To Know or Not To Know? Analyzing Self-Consistency of Large Language Models under Ambiguity

Anonymous ACL submission

Abstract

One of the major aspects contributing to the striking performance of large language models (LLMs) is the vast amount of factual knowledge accumulated during pre-training. Yet, many LLMs suffer from self-inconsistency, which raises doubts about their trustworthiness and reliability. In this paper, we focus on entity type ambiguity and analyze current state-of-the-art LLMs for their proficiency and consistency in applying their factual knowledge when prompted for entities under ambiguity. To do so, we propose an evaluation protocol that disentangles knowing from applying knowledge, and test state-of-the-art LLMs on 49 entities. Our experiments reveal that LLMs perform poorly with ambiguous prompts, achieving only 80% accuracy. Our results further demonstrate systematic discrepancies in LLM behavior and their failure to consistently apply information, indicating that the models can exhibit knowledge without being able to utilize it, significant biases for preferred readings, as well as self-inconsistencies. Our study highlights the importance of handling entity ambiguity in future for more trustworthy LLMs.

1 Introduction

Large language models (LLMs) have recently demonstrated remarkable performance in a variety of natural language processing tasks (OpenAI, 2024; Meta, 2024; Touvron et al., 2023), also largely due to the extensive factual knowledge they accumulate during pre-training. LLMs frequently produce unreliable responses, for example when externally retrieved knowledge conflicts with internal parametric knowledge (Xie et al., 2024; Pan et al., 2023) or when models are exposed to misinformation during pretraining (Zhao et al., 2024; Chen et al., 2023). Here, we identify entity ambiguity as a source of unreliability. Such conflicts, especially the latter, often lead to inconsistencies in model responses, reducing LLMs trustworthiness (Sun et al., 2024; Litschko et al., 2023). A crucial factor in building trust in models is their capacity to generate consistent and reliable outputs—especially in light of ambiguity—and, being consistent with their internal knowledge (Li et al., 2024; Zhao et al., 2024). Our work is similar to KoLA (Yu et al., 2024), which is a large-scale quantitative benchmark used to evaluate how well LLMs can apply their world knowledge. In contrast, we focus on on an *in-depth qualitative analysis* to understand model behaviour under ambiguity.

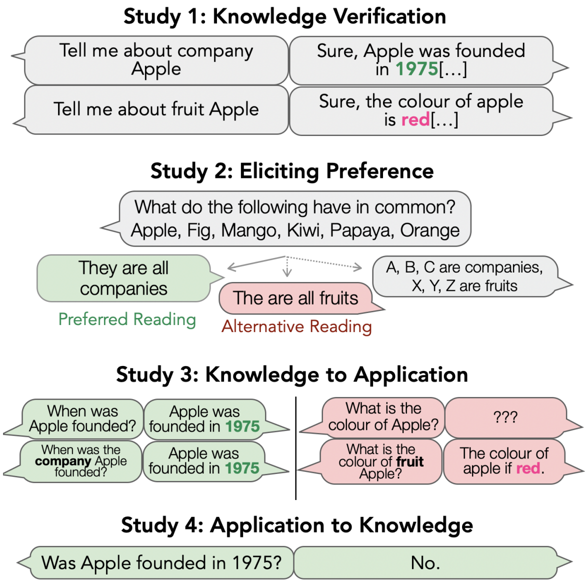

Figure 1: We focus on entity type ambiguity to study self-consistency of LLMs. Overview of our 4 studies.

Our study examines the self-consistency[1] of state-of-the-art LLMs—i.e., how well they align with their intrinsic knowledge while avoiding contradictory responses (Chen et al., 2024)—by evaluating their reasoning abilities in contexts involving *entity type ambiguity*, a commonly encountered challenge for LLMs (Parcalabescu and Frank, 2024; Kim et al., 2024; Parrish et al., 2022). Impor-

[1]This paper examines consistency in "internal knowledge retrieval" on straightforward, fact-intensive tasks that do not necessitate CoT prompting as in, e.g., Wang et al. (2023).

Entity Type	List of Entities	Entity Property
animal	Jaguar, Puma, Penguin, Greyhound, Dove, Fox, Lynx	speed
fruit	Apple, Fig, Mango, Kiwi, Papaya, Orange	color
myth	Amazon, Nike, Midas, Mars, Hermes, Hyperion, Vulcan, Pegasus	gender
person	Ford, Disney, Tesla, Boeing, Dell, Ferrero, Benetton, Levi Strauss, Versace, Philips	date of birth
location	Amazon, Cisco, Montblanc, Patagonia, Hershey, Nokia, Eagle Creek, Prosper	area in m^2
abstract	Triumph, Harmony, Genesis, Vision, Pioneer, Vanguard, Zenith, Allure, Tempo, Fidelity	level of abstractness
company	*all entities listed above*	*founding year*

Table 1: Overview of ambiguous entities. We use a total of 49 entities belonging to 7 entity types. The entities are chosen such that have at least two readings: the listed *entity type* and *company*. Entity properties are chosen such that the entity type can be uniquely inferred from it.

tantly, in the scope of our study we provide an operationalization to disentangle LLM's capabilities of ***Knowing***[2] (i.e., how aware and sensitive a model is to the possible interpretations, or *readings*, of ambiguous entities), and ***Applying knowledge*** (i.e., how well a model can identify the correct reading when prompted with entity-specific questions *and* use their parametric knowledge to provide accurate responses about that entity). The overarching goal of this work is thus to study the interplay between a model's knowledge about different entity readings and their ability to infer the correct reading for a given prompt. For example, as shown in Figure 1, if a model "knows" that Apple can be a fruit and a company, to what extent can we assume that they also infer the company meaning when asked about its founding year? Knowing if an LLM can disambiguate an entity[3] allows us to minimize the number of clarification questions (Xu et al., 2019; Lee et al., 2023) and facilitate more natural conversations. Similarly, if a model responded with *"Apple was founded in 1976"* can we assume that it is self-consistent with its own answer? We systematically investigate these questions by providing a testing suite, thereby characterizing the behaviour of LLMs under entity ambiguity.

More specifically, we aim to answer the following three research questions: Assuming a model "knows" about different entity types, how well can it disambiguate them in a given prompt (**RQ1**)? Can LLMs self-verify their answers for entity-related questions, given they have successfully disambiguated it (**RQ2**)? To what extent is the ability to infer the correct entity type biased towards "preferred readings"? Can this preference be explained by entity popularity (**RQ3**)?

To this end, we analyze the behaviour of six state-of-the-art LLMs (that differ in size, type and open vs proprietary) on 49 entities (see §2): Gemma-1.1-7B-IT (Google, 2024), Mistral-7B-Instruct (Jiang et al., 2023), LlaMa-3 (Meta, 2024), Mixtral-8x7B (Jiang et al., 2024), GPT-3.5 (OpenAI, 2022) and GPT-4o (OpenAI, 2024). Our results show that, despite the seemingly simple task, LLMs fail to disambiguate and handle entities consistently.[4]

2 Methodology

To study the ability of LLMs to implicitly infer the correct entity meanings, we use a set of forty-nine entities, as shown in Table 1. All entities can be interpreted as (1) one of the six listed entity types or (2) company names. That is, each entity has *at least two entity types* and can therefore be interpreted in at least two different ways. We adopt this framework to distinguish between a preferred and an alternative reading, which allows us to investigate if the disambiguation ability of LLMs is consistent or biased across different entity types.

Our research comprises four studies (see Figure 1). Study 1 verifies knowledge possession in models; Studies 2 and 3 assess the models' abilities to *apply* this knowledge ($K \rightarrow A$); and Study 4 evaluates the knowledge possession post-application ($A \rightarrow K$). Collectively, the results of our four experiments provide us a way to gain knowledge on how LLMs treat entity level ambiguity, i.e., the mutual relationship $K \leftrightarrow A$. We discuss each study and our results in more detail next.

Study 1: Knowledge Verification (K). First, we analyze the models' *knowledge* by verifying their awareness of different entity readings. To this end, we use the prompt template *"Tell me about <entity-*

[2]Here and further, we use the term "knowing" to refer to parametric knowledge as discussed in (Mallen et al., 2023; Litschko et al., 2023).

[3]Importantly, we measure the ability to disambiguate entities empirically by comparing their question answering performance on an ambiguous question (*"What is the founding year of X?"*) against a non-ambiguous question (*"What is the founding year of company X?"*).

[4]All the prompts and model responses are provided as supplementary materials for this submission.

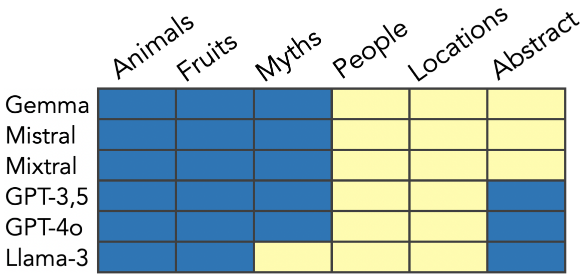

Figure 2: Preferred readings by the models for each entity type (blue for non-company, yellow for company).

type> <*entity*>*”* to manually verify that all LLMs generate meaningful output conforming to world knowledge. We cure the list of entities (see Table 1) to make sure they all pass Study 1. Apart from that, we directly ask the models whether they are aware of ambiguity (*“Can <entity> mean anything else but <entity-type>? Answer only with Yes or No.”*) - the results are provided in Appendix B.

Study 2: Eliciting Preferences (K + A). As mentioned above, each entity has been chosen such that it has at least two entity types. Intuitively, if a model has been exposed to the company Cisco far more often than the location Cisco (city in Texas), we would assume that it is biased towards the former interpretation. We refer to it as its preferred reading. To investigate if a model's behaviour is affected by a preferred reading (RQ3), i.e., if the answer correctness increases (decreases) if the question refers to a preferred (alternative) entity interpretation, we prompt LLMs with *"Group the following entities according to what they all have in common: <entities>"*, where <*entities*> refers to all members of a given category. To ensure robust results, we rephrased each prompt four times and then aggregate the model replies by majority voting. To assess the LLM output, each prompt answer was manually checked (see Appendix C for details and further discussion). In Figure 2 we show the preferred interpretation of each entity group by each model (compared to *company*). Interestingly, except for Llama-3, all LLMs display a clear entity type preference. All LLMs prefer the animal and fruit reading over the company interpretation.

Study 3: Knowledge to Application (K → A). We proceed to test the *knowledge application* ability by examining if LLMs adopt the correct reading for ambiguous entities (after knowledge of both readings is confirmed in Study 1), and whether LLMs accurately answer simple questions related to entity properties. We use the prompt template *“Provide the <entity-property> for <entity>.”* to evaluate if LLMs are capable to implicitly infer <*entity-type*>. For example, a model should infer company when prompted for founding year. We compare their performance against a non-ambiguous baseline with explicit entity hint, which serves as an upper bound: *“Provide the <entity-property> for <entity-type><entity>.”*

Study 4: Applying to Knowing (A → K). Finally, we aim to determine how consistent the models are to their own internal knowledge. For that, we manually retrieve the factual information from the model replies in Study 3 (further referred to as <*info*>) and prompt the same model back to see if they either confirm or deny the correctness of provided information. For example, the knowledge about the non-company reading of “animals” entities is checked with the prompt *"Does an animal X have <info> speed?"* (see also Table 9 in Appendix).Thus, in this setup we operate under a closed world assumption and focus only on the consistency within the model's internal knowledge, ensuring a fair comparison across models of different sizes.

3 Results and Discussion

RQ1: How well can LLMs implicitly disambiguate entity types? By design, we used entities that passed Study 1, i.e., LLMs are able to generate output that conforms to external word knowledge. We present our main results (Study 3) in Table 2. On average, LLMs are able to respond with the correct property value for 80% of all entities. Even if we use a prompt with hint so that the entity type is non-ambiguous (e.g., *“Provide the founding year for company Apple”*) LLMs reach 90.5, thus fail in ∼10% of all entities.

We observe striking differences when we break the results further down into preferred and alternative readings. For preferred readings, LLMs reach 85.4% accuracy with ambiguous prompts, and this increases to almost perfect performance in non-ambiguous prompts with hints (99%). However, the results are substantially lower for non-preferred (alternative readings), where performance drops to 74.5/85.1%. This shows a clear bias of all models to preferred readings. We further look at the correlation between model size and the amount of incorrect readings, finding remarkable trends: e.g., Gemma is the smallest and worst performing

Model	Preferred Reading		Alternative Reading		Average		
	`prop X`	`prop type X`	`prop X`	`prop type X`	`prop X`	`prop type X`	Agg.
Gemma (Google, 2024)	87.8	95.9	63.3	69.4	75.6	82.7	77.6
Mistral (Jiang et al., 2023)	77.6	**100.0**	63.3	87.8	70.5	93.9	82.2
Mixtral (Jiang et al., 2024)	77.6	**100.0**	75.5	85.7	76.6	92.9	84.8
GPT-3.5 (OpenAI, 2022)	87.8	**100.0**	75.5	77.6	81.7	88.8	85.3
GPT-4o (OpenAI, 2024)	**93.9**	**100.0**	83.7	89.8	**88.8**	**94.9**	91.9
Llama-3 (Meta, 2024)	87.8	98.0	**85.7**	**100.0**	86.8	99.0	89.9
Average	85.4	**99.0**	74.5	85.1	80.0	90.5	85.3

Table 2: Results of Study 3: Knowledge to Application (% of correct replies). "`prop`" stands for reading-specific property, "`type`" - for the corresponding entity type (see Table 1). An example of "`prop X`" prompt: *"Provide the founding year of Apple"*, an example of "`prop type X`" prompts: *"Provide the founding year of* ***company*** *Apple."*

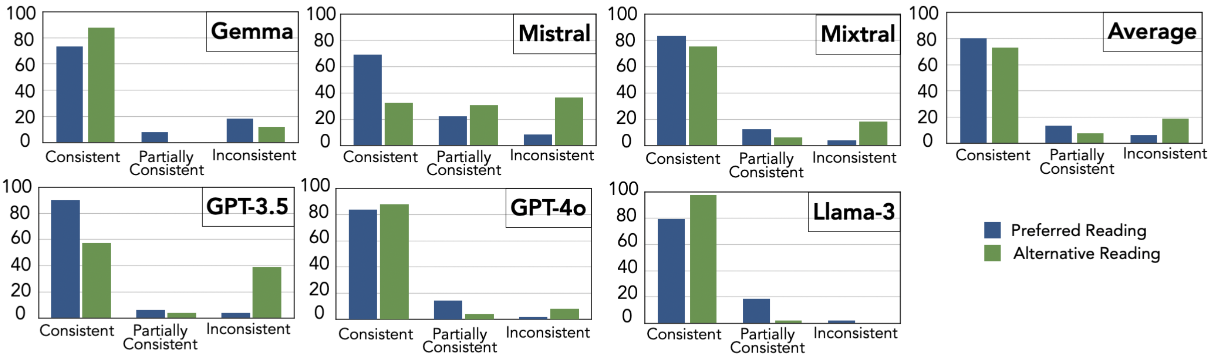

Figure 3: Results of Study 4 (% of all replies). "*Consistent*" means the model reaffirmed all knowledge provided in Study 3, "*partially consistent*" - some but not all, and "*inconsistent*" shows denial of all previous information. The exact numbers are provided in Table 5.

model with only 77.6% of correctly picked readings, while for Llama-3 and GPT-4o are the largest and best performing models with ∼90%. These results reinforces our point that the models often have difficulties in applying knowledge they actually possess, as demonstrated by Study 1, and LLMs consistency is largely affected by preferred reading (RQ3). We observed a notable correlation between the models performance on the individual entities and their popularity - we elaborate more on it in Appendix D.

RQ2: Can LLMs self-verify their answers, given that they successfully disambiguated them? We now investigate whether successful disambiguation implies that their answer can be self-verified (Study 4). Preliminary experiments revealed that closed-source LLMs yielded inconsistent results with multiple runs of the same prompt; therefore, we conducted five trials per prompt and considered the knowledge confirmed if it is confirmed in at least one run (see Table 9 for more details). As Figure 3 shows, *none* of the tested models confirmed all the knowledge provided in the previous study. On average, LLMs show a higher (partial) consistency under preferred readings. Consistent with our previous findings, Llama-3 emerged as the most self-consistent model, being consist in about 89% of its responses, while Mistral performed worst (>30% answers in alternative reading could not be self-verified).

RQ3: Does entity popularity explain the "default reading" of an LLM? We hypothesize that a model's preferred reading is influenced by its frequency in the pre-training corpus. For example, if apple mostly appears in the context of fruits, we would expect this meaning to dominate over other readings. We follow Mallen et al. (2023) and use Wikipedia popularity as a proxy for entity type frequency. In only three out of six entity types (fruit, myth, location) a higher popularity coincides with a better model performance (see Appendix A).

4 Conclusion

We find that state-of-the-art LLMs perform poorly on on prompts that require to implicitly disambiguate entity types. Furthermore, their performance is biased by a preferred reading. Finally, we find that LLMs cannot self-verify their own answers. Our results highlight the lack of self-consistency as an open challenge of current LLMs.

5 Limitations

In this study, we adopt a very generic definition of ambiguity, distinguishing between company-related and non-company-related company vs. non-company readings across different entity types. A more thorough investigation into the degrees of polysemy associated with different entity types should be included in a follow up study. Moreover, the properties of the entities might also contain a certain level of ambiguity that we are not thoroughly addressing in this work.

References

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. 2024. Two failures of self-consistency in the multi-step reasoning of LLMs. *Transactions on Machine Learning Research*.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say what you mean! large language models speak too positively about negative commonsense knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9890–9908, Toronto, Canada. Association for Computational Linguistics.

Google. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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

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

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang goo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. *Preprint*, arXiv:2404.11972.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking clarification questions to handle ambiguity in open-domain QA. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11526–11544, Singapore. Association for Computational Linguistics.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2024. Benchmarking and improving generator-validator consistency of language models. In *The Twelfth International Conference on Learning Representations*.

Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber-Genzel, and Barbara Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Singapore. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Hello gpt-4o.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539, Nusa Dua, Bali. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. *Preprint*, arXiv:2311.07466.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models. *Preprint*, arXiv:2401.05561.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. 2024. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Knowing what llms do not know: A simple yet effective self-detection method. *Preprint*, arXiv:2310.17918.

A Entity Popularity and Ambiguity

In Table 3, we present additional information about all entities utilized in our experiments. Following Mallen et al. (2023), we assess the popularity of each entity (in our context: each entity's interpretation, such as company-related and non-company-related) based on Wikipedia page views over the past nine years. In instances of ambiguity within a single interpretation (e.g., multiple companies sharing the same name, or multiple individuals with the same surname), we selected the most popular one. Furthermore, we estimated the ambiguity of each entity using its corresponding Wikipedia disambigation page, for example: `https://en.wikipedia.org/wiki/Jaguar_(disambiguation)`. Specifically, we counted the number of pages listed on the disambiguation page, providing a preliminary estimate of the number of real-world entities to which the term could refer.

Additionally, in order to evaluate correlation between the performance of the models on individual entities and the popularity of these entities, we aggregated the results of Study 3 across all models for each entity. Specifically, for each of the entity readings, we counted how many times each model selected a correct interpretation when providing response to a relevant prompt and calculated the average. For example, the performance of the models for entity *Jaguar* in its company reading was aggregated from the replies of all models to the prompt *"Provide the founding year for the company Jaguar"*.

The plots representing the entities' popularity are presented in Figure 4.

B Study 1: Further Discussion

The results from directly prompting the model to determine its awareness of ambiguity, using the prompt *"Can <entity> mean anything else but <entity-type>? Answer only with Yes or No."*, are

Type	Entity	Ambiguity	Company Reading			Non-Company Reading		
			Popularity (views)	`prop X`	`prop type X`	Popularity (views)	`prop X`	`prop type X`
Animal	Penguin	55	1,330,112	100.0	100.0	8,965,921	100.0	100.0
	Jaguar	53	7,989,902	100.0	100.0	11,939,755	0.0	100.0
	Greyhound	36	1,823,476	100.0	100.0	3,380,437	33.3	100.0
	Fox	89	3,648,500	100.0	100.0	9,301,784	100.0	100.0
	Dove	50	3,796	100.0	100.0	4,244,904	50.0	83.3
	Lynx	78	1,057,210	100.0	100.0	6,650,833	83.3	100.0
	Puma	45	4,701,402	100.0	100.0	11,554,347	83.3	100.0
	Avg	**58.0**	**2,936,343**	**100**	**100**	**8,005,426**	**64**	**98**
Fruit	Apple	49	40,325,969	100.0	100.0	10,948,070	33.3	100.0
	Fig	15	129,832	100.0	83.3	2,248,635	83.3	100.0
	Mango	43	823,939	100.0	100.0	8,713,110	100.0	100.0
	Kiwi	36	293,874	100.0	100.0	6,245,271	100.0	100.0
	Papaya	12	-	100.0	100.0	4,770,845	100.0	100.0
	Orange	103	2,007,461	100.0	100.0	7,409,145	66.7	83.3
	Avg	**43.0**	**8,716,215**	**100.0**	**97.2**	**6,722,513**	**80.6**	**97.2**
Myth. Character	Pegasus	86	1,773,226	33.3	83.3	4,853,706	100.0	100.0
	Vulcan	79	635,380	66.7	100.0	2,673,387	0.0	100.0
	Midas	38	187,394	83.3	83.3	3,687,467	100.0	100.0
	Nike	34	18,187,528	100.0	100.0	4,375,918	33.3	100.0
	Mars	134	259,189	33.3	100.0	19,365,488	66.7	100.0
	Hyperion	62	58,794	66.7	100.0	1,316,548	83.3	100.0
	Hermes	56	3,426,101	83.3	100.0	10,337,899	100.0	100.0
	Amazon	64	38,684,687	100.0	100.0	5,119,820	16.7	100.0
	Avg.	**69.1**	**7,901,537**	**70.8**	**95.8**	**6,466,279**	**62.5**	**100.0**
Person	Versace	13	7,095,079	100.0	100.0	22,180,811	100.0	66.7
	Boeing	-	10,754,848	100.0	100.0	681,877	0.0	33.3
	Ford	104	14,643,256	100.0	100.0	13,966,210	83.3	50.0
	Philips	6	5,948,052	100.0	100.0	331,229	16.7	33.3
	Levi Strauss	13	3,744,382	100.0	100.0	2,320,188	100.0	100.0
	Ferrero	4	3,447,282	100.0	100.0	409,662	66.7	66.7
	Tesla	21	23,462,104	100.0	100.0	37,395,340	83.3	83.3
	Disney	58	20,938,263	100.0	100.0	31,693,370	100.0	50.0
	Dell	22	7,310,499	100.0	100.0	3,558,086	16.7	33.3
	Benetton	5	1,864,193	100.0	100.0	378,208	50.0	50.0
	Avg.	**27.3**	**9,920,796**	**100.0**	**100.0**	**11,291,498**	**61.7**	**56.7**
Location	Cisco	26	1,738,862	100.0	100.0	-	0.0	100.0
	Prosper	10	276,714	100.0	83.3	419,461	33.3	100.0
	Patagonia	12	1,055,737	100.0	100.0	11,426,844	100.0	100.0
	Montblanc	5	1,306,077	100.0	100.0	5,671,509	100.0	100.0
	Amazon	64	38,684,687	100.0	100.0	6,509,535	33.3	100.0
	Nokia	13	11,446,036	100.0	100.0	332,572	0.0	83.3
	Hershey	24	3,929,199	100.0	100.0	1,419,873	100.0	100.0
	Eagle Creek	24	55,717	100.0	100.0	2,248	83.3	100.0
	Avg.	**22.3**	**7,311,629**	**100.0**	**97.9**	**3,683,149**	**58.3**	**95.8**
Abstract	Harmony	119	143,865	83.3	83.3	1,847,278	100.0	100.0
	Fidelity	29	3,648,171	100.0	100.0	633,474	100.0	100.0
	Allure	17	832,160	100.0	100.0	728,597	50.0	100.0
	Vision	102	29,660	100.0	100.0	1,810,577	100.0	100.0
	Genesis	141	2,809,401	50.0	100.0	6,338,641	100.0	100.0
	Tempo	59	27,507	100.0	100.0	5,416,890	66.7	100.0
	Triumph	45	351,267	100.0	100.0	1,132,962	83.3	100.0
	Vanguard	128	6,661,130	100.0	100.0	1,059,408	16.7	83.3
	Pioneer	95	1,058,945	100.0	100.0	521,227	66.7	100.0
	Zenith	64	753,374	100.0	100.0	1,602,303	100.0	100.0
	Avg	**79.9**	**1,631,548**	**93.3**	**98.3**	**2,109,136**	**78.3**	**98.3**

Table 3: Summary of entity types their characteristics: ambiguity and popularity. Following Mallen et al. (2023), we evaluate the popularity and ambiguity of each entity based on Wikipedia page views and the number of pages references to on the Wikipedia entity disambiguation page, respectively. Dashes are used in cases where Wikipedia disambiguation page is absent for the specific entity. Additionally, we provide the model performance on each entity demonstrated in Study 3, aggregated across the models.

Model	Animals	Fruits	Myths	People	Locations	Abstract	**Average**
Gemma	100.0	100.0	37.5	0.0	12.5	10.0	43.3
Mistral	100.0	83.8	75.0	10.0	75.0	90.0	72.3
Mixtral	71.4	50.0	0.0	0.0	30.0	50.0	93.1
GPT-3.5	57.1	100.0	0.0	10.0	12.5	10.0	31.6
GPT-4o	100.0	100.0	100.0	60.0	100.0	90.0	91.7
LLaMa-3	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Average	80.1	89.0	52.1	50.0	52.8	50.0	72.0

Table 4: The results of experiments with direct prompting the model about the ambiguity (*"Can <entity> mean anything else but <entity-type>? Answer only with Yes or No."*).

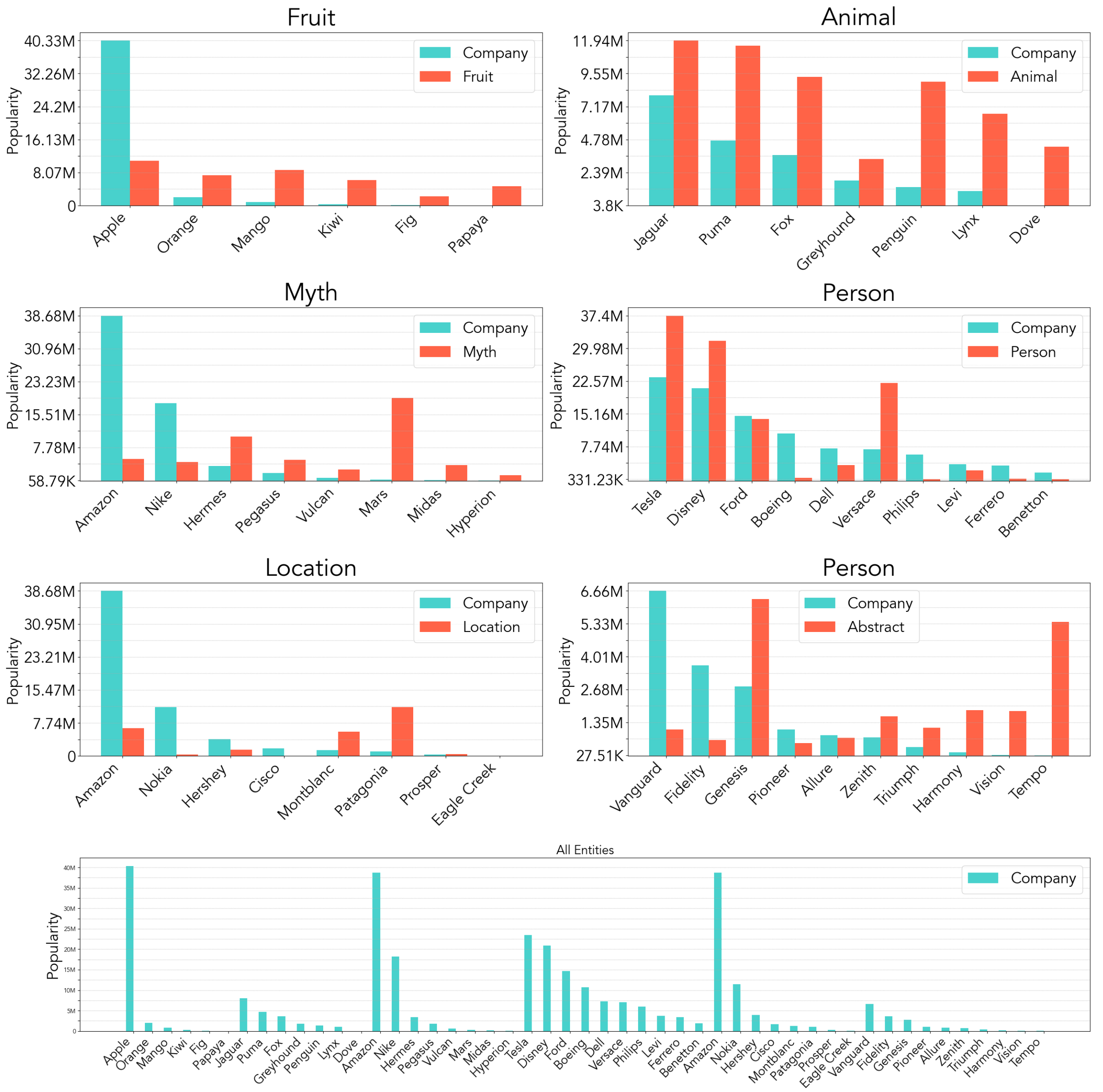

Figure 4: Popularity distribution of both, company and non-company readings of all 49 entities involved in our studies.

provided in Table 4. As it becomes clear, despite possessing knowledge about the different meanings of each entity (as proven by Study 1, see Section 3), the models tend to struggle to provide this information when asked directly. For example, Mixtral often denies other interpretations, justifying this by claiming that there is one clear meaning of the entity, although it can be used for other purposes (for example, *"No, Eagle Creek cannot mean anything else in general usage. It is primarily a geographic location, specifically a creek name occurring in various places in the United States. However, like many place names, it can be used as a proper noun in other contexts, such as brand names (e.g., Eagle Creek luggage)."*). From this observation, we make two assumptions: (1) each model may have a preferred interpretation for each entity and entity type, an hypothesis we intend to explore in Study 2, and (2) a more carefully considered experimental setup is required, rather than straightforwardly querying the model about ambiguity, which was one of the motivations behind the grouping task approach we adopted for Study 2.

C Study 2: Further Discussion

We noticed that for most of the entity groups (*Fruits*, *Locations*, *Animals*, and *People*), all analyzed models clearly prefer one reading over the other. Notably, large models like Llama-3 and Mixtral, even though ultimately grouping based on one single reading, demonstrate an understanding of entity ambiguity (e.g., Mixtral: *"All of the words you've listed are common names for either a type of animal or a brand..."*), (e.g., Mixtral: *"All of the words you've listed are common names for either a type of animal or a brand. To be more specific, they are all common names for either a type of mammal or a type of bird..."*, or LlaMa-3: *"After examining the list, I noticed that all the mentioned animals have one thing in common: they are all names of car models or brands at some point in history..."*). However, in these replies, the model still prefers one reading over the other and use it for grouping the entities; in such cases, we consider this reading to be a *preferred reading*.

The categories *Abstract* and *Myths* elicit the most diverse responses from models. This could be explained by the particularly high ambiguity associated with the entities in these categories, beyong merely "companies" and "entity types" - e.g., in the latter, "planets and moons" (e.g., Mars, Vulcan, Hyperion). Indeed, Table 3, where the ambiguity of each entity is estimated based on Wikipedia disambiguation pages (a potentially conservative measure, as not all objects a particular entity may refer to have Wikipedia pages), shows the highest average rate of ambiguity across entities within these categories: 79.9 for abstract entities and 69.1 for myths entities[5]. As a result, for these categories, models frequently mix readings, distinguishing "companies" as a separate group while also identifying non-company meanings, leading to groupings like: "Greek Gods", "Roman Gods", and "Companies". However, such responses do not clarify if the model recognizes ambiguity, as adopting both "company" and "non-company" reading in the same response could indicate either a misunderstanding of entities ambiguity (i.e., the model recognize some entities as companies and others not, despite evidence to the contrary from Study 1) or simply a preference for a specific reading for specific entities.

Interestingly, in all models, these groups consist of either only the entity *Amazon* or also include *Nike*. This disparity can be logically attributed to the significantly higher popularity of these companies compared to others on the list. For instance, the Wikipedia page for Amazon has garnered over 38 million views, and Nike's page has received more than 18 million views. In contrast, Hermes, the third most popular company within the *myths* category, has only 3 million views. This observation suggests that the model's behavior towards specific entities may be influenced by their popularity, and there appears to be a tendency for the models to favor one interpretation over another for a given entity, regardless of the prompt provided.

D Study 3: Further Discussion and Case Study

Some of the systematic errors of LLMs can be explained from observed model behavior. For example, when prompted with *"Provide the speed of jaguar,"* intended to draw an animal reading, multiple models erroneously reference the car and offer speeds of Jaguar automobiles instead; a similar error occurs with prompts about Greyhound (a well-known bus company in North America). Although this mistake can be easily explained, we still classified such cases as incorrect in our evaluation. Another notable trend is that for *people* entities in their non-company reading, all models perform better with underspecified prompts (*"What is the date of birth of X?"* compared to *"What is the date of birth of a person X?"*) and often decline to answer more detailed prompts, referring to the absence of "sensitive" information about specific

[5]Please note that these numbers are not normalized and calculated as an average of the absolute numbers for each entity.

individuals. However, that the models indeed possess this knowledge was confirmed by an additional sanity check using highly unambiguous prompts, where the entity X is presented with both first and last names (e.g., *"Provide the date of birth of Henry Ford"*), clearly indicating a preferred *people* interpretation.

The other systematic errors observed across the outputs of various models are also evident and indicate a consistent misinterpretation of the provided prompts. For instance, multiple models often respond with company foundation dates when prompted for a *date-of-birth*, where the expected behavior is for the model to recognize that it should provide the birth date of a person, thus adopting the *person* interpretation of the entity in question. For example:

- Prompt: *"Provide the date of birth for Boeing."*
- Llama-3 Response: *Boeing's founding date is July 15, 1916, when William E. Boeing and Conrad Westervelt founded Pacific Aero Products Co., which later became The Boeing Company.*

Sometimes the models even mention the name of the founder, yet still adhere to the *company* interpretation when providing the response. For example:

- Prompt: *"Provide the date of birth for Disney"*
- Mixtral Response: *"Disney is not a person, so it doesn't have a date of birth. The Disney company was founded on October 16, 1923, by Walt Disney and his brother Roy O. Disney, as the Disney Brothers Cartoon Studio."*,

This behavior is likely due to the significantly higher popularity of the *company* interpretation of individual entities compared to their recognition as *persons*. For instance, as indicated in Table 3, the popularity of the Boeing entity as a *company* is almost 1.5 times greater than that of its founder, William E. Boeing, suggesting that the *person* interpretation was *never* adopted by models when prompted to *"Provide the date of birth for Boeing"*. However, while this is a noticeable trend for many entities in the *person* group, it cannot be generalized as a universal trend for this entity type. In cases where the popularity of the *person*-entity exceeds that of the *company*-entity (such as Versace, Tesla, Disney), the performance across models for those entities is markedly better.

Model	Preferred Reading			Alternative Reading		
	Consistent	Partially Consistent	Inconsistent	Consistent	Partially Consistent	Inconsistent
	Companies Reading					
Gemma	38.8	0	18.4	32.7	0	10.2
Mistral	49.0	4.1	4.1	24.5	10.2	8.2
MiXtral	53.1	0	4.1	34.7	2.0	6.1
GPT-3.5	32.7	0	4.1	38.8	4.1	20.4
GPT-4o	36.7	0	0	57.1	4.1	2.0
LLaMa-3	53.1	0	0	46.9	0	0
	Animals Reading					
Gemma	57.1	42.9	0.0	-	-	-
Mistral	28.6	42.9	29.0	-	-	-
Mixtral	100.0	0.0	0.0	-	-	-
GPT-3.5	85.7	14.3	0.0	-	-	-
GPT-4o	86.0	0.0	14.0	-	-	-
LLaMa-3	57.1	42.9	0.0	-	-	-
	Fruits Reading					
Gemma	83.3	16.7	0.0	-	-	-
Mistral	0.0	100.0	0.0	-	-	-
Mixtral	16.7	83.3	0.0	-	-	-
GPT-3.5	66.7	33.3	0.0	-	-	-
GPT-4o	83.3	16.7	0.0	-	-	-
LLaMa-3	33.3	66.7	0.0	-	-	-
	Myths Reading					
Gemma	100	0.0	0.0	-	-	-
Mistral	100	0.0	0.0	-	-	-
Mixtral	87.5	12.5	0.0	-	-	-
GPT-3.5	100.0	0.0	0.0	-	-	-
GPT-4o	100.0	0.0	0.0	-	-	-
LLaMa-3	-	-	-	100.0	-	-
	People Reading					
Gemma	-	-	-	90.0	0.0	10.0
Mistral	-	-	-	40.0	0.0	60.0
Mixtral	-	-	-	60.0	0.0	40.0
GPT-3.5	-	-	-	60.0	0.0	40.0
GPT-4o	-	-	-	70.0	0.0	30.0
LLaMa-3	-	-	-	90.0	0.0	10.0
	Locations Reading					
Gemma	-	-	-	100.0	0.0	0.0
Mistral	-	-	-	0.0	37.0	63.0
Mixtral	-	-	-	25.0	0.0	75.0
GPT-3.5	-	-	-	37.0	0.0	63.0
GPT-4o	-	-	-	62.0	0.0	38.0
LLaMa-3	-	-	-	100.0	0.0	0.0
	Abstract Reading					
Gemma	-	-	-	100.0	0.0	0.0
Mistral	-	-	-	30.0	50.0	20.0
Mixtral	-	-	-	80.0	20.0	0
GPT-3.5	100.0	0.0	0.0	-	-	-
GPT-4o	40.0	60.0	0.0	-	-	-
LLaMa-3	70.0	20.0	10.0	-	-	-

Table 5: Results of Study 4: The interpretation of these numbers is illustrated in Figure 3. Additionally, for non-company readings, we present the results for each group separately, based on whether this reading was preferred by the model for this entity type or not.

Tell me about a company called <entity>.
Tell me about an animal <animal-entity>.
Tell me about a fruit <fruit-entity>.
Tell me about a geographic location of <location-entity>.
Tell me about a mythological character <myth-entity>.
Tell me about a person <person-entity>.
Tell me about a concept <abstract-entity>.
Can <animal-entity> mean anything else but an animal? Answer only with Yes or No.
Can <fruit-entity> mean anything else but a fruit? Answer only with Yes or No.
Can <location-entity> mean anything else but a geographic location? Answer only with Yes or No.
Can <myth-entity> mean anything else but a mythological character? Answer only with Yes or No.
Can <person-entity> mean anything else but person? Answer only with Yes or No.
Can <abstract-entity> mean anything else but an abstract concept? Answer only with Yes or No.

Table 6: Prompts used in Study 1.

Group the following according to what they all have in common: <animal-entities>. Please provide an explanation.
Group the following according to what they all have in common: <fruit-entities>. Please provide an explanation.
Group the following according to what they all have in common: <location-entities>. Please provide an explanation.
Group the following according to what they all have in common: <myth-entities>. Please provide an explanation.
Group the following according to what they all have in common: <person-entities>. Please provide an explanation.
Group the following according to what they all have in common: <abstract-entities>. Please provide an explanation.
Group the following entities: <animal-entities>. Please provide an explanation.
Group the following entities: <fruit-entities>. Please provide an explanation.
Group the following entities: <location-entities>. Please provide an explanation.
Group the following entities: <myth-entities>. Please provide an explanation.
Group the following entities: <person-entities>. Please provide an explanation.
Group the following entities: <abstract-entities>. Please provide an explanation.
Group the following entities: <animal-entities>.
Group the following entities: <fruit-entities>.
Group the following entities: <location-entities>.
Group the following entities: <myth-entities>.
Group the following entities: <person-entities>.
Group the following entities: <abstract-entities>.
What is the date of birth of <entity of type person>?

Table 7: Prompts used in Study 2.

Provide the founding year for <entity>.
Provide the founding year for the company <entity>.
Provide the speed for <animal-entity>.
Provide the speed for an animal <animal-entity>.
Provide the color for <fruit-entity>.
Provide the color for a fruit <fruit-entity>.
Provide the area in square meters for <location-entity>.
Provide the area in square meters for a geographic location of <location-entity>.
Provide the gender for <myth-entity>.
Provide the gender for a mythological character <myth-entities>.
Provide the date of birth for <person-entity>.
Provide the date of birth for a person <person-entity>.
Provide the level of abstractness for <abstract-entity>.
Provide the level of abstractness for a concept <abstract-entity>.

Table 8: Prompts used in Study 3.

Was there any company with the name (or part of the name) <entity> founded/established/launched/started in the year <retrieved-info>? Answer ONLY with Yes or No. If you cannot answer this question, answer No.
Does a concept <entity> has a <retrieved-info> of abstractness? Answer ONLY with Yes or No. If you cannot answer this question, answer No.
Does a mythological character <entity> have a <retrieved-info> gender? Answer ONLY with Yes or No. If you cannot answer this question, answer No.
Is there a geographic location <entity> with an approximate area of <retrieved-info>? Answer ONLY with Yes or No. If you cannot answer this question, answer No.
Does a fruit <entity> have <retrieved-info> color? Answer ONLY with Yes or No. If you cannot answer this question, answer No.
Does an animal <entity> have <retrieved-info> speed? Answer ONLY with Yes or No. If you cannot answer this question, answer No.
Is <retrieved-info> the date of birth of a person <entity>? Answer ONLY with Yes or No. If you cannot answer this question, answer No.

Table 9: Prompts used in Study 4.