GPT (Z)	<p>You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective. Follow the given structure. TASK: Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? Give the chosen concept by surrounding it with '##'. Let's think step by step.</p>

Table 7: Prompts used for the GPT models. The letters in brackets after the model names refer to the experimental condition (**R**egular vs. **A**nalogical. vs. **Z**ero-shot.)

Model	Prompt
Molmo (R)	<p>Example task: Consider the event ‘threw’ and the four images below (SHTL, XHWK, AKRC, ZHRN). Which of the images best represents the event? Image: XHWK</p> <p>Task: Consider the event ‘[event]’ and the four images below (SHTL, XHWK, AKRC, ZHRN). Which of the images best represents the event? Image:</p>
Qwen2-VL (R)	<p>Example task: Consider the event ‘stopped’ and these four images: SHTL [image], XHWK [image], AKRC [image], ZHRN [image]. Which of the images best represents the event? Image: SHTL</p> <p>Task: Consider the event ‘stopped’ and these four images: [image label][image], [image label]<image>, [image label]<image>, [image label]<image>. Which of the images best represents the event? Image:</p>
Molmo (A)	<p>Example task: Consider the event ‘threw’ and the four images below (SHTL, XHWK, AKRC, ZHRN). Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy: ‘throwing’ often involves launching an object in a horizontal direction. The trajectory followed by the object could be represented as a rightward-pointing arrow. Image: XHWK</p> <p>Task: Consider the event [event] and the four images below (SHTL, XHWK, AKRC, ZHRN). Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy:</p>
Qwen2-VL (A)	<p>Example task: Consider the event ‘stopped’ and these four images: SHTL <image>, XHWK <image>, AKRC <image>, ZHRN <image>. Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy: ‘stopping’ often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone for example. Image: SHTL</p> <p>Task: Consider the event ‘[event]’ and these four images: [image label]<image>, [image label]<image>, [image label]<image>, [image label]<image>. Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy:</p>

Table 8: Prompts used for the vision-language models. The letters in brackets after the model names refer to the experimental condition (**R**egular vs. **A**nalogical.)

	Llama-70B		Llama-70B-Inst			R1-Distill-Llama-70B		
	R	A	R	A	Z	R	A	
Richardson	Up	0.45*	0.53* (+)	0.67*	0.57* (-)	0.61* (-)	0.48*	0.63* (+)
	Down	0.47*	0.31 (-)	0.31	0.27 (-)	0.33 (+)	0.37*	0.44* (+)
	Left	0.34	0.44* (+)	0.36	0.46* (+)	0.07 (-)	0.25	0.47* (+)
	Right	0.58*	0.56* (-)	0.58*	0.57* (-)	0.62* (+)	0.62*	0.61* (-)
	↑	0.67*	0.58* (-)	0.72*	0.66* (-)	0.68* (-)	0.68*	0.57* (-)
	↓	0.66*	0.38* (-)	0.48*	0.49* (+)	0.48* (=)	0.58*	0.62* (+)
	←	0.12	0.61* (+)	0.42*	0.44* (+)	0.33 (-)	0.43*	0.62* (+)
	→	0.47*	0.61* (+)	0.67*	0.72* (+)	0.77* (+)	0.69*	0.68* (-)
	Hor./Vert. ^T	0.56*	0.73* (+)	0.72*	0.70* (-)	0.72* (=)	0.76*	0.79* (+)
Hor./Vert. ^P	0.81*	0.76* (-)	0.89*	0.86* (-)	0.88* (-)	0.85*	0.87* (+)	
Ours	Up	0.57*	0.58* (+)	0.56*	0.51* (-)	0.48* (-)	0.47*	0.58* (+)
	Down	0.47*	0.45* (-)	0.43*	0.40* (-)	0.40* (-)	0.53*	0.57* (+)
	Left	0.38*	0.42* (+)	0.39*	0.36 (-)	0.17 (-)	0.36*	0.49* (+)
	Right	0.47*	0.41* (-)	0.37*	0.35 (-)	0.37* (=)	0.36*	0.33 (-)
	↑	0.70*	0.52* (-)	0.64*	0.60* (-)	0.66* (+)	0.64*	0.50* (-)
	↓	0.60*	0.51* (-)	0.52*	0.53* (+)	0.52* (=)	0.50*	0.50* (=)
	←	0.12	0.59* (+)	0.38*	0.53* (+)	0.44* (+)	0.44*	0.55* (+)
	→	0.37*	0.45* (+)	0.50*	0.52* (+)	0.56* (+)	0.53*	0.41* (-)
	Hor./Vert. ^T	0.57*	0.70* (+)	0.64*	0.65* (+)	0.62* (-)	0.64*	0.64* (=)
Hor./Vert. ^P	0.82*	0.60* (-)	0.74*	0.77* (+)	0.70* (-)	0.66*	0.65* (-)	

Table 9: Spearman correlations between concept distributions by humans and the open-source models (Llama3.1 and DeepSeek R1 Distill Llama). The last four rows report results aggregated into two main directions (‘up’ and ‘down’ into ‘vertical’ and ‘left’ and ‘right’ as ‘horizontal’), for textual (*T*) and pseudo-visual (*P*) concepts. Values in the ‘R’ column refer to the *regular* prompting condition, while ‘A’ indicates *analogy* prompting, and ‘Z’ indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results in an improved correlation with respect to regular prompting (+), remained the same (=), or didn’t improve (-). Asterisks mark statistical significance ($p < 0.05$).

		GPT-3.5		GPT-4o		GPT-4o-Mini		GPT-o1-Preview		
		R	A	R	A	R	A	R	A	Z
Richardson	Up	0.63*	0.48* (-)	0.59*	0.61* (+)	0.61*	0.63* (+)	0.60*	0.58* (-)	0.57* (-)
	Down	0.51*	0.35 (-)	0.41*	0.45* (+)	0.26	0.22 (-)	0.41*	0.34 (-)	0.35 (-)
	Left	0.43*	0.52* (+)	0.32	0.45* (+)	0.36	0.47* (+)	0.35	0.45* (+)	0.26 (-)
	Right	0.69*	0.68* (-)	0.52*	0.65* (+)	0.59*	0.60* (+)	0.59*	0.69* (+)	0.55* (-)
	↑	0.58*	0.47* (-)	0.73*	0.68* (-)	0.69*	0.63* (-)	0.64*	0.69* (+)	0.66* (+)
	↓	0.55*	0.32 (-)	0.59*	0.52* (-)	0.56*	0.36 (-)	0.59*	0.52* (-)	0.47* (-)
	←	0.23	0.29 (+)	0.36	0.49* (+)	0.52*	0.43* (-)	0.46*	0.53* (+)	0.21 (-)
	→	0.69*	0.63* (-)	0.68*	0.64* (-)	0.74*	0.76* (+)	0.70*	0.68* (-)	0.67* (-)
	Hor./Vert. ^T	0.72*	0.73* (+)	0.65*	0.77* (+)	0.71*	0.77* (+)	0.71*	0.85* (+)	0.74* (+)
Hor./Vert. ^P	0.72*	0.71* (-)	0.85*	0.85* (=)	0.85*	0.87* (+)	0.89*	0.90* (+)	0.86* (-)	
Ours	Up	0.60*	0.44* (-)	0.63*	0.58* (-)	0.61*	0.56* (-)	0.55*	0.49* (-)	0.49* (-)
	Down	0.62*	0.44* (-)	0.49*	0.41* (-)	0.33	0.37* (+)	0.54*	0.48* (-)	0.45* (-)
	Left	0.36*	0.56* (+)	0.38*	0.38* (=)	0.38*	0.50* (+)	0.24	0.36 (+)	0.10 (-)
	Right	0.47*	0.50* (+)	0.37*	0.36 (-)	0.40*	0.40* (=)	0.43*	0.57* (+)	0.44* (+)
	↑	0.54*	0.46* (-)	0.63*	0.67* (+)	0.59*	0.64* (+)	0.55*	0.56* (+)	0.58* (+)
	↓	0.54*	0.36 (-)	0.55*	0.54* (-)	0.58*	0.45* (-)	0.59*	0.51* (-)	0.39* (-)
	←	0.25	0.28 (+)	0.34	0.54* (+)	0.50*	0.42* (-)	0.51*	0.54* (+)	0.35 (-)
	→	0.52*	0.42* (-)	0.44*	0.47* (+)	0.50*	0.54* (+)	0.48*	0.48* (=)	0.50* (+)
	Hor./Vert. ^T	0.76*	0.58* (-)	0.64*	0.59* (-)	0.69*	0.65* (-)	0.64*	0.72* (+)	0.59* (-)
Hor./Vert. ^P	0.58*	0.56* (-)	0.74*	0.67* (-)	0.72*	0.73* (+)	0.71*	0.67* (-)	0.65* (-)	

Table 10: Spearman correlations between concept distributions by humans and the GPT models. The last four rows report results aggregated into two main directions (‘up’ and ‘down’ into ‘vertical’ and ‘left’ and ‘right’ as ‘horizontal’), for textual (T) and pseudo-visual (P) concepts. Values in the ‘R’ column refer to the *regular* prompting condition, while ‘A’ indicates *analogy* prompting, and ‘Z’ indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results in an improved correlation with respect to regular prompting (+), remained the same (=), or didn’t improve (-). Asterisks mark statistical significance ($p < 0.05$).

		Molmo-7B		Molmo-72B		Qwen2-VL-7B		Qwen2-VL-72B	
		R	A	R	A	R	A	R	A
Richardson	Up	0.11	0.29 (+)	0.19	0.32 (+)	0.22	0.56* (+)	0.53*	0.37* (-)
	Down	0.36	-0.17 (-)	–	-0.04	0.45*	0.52* (+)	0.50*	0.42* (-)
	Left	–	–	-0.27	-0.07 (+)	0.05	0.11 (+)	0.31	0.36 (+)
	Right	0.34	-0.26 (-)	–	0.15	0.19	0.22 (+)	0.44*	0.52 (+)
	Hor./Vert.	0.33	-0.25 (-)	0.30	0.52*(+)	0.66*	0.79* (+)	0.71*	0.67* (-)
Ours	Up	-0.05	0.03 (+)	0.17	0.30 (+)	0.30	0.46* (+)	0.44*	0.28 (-)
	Down	0.11	-0.08 (-)	–	-0.11	0.26	0.44* (+)	0.31	0.37* (+)
	Left	–	–	-0.15	-0.10 (-)	0.25	0.13 (-)	0.41*	0.37* (-)
	Right	0.30	-0.24 (-)	–	0.05	0.06	0.06	0.30	0.33 (+)
	Hor./Vert.	0.23	-0.22(-)	0.28	0.37* (+)	0.61*	0.73* (+)	0.59*	0.56* (-)

Table 11: Spearman correlations between concept distributions by humans and vision-and-language models. Results are reported both per-concept and per-direction, i.e., aggregating ‘up’ and ‘down’ into ‘vertical’ and ‘left’ and ‘right’ into ‘horizontal’. Values in the ‘R’ columns refer to the *regular* prompting condition, while ‘A’ indicates *analogy* prompting. The signs in brackets signal whether analogy prompting results in an improved correlation with respect to regular prompting (+) or not (-). Asterisks mark statistical significance ($p < 0.05$).

		Llama-70B		Llama-70B-Inst			R1-Distill-Llama-70B	
		R	A	R	A	Z	R	A
Richardson	Concept ^T	0.51	0.40 (-)	0.41	0.48 (+)	0.36 (-)	0.53	0.58 (+)
	Concept ^P	0.44	0.51 (+)	0.60	0.63 (+)	0.60 (=)	0.69	0.63 (-)
	Direction ^T	0.73	0.64 (-)	0.65	0.72 (+)	0.53 (-)	0.83	0.87 (+)
	Direction ^P	0.60	0.70 (+)	0.83	0.90 (+)	0.80 (-)	0.93	0.90 (-)
Ours	Concept ^T	0.50	0.38 (-)	0.33	0.37 (+)	0.33 (=)	0.45	0.41 (-)
	Concept ^P	0.34	0.47 (+)	0.46	0.49 (+)	0.42 (-)	0.49	0.45 (-)
	Direction ^T	0.67	0.71 (+)	0.58	0.72 (+)	0.67 (+)	0.77	0.73 (-)
	Direction ^P	0.52	0.70 (+)	0.70	0.77 (+)	0.67 (-)	0.73	0.70 (-)

Table 12: Weighted F1 scores between human and the open-source models’ concept preferences. The first two rows report results considering all four concepts (up, down, left, right) for textual (T), and (\uparrow , \downarrow , \leftarrow , \rightarrow) for pseudo-visual (P), while the last two rows aggregating them into two main directions (horizontal and vertical). Values in the ‘R’ column refer to the *regular* prompting condition, while ‘A’ indicates *analogy* prompting, and ‘Z’ indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results improved F1 score with respect to regular prompting (+), remained the same (=), or didn’t improve (-).

		GPT-3.5		GPT-4o		GPT-4o-Mini		GPT-o1-Preview		
		R	A	R	A	R	A	R	A	Z
Richardson	Concept ^T	0.60	0.63 (+)	0.40	0.45 (+)	0.45	0.40 (-)	0.35	0.49 (+)	0.40 (+)
	Concept ^P	0.53	0.61 (+)	0.58	0.63 (+)	0.64	0.63 (-)	0.64	0.67 (+)	0.67 (+)
	Direction ^T	0.87	0.90 (+)	0.76	0.76 (=)	0.55	0.68 (+)	0.55	0.64 (+)	0.60 (+)
	Direction ^P	0.80	0.90 (+)	0.90	0.87 (-)	0.90	0.76 (-)	0.80	0.90 (+)	0.83 (+)
Ours	Concept ^T	0.46	0.49 (+)	0.33	0.29 (-)	0.46	0.35 (-)	0.35	0.44 (+)	0.35 (=)
	Concept ^P	0.35	0.50 (+)	0.41	0.42 (+)	0.48	0.45 (-)	0.50	0.46 (-)	0.46 (-)
	Direction ^T	0.80	0.63 (-)	0.62	0.55 (-)	0.62	0.61 (-)	0.62	0.71 (+)	0.67 (+)
	Direction ^P	0.67	0.76 (+)	0.77	0.67 (-)	0.76	0.69 (-)	0.73	0.70 (-)	0.70 (-)

Table 13: Weighted F1 scores between human and GPT’s concept preferences. The first two rows report results considering all four concepts (up, down, left, right) for textual (T), and (\uparrow , \downarrow , \leftarrow , \rightarrow) for pseudo-visual (P), while the last two rows aggregating them into two main directions (horizontal and vertical). Values in the ‘R’ column refer to the *regular* prompting condition, while ‘A’ indicates *analogy* prompting, and ‘Z’ indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results improved F1 score with respect to regular prompting (+), remained the same (=), or didn’t improve (-).

		Molmo-7B		Molmo-72B		Qwen2-VL-7B		Qwen2-VL-72B	
		R	A	R	A	R	A	R	A
Rich.	Concept	0.30	0.15 (-)	0.05	0.15 (+)	0.18	0.34 (+)	0.41	0.51 (+)
	Direction	0.39	0.25 (+)	0.33	0.68 (+)	0.60	0.55 (-)	0.60	0.90 (+)
Ours	Concept	0.20	0.12 (-)	0.05	0.16 (+)	0.23	0.22 (-)	0.35	0.38 (+)
	Direction	0.44	0.32 (-)	0.40	0.61 (+)	0.60	0.62 (+)	0.52	0.69 (+)

Table 14: Weighted F1 scores between VLM and human concept preferences from both Richardson’s and our dataset. Results are reported for both concept preferences and direction preferences. Values in the ‘R’ columns refer to the *regular* prompting condition, while ‘A’ indicates *analogy* prompting. The signs in brackets signal whether analogy prompting results in an improved F1 score with respect to regular prompting (+) or not (-).

Model	% Inv. Resp. ↓			# AWs w/ Inv. Resp. ↓			# Removed AWs ↓		
	R	A	Z	R	A	Z	R	A	Z
Llama-70B ^T	9.44	13.89	–	14	18	–	0	0	–
Llama-70B ^P	2.50	9.72	–	10	14	–	0	0	–
Llama-70B-Inst ^T	0	0.69	1.94	0	2	9	0	0	0
Llama-70B-Inst ^P	0	0.28	6.94	0	2	16	0	0	0
R1-Distill-Llama-70B ^T	0	0.28	–	0	1	–	0	0	–
R1-Distill-Llama-70B ^P	0.14	0.69	–	1	2	–	0	0	–
GPT-3.5 ^T	0.14	1.53	–	1	3	–	0	0	–
GPT-3.5 ^P	0	0.42	–	0	1	–	0	0	–
GPT-4o ^T	2.22	0	–	1	0	–	0	0	–
GPT-4o ^P	0	0	–	0	0	–	0	0	–
GPT-4o-Mini ^T	0	0	–	0	0	–	0	0	–
GPT-4o-Mini ^P	0	0	–	0	0	–	0	0	–
GPT-o1-Preview ^T	0	0	0	1	0	0	0	0	0
GPT-o1-Preview ^P	0	0	0	1	0	0	0	0	0
Molmo-7B ^V	17	0	–	5	0	–	5	0	–
Molmo-72B ^V	0	0	–	0	0	–	0	0	–
Qwen2-VL-7B ^V	0	0	–	0	0	–	0	0	–
Qwen2-VL-72B ^V	0	0	–	0	0	–	0	0	–

Table 15: Overview of invalid responses in the Regular, Analogy, and Zero-shot prompting conditions, for the textual (*T*), pseudo-visual (*P*), and visual (*V*) conditions. The first column contains the overall percentage of invalid responses, the second the number of action words for which at least one invalid response was generated, and the last the number of action words that were removed because none of the generated answers was valid. A “–” indicates that the model was not evaluated under the corresponding prompting condition.

Author Index

- Al-Azary, Hamad, 37
Arndt-Lappe, Sabine, 1
- Bambini, Valentina, 37
Bansal, Swayam, 81
Barattieri Di San Pietro, Chiara, 37
Basile, Valerio, 22
Basili, Roberto, 22
Bavaresco, Anna, 108
Bel-Enguix, Gemma, 45
Belosevic, Milena, 1
Blessing, Andre, 1
Bosco, Cristina, 22
Bueno, Ivo, 108
- Croce, Danilo, 22
Cunha, João Miguel, 108
- Deeg, Tana, 1
- Eichel, Annerose, 1
- Flek, Lucie, 92
Frau, Federico, 37
- Gupta, Ashray, 57
- Ibrohim, Muhammad Okky, 22
- Joseph, Rohan, 57
- Kankhara, Mounil Hiren, 81
Kapoor, Rishit, 81
Kumar, Bagesh, 81
- Lal, Yash Kumar, 92
- M, Himesh Reddy, 81
Maheshwari, Mridul, 81
Mangiaterra, Veronica, 37
- Ojeda-Trueba, Sergio-Luis, 45
- Rai, Sunny, 57
Reddy, G.Sai Linisha, 81
Reimann, Sebastian, 10
- Sadiri Javadi, Vahid, 92
Scheffler, Tatjana, 10
Schulte Im Walde, Sabine, 1
Sierra, Gerardo, 45
Sánchez-Montero, Alec, 45
- Trippas, Johanne, 92
- Wicke, Philipp, 108