Using Analogical Reasoning to Prompt LLMs for their Intuitions of Abstract Spatial Schemas

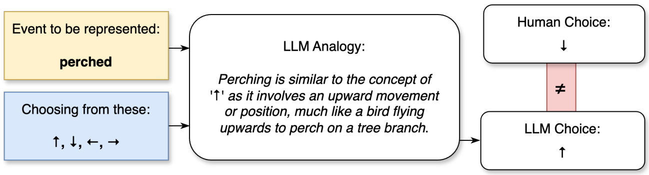
Philipp Wicke[1,3] , **Lea Hirlimann**[1] , **João Miguel Cunha**[2]
[1]LMU Munich
[2]University of Coimbra, CISUC/LASI, DEI
[3]Munich Center for Machine Learning (MCML)
pwicke@cis.lmu.de

Abstract

Abstract notions are often comprehended through analogies, wherein there exists correspondence or partial similarity with more concrete concepts. A fundamental aspect of human cognition involves synthesising embodied experiences into spatial schemas, which profoundly influence conceptualisation and underlie language acquisition. Recent studies have demonstrated that Large Language Models (LLMs) exhibit certain spatial intuitions akin to human language. For instance, both humans and LLMs tend to associate ↑ with `hope` more readily than with `warn`. However, the nuanced partial similarities between concrete (e.g., ↑) and abstract (e.g., hope) concepts, remain insufficiently explored. Therefore, we propose a novel methodology utilising analogical reasoning to elucidate these associations and examine whether LLMs adjust their associations in response to analogy-prompts. We find that analogy-prompting is slightly increasing agreement with human choices and the answers given by models include valid explanations supported by analogies, even when in disagreement with human results[1].

1 Introduction

In recent years, the development of LLMs has led to remarkable advancements in natural language processing (NLP) and many other fields. These models, trained on vast amounts of text data, exhibit impressive language understanding capabilities and demonstrate a variety of emergent abilities similar to human reasoning capabilities [Wei *et al.*, 2022a]. However, amidst their success, many fundamental questions remain: whether these abilities are accurately measured [Schaeffer *et al.*, 2024; Katzir, 2023] and whether their internal conceptualisations are comparable to human concept formation. Specifically, the conceptualisation of abstract notions is an understudied but relevant measure that can unveil more profound insights into the inner workings of LLMs.



Perching is similar to the concept of '↑' as it involves an upward movement or position, much like a bird flying upwards to perch on a tree branch.

Figure 1: Example of an instance in which the LLM response disagrees with human judgement given the following prompt: Given the concepts: ↑, ↓, ←, →. For the CONCEPT that best represent the event "perched", what would you choose? Explain the analogy (ANALOGY:), then provide one concept (↑, ↓, ←, →) (CONCEPT:)

Abstract notions are often comprehended through analogies, wherein there exists correspondence or partial similarity with more concrete concepts. This analogical reasoning is a fundamental aspect of human cognition [Holyoak *et al.*, 2001; Hofstadter, 2001]. Conceptualisation involves synthesising embodied experiences into spatial (or image) schemas, which influence conceptualisation and underlie human language acquisition [Geeraerts and Cuyckens, 2007].

Recent research has delved into this inquiry, focusing on the non-embodiment of LLMs and their capacity to capture implicit human intuitions regarding abstract concepts [Wicke, 2023; Wicke and Wachowiak, 2024; Wachowiak and Gromann, 2023]. While these studies offer initial results and evidence regarding the commonalities and differences between LLM abilities and human cognition, they fall short of providing deeper insights into why LLMs sometimes completely oppose our thinking strategies and how this can be explained.

Drawing inspiration from analogical abstraction, we propose a novel methodology to prompt LLMs for their intuitions of abstract spatial schemas. By leveraging analogical reasoning, we seek to bridge the gap between human intuitions of spatial concepts and computational representations encoded within LLMs. Our approach offers the following: **(I)** a novel methodology that evokes LLMs to use analogical reasoning strategies to solve a task; **(II)** the quantitative analysis of the proposed methodology, compared to human and baseline results; **(III)** a qualitative analysis of the individual differences within the intuitions of spatial schemas.

[1] All results at: https://cisnlp.github.io/analogical_reasoning/

J.M.C's work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020; and supported by the Portuguese PRR through project C645008882-00000055, Center for Responsible AI.

2 Related Works

LLMs, trained on extensive human-generated data, often yield output results similar to human answers, indicative of human tendencies and preferences [Hagendorff *et al.*, 2022; Dasgupta *et al.*, 2022]. Recently, this has motivated psychology studies that aim to gain insight into human cognition through LLMs [Dillion *et al.*, 2023; Harding *et al.*, 2023; Aher *et al.*, 2023]. In a study by [Dillion *et al.*, 2023], they investigate if LLMs can replace humans in moral judgement tasks. They find a strong correlation between human and LLM responses but note demographic biases, limiting diversity capture. Critics argue LLMs lack sufficient evaluation to replace humans in experiments [Harding *et al.*, 2023], and LLMs' accuracy is questioned due to reliance on static training data, unable to adapt to evolving human judgements.

Despite contrasting views psychological experiments explore if LLMs can emulate humans, driven by interest in comparing their performance on a behavioral level to human performance [Lieto, 2021]. In psycholinguistics, such experiments highlight LLMs' language processing abilities [Houghton *et al.*, 2023]. They exhibit similarities to humans in discerning grammatical sentences [Dentella *et al.*, 2023] and cognitive biases [Hagendorff *et al.*, 2022]. Recent models like ChatGPT and GPT-4 minimize these biases [Hagendorff *et al.*, 2022].

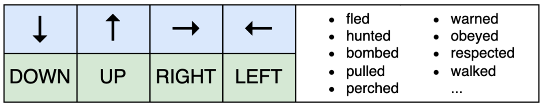
Analogical Reasoning will be the tool with which we plan to investigate the LLMs spatial conceptualisations in this work-in-progress study. [Thaler *et al.*, 2022] describes analogical reasoning as the mapping process between example solutions and new problems, their research involves testing students on their conceptualisation and knowledge of SQL, in regard to example solutions they are shown. In a survey on anti-unification, [Cerna and Kutsia, 2023] name analogical reasoning as one of the key applications for generalisation. The Analogical Reasoning Framework (ARN) presented in [Sourati *et al.*, 2023] focuses on the ability of LLMs to pick the correct analogy on a narrative level. Enhanced through few shot and Chain-of-Thought (CoT) prompting, the models employ analogical reasoning to decide on the correct answer given a query narrative. [Jiang *et al.*, 2024] presents the results and challenges of various models on the BRAINTEASER(S) benchmark, which was introduced to test LLMs on their lateral reasoning skills. The authors regard analogical reasoning as a future direction for challenge improvement.

2.1 Background

In a recent study [Wicke and Wachowiak, 2024], LLMs' spatial intuitions were investigated by replicating three psycholinguistic experiments [Gibbs Jr *et al.*, 1994; Beitel *et al.*, 2001; Richardson *et al.*, 2001]. The study [Gibbs Jr *et al.*, 1994] examines how bodily experiences influence our understanding of "stand," identifying related image schemas, evaluating similarity judgements, and exploring alternative interpretations. [Beitel *et al.*, 2001] replicates these experiments focusing on the preposition "on." The third experiment [Richardson *et al.*, 2001] takes a slightly different approach, because it asks participants to choose an abstract representation (↑, ↓, →, ←) for a list of action words that are abstract (e.g. hope) or concrete (e.g. pull). Overall, model responses often align with human responses, particularly in larger models, indicating a correlation, albeit with discrepancies in certain image schemas. These insights suggest that models reflect spatial primitive intuitions, potentially attributed to their ability to model words, their contextual use, and relation to schema definitions. However, the study highlights limitations in explanatory value, especially regarding cases where model answers diverge significantly from human responses.

This workshop paper aims to address the explanatory gap observed in this research on conceptualisations by leveraging analogical reasoning within LLMs to provide further explanations. In particular, we focus on the third experiment (see Fig. 2). Producing new results using analogical reasoning may also produce different agreement between model and human response. Previous studies have demonstrated that CoT prompting LLMs to provide step-by-step instructions yields improved performance across various tasks [Zhang *et al.*, 2022; Wei *et al.*, 2022b]. Similar to the *let's think step-by-step* instruction provided to evoke CoT reasoning, we propose an *explain the analogy* instruction as a new methodology wherein the model is tasked to provide explanatory analogies.

The approach makes two assumptions: linking abstract symbols (e.g., ↑, ↓, →, ←) with concepts (e.g., hope, perch, pull) requires finding similarities, and misalignment in analogical abstraction may lead to differences with human choices. The study aims to: i) provide insights through analogies, and ii) assess if prompting LLMs to generate analogies, like CoT prompts, improves alignment with humans.



↓	↑	→	←
DOWN	UP	RIGHT	LEFT

- fled
- hunted
- bombed
- pulled
- perched

- warned
- obeyed
- respected
- walked
- ...

Figure 2: Two LLMs (GPT-3.5 and GPT-4) are asked to pick either from the Unicode arrows or the textual options ('DOWN', 'UP' etc.) on the left to best represent the concepts on the right (e.g. *fled*).

3 Experimental Design

To summarize, we replicate one of the three experiments conducted by [Wicke and Wachowiak, 2024] to explore analogical reasoning's potential for providing explanations and enhancing alignment. Given our focus on testing this aspect, we designate our effort as work in progress and solely execute the third experiment. Notably, we restrict our selection to models demonstrating the highest alignment, following [Wicke and Wachowiak, 2024], to ensure a more dependable assessment of the analogical reasoning effect. Furthermore, we opt to exclude vision language models (VLMs) (e.g., GPT-4 vision) as in the third condition of the original experiment due to its cost, deeming it unnecessary for a preliminary investigation.

3.1 Methodology

Selected Models and Items The highest alignment between LLMs and humans for the third experiment was measured for *OpenAI* models GPT-3.5 and GPT-4. Specifically, we choose the `gpt-3.5-turbo-instruct` and `gpt-4`

Choice	GPT-3.5	GPT-3.5^{A}	GPT-4	GPT-4^{A}
`Up`	0.63	0.57 (-)	0.66	0.65 (-)
`Down`	0.31	0.44 (+)	0.33	0.41 (+)
`Left`	0.37	0.30 (-)	0.24	0.18 (-)
`Right`	0.56	0.58 (+)	0.41	0.65 (+)
↑	0.49	0.68 (+)	0.70	0.64 (-)
↓	0.42	0.50 (+)	0.49	0.60 (+)
←	0.31	0.34 (+)	0.18	0.38 (+)
→	0.56	0.63 (+)	0.69	0.60 (-)
Average	0.46	**0.51 (+)**	0.46	**0.51 (+)**

Table 1: Spearman correlation between model answers and human answers. Average **correlation is higher** for analogy-prompting ($^{\mathbf{A}}$).

endpoints in the Chat Completions API. The temperature parameter was set to 0 and the max token number to 80. Total inference cost were \$11.88. Reproducing the third experiment includes the list of 30 verbs (see examples in Figure 2 and the original reseach [Richardson *et al.*, 2001; Wicke and Wachowiak, 2024]). We exclude the third condition requiring VLMs, because this initial study is text-only.

Selected Prompts and Modalities We adopt the same prompts used by [Wicke and Wachowiak, 2024], but we add the prompt for an analogy. Due to the increased performance of LLMs with an instruction-following objective [Ouyang *et al.*, 2022], we can make use of certain instructive patterns. Here, we include the 'CONCEPT:' and 'ANALOGY:' tags to enforce a specific output format. As presented in [Wicke and Wachowiak, 2024], the models are sensitive towards textual or pseudo-visual options (words or unicode, e.g. UP or ↑) as well as order of options [Pezeshkpour and Hruschka, 2023].

The resulting prompt is assessed for all 24 permutations of the option order (24 ways of ordering the 4 directions), for both modalities (word/unicode), all 30 items and both models resulting in a total of 2880 model inferences. The prompt:

> Given the concepts: *\$options\$*. For the CONCEPT that best represent the event *\$item\$*, what would you choose? Explain the analogy (ANALOGY:), then provide one concept (*\$options\$*) (CONCEPT:)

See Fig. 1 for an example of the prompt with options and item. In this preliminary study, we simply adopted an augmented prompt used in [Wicke and Wachowiak, 2024]. Considering our aim to validate the concept, we refrained from conducting extensive prompt optimization, which typically mandates the LLM to produce responses in predetermined formats (e.g., analogies following "ANALOGY" and choices following "CONCEPT"). Instead, our analysis involved manual data processing to ensure the validity of responses.

3.2 Analysis and Results

Data processing Following analysis of model responses, instances where no valid choice could be identified automatically or multiple choices were provided (e.g., "CONCEPT: left or right") were excluded and we manually added the chosen concept if present in the models' analogy. Processing yielded 2846 valid choices of 2880 inferences. Only 34 inferences lacked valid choices, yielding an error rate of 1.18%, which we deem negligible for our preliminary findings.

Quantitative Results In the analysis of all 24 permutations of options, a distribution of responses from each model is derived for every item. These distributions are subsequently correlated with human choices obtained from the original psycholinguistic studies and the baselines articulated in [Wicke and Wachowiak, 2024]. The spearman correlations assessing agreement with human choices are contrasted with the baseline (without analogy-prompting) in Table 1. This quantitative assessment elucidates that for certain selections (e.g., Up / ↑), the utilisation of analogy-prompting neither elevates nor diminishes correlations. On average however, when compared to baseline outcomes, an overall enhancement in agreement is discernible: GPT-3.5/GPT-4 (0.46/0.46) vs. analogical prompts with GPT-3.5^{A}/GPT-4^{A} (0.51/0.51).

Qualitative Results The qualitative analysis examined items with deviations between model and human choices. We analysed the analogies accompanying these differences.

Comparing human results to those of the models revealed two scenarios: (i) words where all four model setups strongly agreed on a concept different from humans' choice (e.g., *perched* and *obeyed*); and (ii) words where humans identified two potential concepts while the model setups consistently leaned toward one (e.g., *flew*, *floated*, and *rested*). Some analogies involved a change in perspective. For instance, regarding *perched*, most results used the analogy of a bird perching high up to justify *up*, with some models even linking it to sitting humans, likely leading humans to choose *down*. Still, the models chose *up* because the bird is sitting high up.

We noticed GPT-4 provides more complex analogies. For instance, GPT-3.5 relates *flew* with *up* simply as the direction of flight, while GPT-4 resorts to a bird/plane analogy. Similarly, for *floated* most explanations rely on *defying gravity* but GPT-4 provides a *balloon* analogy. A common theme is associating directions with sentiment: *down* often implies danger and negativity; *left* suggests the wrong path or regression; *right* signifies the correct path, progress, and safety. Additionally, arrows represent more than directions, such as the subject (*self*). For instance, *pulled* in Unicode approaches is linked to ←, explained as *moving something towards oneself*, while text approaches favour *down*, explained by a gravity analogy. We will release all data upon acceptance (more at[1]).

4 Conclusion

This preliminary study suggests that analogical reasoning can influence alignment between humans and LLMs in representation. It also offers initial insights into differences between human and LLM conceptualisation of concrete and abstract concepts by means of a qualitative analysis.

Limitations and Future Work This study is limited in the number of experiments, models, and paper length, indicating the need for further investigation. Specifically, we aim to gather human analogical reasoning of the same task for comparison with LLMs more comprehensively. Additionally, we plan to incorporate VLMs and conduct more experiments, including refining prompt engineering with no error tolerance.

Acknowledgments & Ethical Statement

There are no ethical issues.

References

[Aher *et al.*, 2023] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *ICML*, pages 337–371. PMLR, 2023.

[Beitel *et al.*, 2001] Dinara A Beitel, Raymond W Gibbs Jr, and Paul Sanders. The embodied approach to the polysemy of the spatial preposition on. In *Polysemy in cognitive linguistics*, pages 241–260. John Benjamins, 2001.

[Cerna and Kutsia, 2023] David M Cerna and Temur Kutsia. Anti-unification and generalization: a survey. In *IJCAI*, pages 6563–6573, 2023.

[Dasgupta *et al.*, 2022] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *preprint arXiv:2207.07051*, 2022. DeepMind.

[Dentella *et al.*, 2023] Vittoria Dentella, Fritz Günther, and Evelina Leivada. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51), 2023.

[Dillion *et al.*, 2023] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023.

[Geeraerts and Cuyckens, 2007] Dirk Geeraerts and Hubert Cuyckens. *The Oxford handbook of cognitive linguistics*. OUP USA, 2007.

[Gibbs Jr *et al.*, 1994] Raymond W Gibbs Jr, Dinara A Beitel, Michael Harrington, and Paul E Sanders. Taking a stand on the meanings of stand: Bodily experience as motivation for polysemy. *Journal of Semantics*, 11(4):231–251, 1994.

[Hagendorff *et al.*, 2022] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Thinking fast and slow in large language models. *arXiv preprint arXiv:2212.05206*, 2022.

[Harding *et al.*, 2023] Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. Ai language models cannot replace human research participants. *AI & SOCIETY*, pages 1–3, 2023.

[Hofstadter, 2001] Douglas R Hofstadter. Epilogue: Analogy as the core of cognition. pages 499–538, 2001.

[Holyoak *et al.*, 2001] K Holyoak, Dedre Gentner, and B Kokinov. The place of analogy in cognition. *The analogical mind: Perspectives from cognitive science*, 119, 2001.

[Houghton *et al.*, 2023] Conor Houghton, Nina Kazanina, and Priyanka Sukumaran. Beyond the limitations of any imaginable mechanism: Large language models and psycholinguistics. *The Behavioral and brain sciences*, 2023.

[Jiang *et al.*, 2024] Yifan Jiang, Filip Ilievski, and Kaixin Ma. Semeval-2024 task 9: Brainteaser: A novel task defying common sense. *preprint arXiv:2404.16068*, 2024.

[Katzir, 2023] Roni Katzir. Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Manuscript. Tel Aviv University*, 2023.

[Lieto, 2021] Antonio Lieto. *Cognitive design for artificial minds*. Routledge, 2021.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[Pezeshkpour and Hruschka, 2023] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[Richardson *et al.*, 2001] Daniel C Richardson, Michael J Spivey, Shimon Edelman, and Adam J Naples. "language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 23, 2001.

[Schaeffer *et al.*, 2024] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *NeurIPS*, 36, 2024.

[Sourati *et al.*, 2023] Zhivar Sourati, Filip Ilievski, and Pia Sommerauer. Arn: A comprehensive framework and dataset for analogical reasoning on narratives. *arXiv preprint arXiv:2310.00996*, 2023.

[Thaler *et al.*, 2022] Anna Magdalena Thaler, Antonija Mitrovic, and Ute Schmid. Worked examples as application of analogical reasoning in intelligent tutoring and their effects on sql competencies. 2022.

[Wachowiak and Gromann, 2023] Lennart Wachowiak and Dagmar Gromann. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proc. of ACL'23*, pages 1018–1032, 2023.

[Wei *et al.*, 2022a] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[Wei *et al.*, 2022b] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.

[Wicke and Wachowiak, 2024] Philipp Wicke and Lennart Wachowiak. Exploring spatial schema intuitions in large language and vision models. In *ACl'24 Findings*, 2024.

[Wicke, 2023] Philipp Wicke. Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models. In *ACl'23 Findings*, pages 4899–4913, 2023.

[Zhang *et al.*, 2022] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2022.