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

Anastasiia Sedova[1,2*] **Robert Litschko**[3,4*] **Diego Frassinelli**[3]
Benjamin Roth[1,5] **Barbara Plank**[3,4]
[1]Faculty of Computer Science [2] UniVie Doctoral School Computer Science
[5] Faculty of Philological and Cultural Studies, University of Vienna, Austria
[3] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[4] Munich Center for Machine Learning (MCML), Germany
anastasiia.sedova@univie.ac.at rlitschk@cis.lmu.de

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. This paper focuses on entity type ambiguity, analyzing the proficiency and consistency of state-of-the-art LLMs in applying factual knowledge when prompted with ambiguous entities. To do so, we propose an evaluation protocol that disentangles *knowing* from *applying knowledge*, and test state-of-the-art LLMs on 49 ambiguous entities. Our experiments reveal that LLMs struggle with choosing the correct entity reading, achieving an average accuracy of only 85%, and as low as 75% with underspecified prompts. The results also reveal systematic discrepancies in LLM behavior, showing that while the models may possess knowledge, they struggle to apply it consistently, exhibit biases toward preferred readings, and display self-inconsistencies. This highlights the need to address entity ambiguity in the future for more trustworthy LLMs.[1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance in a variety of natural language processing tasks (Touvron et al., 2023; OpenAI, 2024; Meta, 2024), also largely due to the extensive factual knowledge they accumulate during pre-training. A crucial factor in building trust in models is their capacity to generate consistent and dependable outputs that align with their internal knowledge (Li et al., 2024; Yu et al., 2024).

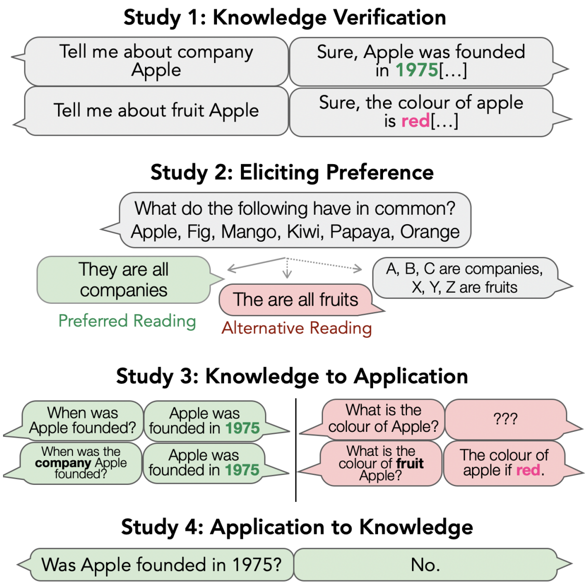

Figure 1: Overview of our four studies on LLMs' self-consistency using prompts with ambiguous entities. Colors indicate preferred (green) and alternative (red) readings implied in the query or adopted by the model.

However, LLMs frequently produce unreliable responses: e.g., when externally provided knowledge conflicts with internal parametric knowledge (Pan et al., 2023; Xie et al., 2024), or when exposed to misinformation during pre-training (Zhao et al., 2024). Such conflicts often lead to inconsistencies in model responses, reducing LLMs' trustworthiness (Litschko et al., 2023; Sun et al., 2024).

In this work, we highlight *entity ambiguity*, a frequent challenge for LLMs (Liu et al., 2023; Stengel-Eskin et al., 2024; Kim et al., 2024), as a *source of unreliability* and conduct an in-depth qualitative analysis to better understand how models behave in its presence. Specifically, we examine the self-consistency[2] of state-of-the-art LLMs –

*Equal contribution

[1]We make the code and the model outputs available: https://github.com/anasedova/ToKnow_or_NotToKnow

[2]We analyze consistency on straightforward "internal

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 17203–17217
November 12-16, 2024

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.

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. Importantly, in our study we provide an operationalization to disentangle LLM's capabilities of ***Knowing***[3] (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). 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 the model infers the company meaning when asked about the founding year of the entity? Similarly, if a model responds with *"Apple was founded in 1976"* can we assume that it remains self-consistent with its own answer? Gaining more insights on the capability of an LLM to deal with entity ambiguity has several benefits. It facilitates more natural conversations, as knowing whether an LLM can disambiguate an entity allows to minimize the number of clarification questions (Lee et al., 2023), and enhances trust in LLMs via its ability to acknowledge previously provided knowledge. The overarching goal of this work is to provide a behavioral test suite for studying how a model's knowledge of different entity readings interacts with its ability to infer and apply the correct reading for a given prompt.

Specifically, we address the following research questions: (**RQ1**) How well do LLMs resolve entity ambiguity within a given prompt context? (**RQ2**) How (much) is the ability to infer the correct entity type biased towards "preferred readings"? (**RQ3**) Can LLMs self-verify their answers for entity-related questions, given they have successfully disambiguated them?

We examine the behavior of six state-of-the-art LLMs, varying in size, type, and whether they are open-source or proprietary: Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Gemma-1.1-7B-IT (Google, 2024), Llama-3-70B (Meta, 2024), Mixtral-8x7B-Instruct-v0.1(Jiang et al., 2024), GPT-3.5 (OpenAI, 2022), and GPT-4o (OpenAI, 2024)[4]. Our results show that, despite the seemingly simple task, LLMs fail to disambiguate and handle entities consistently.

2 Methodology

To study LLMs' ability to implicitly infer correct entity meanings, we devise a behavioral test suite. We use a set of entities that can be interpreted as either (1) one of six entity types or (2) company names (Table 1), meaning each entity has *at least two* possible interpretations. We adopt this approach 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 setup 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.

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-*

knowledge retrieval" tasks that do not necessitate CoT prompting as in, for example, Wang et al. (2023).

[3]The term "knowing" refers to parametric knowledge as discussed in Mallen et al. (2023); Litschko et al. (2023).

[4]Implementation details are provided in Appendix E.

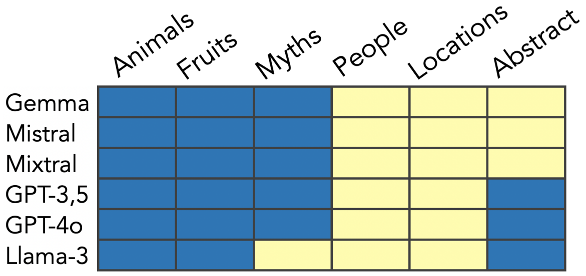

Figure 2: Models preferred readings discovered in Study 2 (blue for non-company, yellow for company; e.g., all analyzed models prefer the 'company' reading for entities from the *people* category).

type> <entity>" (see more details about prompts in Appendix C) to verify that all LLMs generate meaningful output conforming to world knowledge: i.e., when asked about "fruit Apple," the models provides some information about the fruit, and when asked about "company Apple," it should discuss a company named Apple. Next, we ensure that all entities in our experiments pass Study 1, resulting in a final list of 49 entities shown in Table 1. Additionally, we conduct Study 1a, where we directly query the models about their awareness of ambiguity (*"Can <entity> mean anything other than <entity-type>? Answer only with Yes or No."*).

Study 2: Eliciting Preferences (K + A). 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 determine each model's preferred reading (for later analysis on whether a model's behaviour is affected by it (RQ2), i.e., whether answer correctness increases (decreases) with 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. Since the grouping criteria are unspecified (meaning there is no expected *correct* entities interpretation), we assume the model will adopt its preferred reading of the given entity type when forming the grouping.

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 when answering simple questions related to entity properties. Importantly, only the correctness of the chosen reading is evaluated, not the factual accuracy of models' answers. We use the prompt template *"Provide the <entity-property> for <entity>."* to evaluate if LLMs are capable to implicitly infer *<entity-type>* (for example, 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 assess how consistent the models are with their internal knowledge. We automatically retrieve factual information (*<info>*) from the model replies in Study 3 and prompt the same model back to see if it confirms or denies the correctness of information provided. For example, the knowledge about the non-company reading of "animals" entities is checked with prompts like *"Does an animal X have <info> speed?"*. In such setup, we, again, do not verify the factual correctness of the responses, but rather operate under a closed world assumption and focus only on consistency within model's internal knowledge, ensuring fair comparison across models of different sizes.

3 Results and Discussion

RQ1: How well can LLMs implicitly disambiguate entity types? By design, all entities passed Study 1, meaning the LLMs are aware of both entity readings and can provide relevant information about the associated objects or companies. However, when directly prompted about ambiguity awareness (Study 1a), only Llama-3 recognized ambiguity in all entities, while other models struggled with this task (see Table 4 in Appendix). The highest ambiguity confirmation rates were demonstrated for *Animals* and *Fruits* entities, while for other types, models confirmed ambiguity for no more than 53% of entities on average[5].

Figure 2 summarizes the results of Study 2. For four of six entity types, all models favored the same interpretation: non-company for *Animals* and *Fruits*, and company for *People* and *Locations*; for the others, the preferred reading varies by model. Notably, the large Llama-3 and Mixtral models, despite ultimately grouping entities based on a single reading, demonstrate an understanding of their ambiguity (e.g., Mixtral: *"All of the words you've listed are common names for either a type of animal or a brand [...]"*)[6]. Another observation is

[5]Model responses in Studies 1-3 were manually reviewed; see Appendix B for more details about results evaluation.

[6]Here and throughout, the example are presented with

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 replies with correctly adopted reading). "`prop`" stands for reading-specific property, "`type`" - for the corresponding entity type (see Table 1). For example, `prop X` prompt: *"Provide the founding year of Apple"*, `prop type X` prompt: *"Provide the founding year of **company** Apple."*

that all the models tend to provide more varied responses for *Abstract* and *Myths* entities. We hypothise that this is due to their higher ambiguity, which leads to detection of other categories beyond just companies and predefined entity type (e.g., *Mars* and *Vulcan* may be seen neither as companies nor as mythological characters, but as planets, and grouped correspondingly). This hypothesis is supported by the entity ambiguity estimations based on Wikipedia disambiguation pages (the average number of entries on disambiguation pages across all entities in the group): 79.9 for *Abstract* and 69.1 for *Myth* entities, compared to, for example, 22.3 for *Locations* (see Table 3 in Appendix A). However, these mixed responses do not clarify whether the model successfully recognizes ambiguity, as including both "company" and "non-company" reading in the same entity list could indicate either a misunderstanding of entity ambiguity (i.e., the model identifies some entities as companies and others not, despite evidence from Study 1) or a preference for specific readings for certain entities.

The results of Study 3 are presented in Table 2. On average, LLMs adopt the correct reading for 85.3% of entities. However, even with non-ambiguous prompts that include hint so that the entity type is supposed to be clear (e.g., *"Provide the founding year for company Apple"*) LLMs reach 90.5%, thus fail with ∼10% of all entities. We also observe a correlation between model size and the amount of incorrect readings: Gemma, the smallest model, has the lowest performance with only an average of 77.6% correctly picked readings, while the large Llama-3 and GPT-4o models perform best at ∼90%. Overall, we can conclude that while all models have previously demonstrated knowledge about different readings of ambiguous entities, they struggle to apply this knowledge when choosing the correct reading in their responses.

omissions. For the full outputs and additional findings, please refer to Appendix D.

RQ2: How (much) is the ability to infer the correct entity type biased towards "preferred readings"? When we break the results further down into preferred and alternative readings, we observe striking differences. For preferred readings, LLMs achieve an average accuracy of 85.4% with ambiguous prompts, rising to nearly perfect performance (99%) with non-ambiguous prompts that include hints. However, the results are substantially lower for non-preferred (alternative) readings, where performance drops to 74.5%/85.1%. This shows a clear bias across all models toward preferred readings. Among the common systematic errors is the model providing company foundation dates when asked for a *date of birth*, even with non-ambiguous prompts: e.g., *"Provide the date of birth for a person Boeing,"*–Mixtral response: *"I'm sorry for any confusion, but Boeing is not a person, it's a multinational corporation [...]"*. Moreover, certain entities are consistently assigned incorrect readings by the models, correlating with their popularity, as discussed next.

We hypothesize that a model's preferred reading of particular entities is influenced by their frequencies in the pre-training corpus. For example, if *nike* mostly appears in the context of companies (as opposed to the context of a Greek goddess), 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 (see Table 3). Our experiments indeed demonstrate the strong correlation between the entity popularity and a preferred reading. For example, whenever a model created a separate "compa-

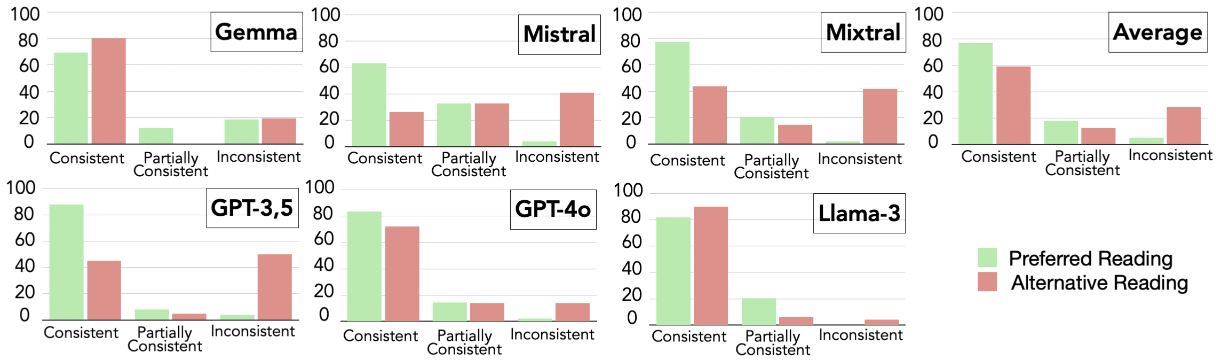

Figure 3: Results of Study 4 (% of entities). "*Consistent*" entities are those for which the model reaffirmed all provided information in Study 3. "*Partially Consistent*" entities are those where some information was reaffirmed but not all, while "*Inconsistent*" entities are those for which all previously provided information was denied. The exact numbers are provided in Appendix D (Table 5).

nies" group while categorizing *Myths* entities, the group consistently included either just *Amazon* or both *Amazon* and *Nike*; e.g., GPT-3.5: *"1. Mythological entities: [...] 2. Companies/brands: Amazon"*, or GPT-4o: *"[...] 1. Brands and Companies: - Nike - Amazon 2. Greek Mythology: [...]*. This disparity can be attributed to the much higher popularity of these companies: Amazon's Wikipedia page has over 38 million views, Nike's 18 million, while Hermes, the third most popular company in *Myths*, has only 3 million. As a result, models tend to favor *company* interpretations, regardless of the prompt. In Study 3, *all* models fail to provide the non-company information about these entities with underspecified prompts. For example, when prompted with *"Provide the gender for Amazon"*, Mixtral responded with: *"Amazon.com, Inc. is a company, and as such, it does not have a gender. [...]"*. The errors in providing people birth dates can also be attributed to the significantly higher popularity of the *company* interpretation of a majority of *People* entities. For instance, Boeing's popularity as a *company* is nearly 1.5 times that of its founder, William E. Boeing, leading models to never adopt the *person* interpretation when prompted with *"Provide the date of birth for Boeing."* Conversely, when the popularity of the *person*-entity exceeds that of the *company*-entity (e.g., Versace, Tesla, Disney), models tend to adopt correct interpretation; e.g., Gemma: *"Tesla was born on July 10, 1856"*, and Llama-3: *"You're referring to Walt Disney! Walter Elias Disney [...] was born on December 5, 1901."* These findings clearly indicate that the models exhibit a bias toward one reading of ambiguous entities over another, influenced by the popularity of the entities in those readings.

RQ3: Can LLMs self-verify their answers, given that they successfully disambiguated them? Finally, we investigate whether successful disambiguation implies that the model's answers can be self-verified - in other words, whether a model remains consistent with its previous statements (Study 4). As Figure 3 shows, *none* of the tested models confirmed all the knowledge provided in the previous study. On average, LLMs show higher consistency with preferred readings, with Llama-3 being the most self-consistent (over 80% across all readings), while Mistral performed the worst (over 30% of preferred and 60% of alternative readings unverified). To analyze the non-confirmation cases, we conducted additional experiments with a modified prompt requesting an explanation (e.g., *"Is December 5 1901 the date of birth of a person Disney? Answer with Yes or No and provide explanation."*). Surprisingly, the majority of results were like: (GPT-3.5) *"No. December 5, 1901 is not the date of birth of Walt Disney. Walt Disney was actually born on December 5, 1901."* These findings show that current state-of-the-art LLMs often fail to self-verify their answers, even when they provide the same information immediately afterward.

4 Conclusion

We find that state-of-the-art LLMs perform poorly at implicitly disambiguating entity types. Furthermore, their performance is biased by a preferred reading that is heavily influenced by the popularity of the entities. Finally, we find that LLMs cannot reliably 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.

6 Acknowledgement

We thank the anonymous reviewers for their helpful suggestions and constructive feedback. This research has been funded by the Vienna Science and Technology Fund (WWTF)[10.47379/VRG19008] "Knowledge-infused Deep Learning for Natural Language Processing" and ERC Consolidator Grant DIALECT 101043235.

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*.

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.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, 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.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Elias Stengel-Eskin, Kyle Rawlins, and Benjamin Van Durme. 2024. Zero and few-shot semantic parsing with ambiguous inputs. In *The Twelfth International Conference on Learning Representations*.

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*.

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

The entity selection process was based on the authors' intuition, followed by filtering the initial list to include only entities that all LLMs recognize in both their main and company readings (that is, the entities have passed Study 1).

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 illustrating the popularity of the entities are shown in Figure 4.

Type	Entity	Ambiguity	Company Reading			Non-Company Reading		
			Popularity	`prop X`	`prop type X`	Popularity	`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
	Average	**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
	Average	**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
	Average	**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
	Average	**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
	Average	**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
	Average	**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. Data collected on: June 10, 2024. Dash means there is no Wikipedia disambiguation page for this entity. Additionally, we provide the model performance on each entity demonstrated in Study 3, aggregated across the models.

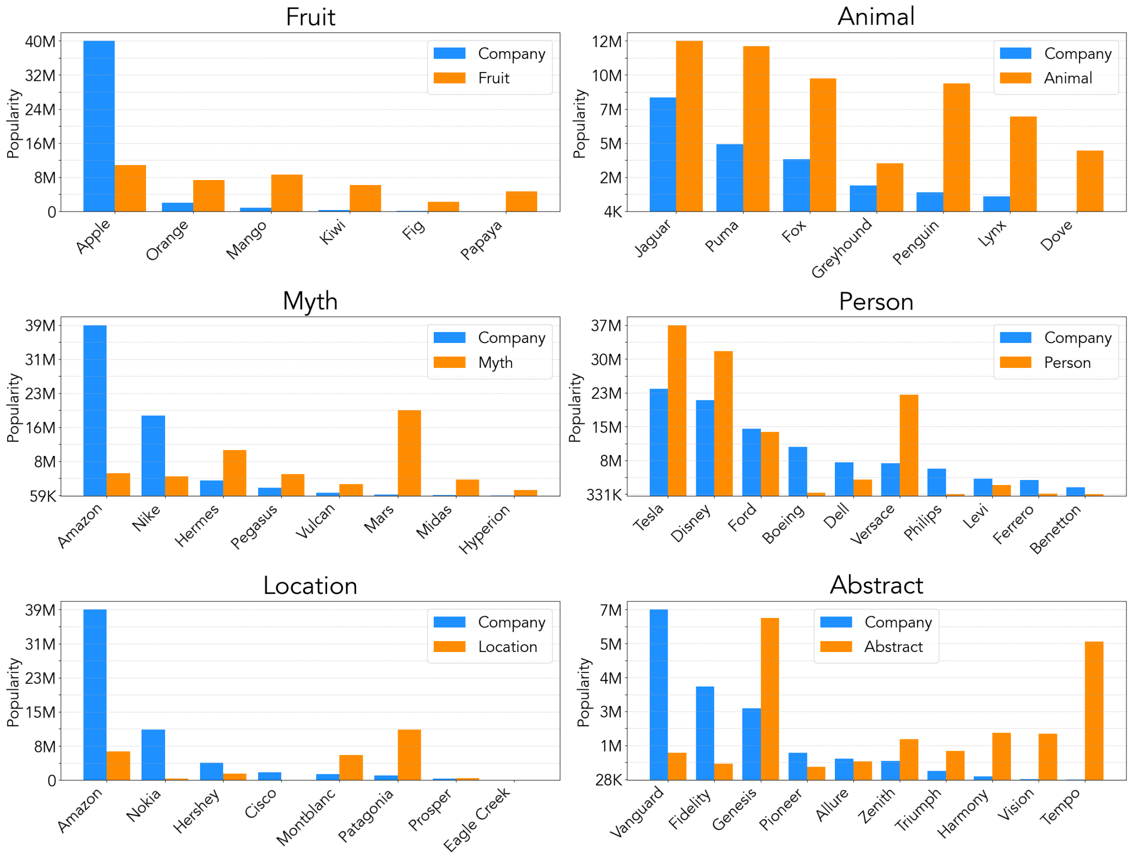

Figure 4: Popularity distribution of company and non-company readings of all 49 entities involved in our studies.

B Annotation Details

To evaluate the LLMs' output, we opted for a manual annotation of all models generations. To keep the annotation process as simple and clear as possible as well as preserve experiments scalability, we designed it to resemble automatic evaluation rather than typical human annotation, ensuring that it remains objective and unambiguous. For example, determining whether a model's response about *apple* refers to the company or the fruit is straightforward by examining the model's reply.

Prompt: *"Provide the color for Apple"*, Llama-3: *"The iconic color often associated with Apple is Silver/Space Gray, but if you're thinking of a more specific brand element, it's likely the bright, bold, and vibrant color: **Red**"* (obviously, this response refers to the entity *apple* as a company).

Prompt: *"Provide the color for a fruit Apple"*, Llama-3: *"The typical color of an Apple is RED, but apples can also come in other colors such as: * Green (Granny Smith, Green Apple)* Yellow (Golden Delicious, Yellow Apple)* Red with Yellow or Green stripes (Gala, Fuji)* Purple (some heirloom varieties) But, if I had to pick one, I'd say the most iconic color for an apple is RED!"* (this response addresses the color of *apple* as a fruit).

The same is valid for the annotation in Study 3:

Prompt: *"Provide the date of birth for Boeing"*, Llama-3: *"Boeing's founding date is July 15, 1916, when William E. Boeing and Conrad Westervelt founded [...], which later became The Boeing Company."* (the annotation is clear: the model adopted the "company" reading. Since we are not assessing factual correctness, the accuracy of dates is irrelevant, and we do not have gold labels for it).

C Prompts

The full lists of prompts used in our experiments are to be found in Table 6. At a high level, all our studies can be viewed as different variants of the same prompt, designed to elicit the same information (entity types), and extensive prompt engineering is beyond the scope of this work. However, we conducted some experiments on prompt variation, considering two types of modification: one at the formatting level and the other at the semantic level.

Formatting level. To assess the variability in model performance across different prompt for-

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."*).

mats, we utilized the tool FormatSpread (Sclar et al., 2023). This tool was initially designed to generate semantically equivalent variations of a given prompt to identify the most effective format. For our study, we randomly selected 20 prompt formats from those generated by FormatSpread, applied these formats to all our original prompts, and repeated the same experimental and evaluation procedures. Examples of these formats include: (original) *Tell me about a company called Amazon*, (1) *Tell me about a company called - Amazon*, (2) *TELL ME ABOUT A COMPANY CALLED::: Amazon*, (3) *TELL ME ABOUT A COMPANY CALLED\nAmazon*, (4) *Tell Me About A Company Called\tAmazon*, (5) *tell me about a company called Amazon*, (6) *TELL ME ABOUT A COMPANY CALLED : Amazon*, etc. However, we did not observe any significant differences in the results compared to those obtained with the original prompts.

Semantic level. At the semantic level, we tested different paraphrasings to select the prompts for the final experimental setup. A notable difference between paraphrasings was observed only in Study 2, where we used four variants of the initial prompt (see Table 6, Study 2) and aggregated the model responses through majority voting.

D Results: Further Discussion

Study 1. The results of 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 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 often fail 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.

Study 2. Complete examples of model responses for *Myths* entities are as follows:

Prompt: *"Group the following entities: Amazon, Mars, Vulcan, Hermes, Hyperion, Nike, Midas, Pegasus."*, GPT-3.5: *"1. Mythological entities: Mars, Vulcan, Hermes, Hyperion, Nike, Midas, Pegasus.* ***2. Companies/brands: Amazon****"*, GPT-4o: *"Certainly! These entities can be grouped based on their categories or mythological references: By category:* ***1. Brands and Companies: - Nike - Amazon*** *2. Greek Mythology: - Nike (Also a brand, but originally the Greek goddess of victory) - Pegasus - Hermes 3. Roman Mythology: - Mars - Vulcan 4. Mythological figures and references from various mythologies: - Hyperion (Greek mythology, a Titan) - Midas (Greek mythology, king with the golden touch) By mythology: 1. Greek Mythology: - Nike - Pegasus - Hermes - Hyperion - Midas 2. Roman Mythology: - Mars (Roman counterpart to the Greek god Ares) - Vulcan (Roman counterpart to the Greek god Hephaestus)* ***3. Modern enti-***

ties named after mythological references: - Nike *(brand named after the Greek goddess of victory)* ***- Amazon*** *(company name inspired by the Amazon warriors of Greek mythology)"*, Mixtral: *"Sure, I will group these entities into the following categories: 1. Planets: Mars, 2. Mythological figures: Hyperion, Midas, Nike, Hermes, Vulcan, Pegasus,* ***3. Companies: Amazon.*** *Explanation: 1. Mars is a planet in our solar system. 2. Hyperion, Midas, Nike, Hermes, Vulcan, and Pegasus are all figures from Greek and Roman mythology. Hyperion is a Titan and the father of the Sun, Moon, and Dawn. Midas was a king who was granted the power to turn everything he touched into gold. Nike is the goddess of victory, Hermes is the god of trade, thieves, travelers, and athletes, Vulcan is the god of fire including the fire of volcanoes, and Pegasus is a winged horse. 3. Amazon is a multinational technology company that focuses on e-commerce, cloud computing, digital streaming, and artificial intelligence. It is one of the world's largest online marketplaces, AI assistant providers, and cloud computing platforms."*

Study 3. As discussed in the paper, models more successfully recognize and provides information for entities in their preferred reading. For example, all models successfully provided the required information about *Animals* and *Fruits* in their non-company readings (e.g., speed and color, respectively)–see Table 5 values in the first *Sum* column. In contrast, several models failed to provide information about the date of birth and geographic locations for some of the *People* and *Locations* entities (the second *Sum* column in Table 5). For instance, only LLaMa-3, Mixtral, and Mistral models managed to provide reasonable answers for all entities related to *Locations*. Other models did not recognize either 2 (GPT-4o) or 3 (Gemma and GPT-3.5) locations from the list, even when given the most detailed prompt:

Prompt: *"Provide the area in square meters for a geographic location of Cisco"*, Gemma: *"I am unable to access real-time information or location-specific data, and am unable to provide information regarding the area of Cisco in square meters. For the most up-to-date and accurate information, please refer to reliable sources such as public maps or official Cisco publications."*, GPT-4o: *"Cisco is a large multinational company that has numerous office locations and various facilities around the world, so it's challenging to specify the exact area in square meters for a specific "Cisco" location without more information. If you are referring to a particular office or campus, please provide additional details so I can help you more accurately. Otherwise, for the company's worldwide presence, there isn't a single, cohesive geographic area that can be quantified."*, GPT-3.5: *"I'm sorry, but I cannot provide the exact area in square meters for the geographic location of Cisco as it's a private company and not a specific geographical region. However, if you have a specific location in mind related to Cisco, please provide more information so that I can assist you with finding the area of that specific location."*

Complete examples of model responses for *People* entities are as follows:

Prompt: *"Provide the date of birth for a person Boeing"*, Mixtral: *"I'm sorry for any confusion, but Boeing is not a person, it's a multinational corporation that designs, manufactures, and sells airplanes, rotorcraft, rockets, satellites, telecommunications equipment, and missiles worldwide. The company was founded on July 15, 1916, but I cannot provide a date of birth for a corporation."* (labeled as incorrect, regardless of whether the provided year is the correct year of the company's founding), Gemma: *"I am unable to access personally identifiable information or sensitive date of birth information of individuals, including Boeing"* (labeled as correct, as the model adopted the correct "person" reading, even though no specific information about this person's date of birth was provided).

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

Prompt: *"Provide the date of birth for Boeing."*, Llama-3: *"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."*

In the meantime, the fact that all the models that were included in the experiments indeed possess this knowledge was confirmed by an additional sanity check using completely unambiguous prompts, where the entity X is presented with both first and

Model	Preferred Reading				Alternative Reading			
	Consistent	Partially Consistent	Inconsistent	Sum	Consistent	Partially Consistent	Inconsistent	Sum
	Companies Reading							
Gemma	18 (38.3%)	1 (2.1%)	9 (19.1%)	28	13 (27.7%)	0 (00.0%)	6 (12.8%)	19
Mistral	21 (42.9%)	7 (14.3%)	0 (00.0%)	28	10 (20.4%)	8 (16.3%)	3 (6.1%)	21
MiXtral	24 (50.0%)	3 (6.2%)	1 (2.1%)	28	12 (25.0%)	5 (10.4%)	3 (6.2%)	20
GPT-3.5	16 (32.7%)	0 (0.0%)	2 (4.1%)	18	19 (38.8%)	2 (4.1%)	10 (20.4%)	31
GPT-4o	18 (38.3%)	0 (0.0%)	0 (0.0%)	18	24 (51.1%)	4 (8.5%)	1 (2.1%)	29
LLaMa-3	24 (49.0%)	2 (4.1%)	0 (0.0%)	26	22 (44.9%)	1 (2.0%)	0 (0.0%)	23
	Animals Reading							
Gemma	3 (42.9%)	4 (57.1%)	0 (0.0%)	7	-	-	-	-
Mistral	2 (28.6%)	3 (42.9%)	2 (28.6%)	7	-	-	-	-
Mixtral	6 (85.7%)	1 (14.3%)	0 (0.0%)	7	-	-	-	-
GPT-3.5	5 (71.4%)	2 (28.6%)	0 (0.0%)	7	-	-	-	-
GPT-4o	6 (85.7%)	0 (00.0%)	1 (14.3%)	7	-	-	-	-
LLaMa-3	4 (57.1%)	3 (42.9%)	0 (0.0%)	7	-	-	-	-
	Fruits Reading							
Gemma	5 (83.3%)	1 (16.7%)	0 (0.0%)	6	-	-	-	-
Mistral	0 (0.0%)	6 (100.0%)	0 (0.0%)	6	-	-	-	-
Mixtral	1 (16.7%)	5 (83.3%)	0 (0.0%)	6	-	-	-	-
GPT-3.5	4 (66.7%)	2 (33.3%)	0 (0.0%)	6	-	-	-	-
GPT-4o	5 (83.3%)	1 (16.7%)	0 (0.0%)	6	-	-	-	-
LLaMa-3	2 (33.3%)	4 (66.7%)	0 (0.0%)	6	-	-	-	-
	Myths Reading							
Gemma	8 (100.0%)	0 (0.0%)	0 (0.0%)	8	-	-	-	-
Mistral	8 (100.0%)	0 (0.0%)	0 (0.0%)	8	-	-	-	-
Mixtral	7 (87.5%)	1 (12.5%)	0 (0.0%)	8	-	-	-	-
GPT-3.5	8 (100.0%)	0 (0.0%)	0 (0.0%)	8	-	-	-	-
GPT-4o	8 (100.0%)	0 (0.0%)	0 (0.0%)	8	-	-	-	-
LLaMa-3	-	-	-	-	8 (100.0%)	0 (0.0%)	0 (0.0%)	8
	People Reading							
Gemma	-	-	-	-	5 (71.4%)	0 (0.0%)	2 (28.6%)	7
Mistral	-	-	-	-	0 (0.0%)	0 (0.0%)	10 (100.0%)	10
Mixtral	-	-	-	-	1 (10.0%)	0 (0.0%)	9 (90.0%)	10
GPT-3.5	-	-	-	-	0 (0.0%)	0 (0.0%)	6 (100.0%)	6
GPT-4o	-	-	-	-	4 (50.0%)	1 (12.5%)	3 (37.5%)	8
LLaMa-3	-	-	-	-	6 (60.0%)	2 (20.0%)	2 (20.0%)	10
	Locations Reading							
Gemma	-	-	-	-	5 (100.0%)	0 (0.0%)	0 (0.0%)	5
Mistral	-	-	-	-	0 (0.0%)	3 (37.5%)	5 (62.5%)	8
Mixtral	-	-	-	-	0 (0.0%)	0 (0.0%)	8 (100.0%)	8
GPT-3.5	-	-	-	-	0 (0.0%)	0 (0.0%)	5 (100.0%)	5
GPT-4o	-	-	-	-	3 (50.0%)	1 (16.7%)	2 (33.3%)	6
LLaMa-3	-	-	-	-	8 (100.0%)	0 (0.0%)	0 (0.0%)	8
	Abstract Reading							
Gemma	-	-	-	-	10 (100.0%)	0 (0.0%)	0 (0.0%)	10
Mistral	-	-	-	-	3 (30.0%)	5 (50.0%)	2 (20.0%)	10
Mixtral	-	-	-	-	8 (80.0%)	2 (20.0%)	0 (0.0%)	10
GPT-3.5	10 (100.0%)	0 (0.0%)	0 (0.0%)	10	-	-	-	-
GPT-4o	4 (40.0%)	6 (60.0%)	0 (0.0%)	10	-	-	-	-
LLaMa-3	10 (100.0%)	0 (0.0%)	0 (0.0%)	10	-	-	-	-

Table 5: Results from Study 4 (absolute numbers and the percentange of all replies; the graphical interpretation of these findings is illustrated in Figure 3). For non-company readings, we present the results for each group separately. Depending on whether a model prefers this reading or not, the results are provided in the corresponding column. Additionally, please note that the percentages are calculated based on the number of entities for which the corresponding model provided information in Study 3, not the total number of entities used in the experiments. For example, the Gemma model has provided information about the location for only 5 entities out of 8; therefore, the percentage of each category (consistent, partially consistent, and inconsistent) would be calculated based on 5, not 8 (in contrast to, for instance, the Mistral model which provided information for all 8 entities).

last names, indicating a preferred *people* interpretation in the most unambiguous way:

Prompt: *"What is the date of birth of William E. Boeing?"*, Llama-3 & GPT-3.5: *"William E. Boeing was born on October 1, 1881."*

Another notable trend is that for *people* entities in their non-company reading, all models perform better with underspecified prompts (*"Provide the date of birth for X"*) and often decline to answer more detailed prompts (*"Provide the date of birth for a person X"*), referring to the absence of "sensitive" information about specific individuals.

Prompt: *"Provide the date of birth for Versace"*, GPT-3.5: *"Versace was born on December 2, 1946."*, Gemma: *"The date of birth for Gianni Versace is October 2, 1955"*.

Prompt: *"Provide the date of birth for a person Versace"*, GPT-3.5 Response: *"I'm sorry, but I am unable to provide the date of birth for a person named Versace as it is not specified. If you have a specific individual in mind, please provide more information so I can assist you further."*, Gemma Response: *"I am unable to access personally identifiable information or sensitive date of birth information of individuals. Sharing such data would violate privacy laws and ethical guidelines"*.

Study 4. In Study 4, we prompted the model about the information it provided in Study 3. For example, if in Study 3 the Mixtral model responded to *"Provide the area in square meters for a geographic location of Nokia"* with *"...The city of Nokia has a total area of approximately 327.74 sq.km. (or 32,774,000 sq.m.)..."*, in Study 4, it will be prompted with *"Is there a geographic location Nokia with an approximate area of 32,774,000 sq.m.? Answer ONLY with Yes or No. If you cannot answer this question, answer No."*. If the model answers "Yes", the knowledge about this entity is reaffirmed. If not all knowledge is reaffirmed—e.g., in Study 3, the model provided varying colors for a fruit entity, and in Study 4, when asked to confirm each color separately, it responded "Yes" to some and "No" to others—the entity is classified as *Partially Consistent*. If the model replies "No" about all information it previously provided about the entity in the group, the entity is counted as *Inconsistent*. Preliminary experiments revealed that closed-source LLMs yielded inconsistent results across multiple runs of the same prompt; therefore, we conducted five trials per prompt and considered the knowledge confirmed if it was confirmed in at least one run. The results are provided in Table 5.

The highest ratio of denying the correctness of provided information was observed in the *People* group, where the models were asked to reconfirm the date of birth using the prompt: *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.* For instance, GPT-3.5, which was among the top three best-performing models in Study 3, did not reconfirm any of the dates (i.e., it replied with a single token "No" to all the requests). To clarify such unexpected model behavior, we additionally ran experiments with a modified prompt (*Is <retrieved-info> the date of birth of a person <entity>? Answer with Yes or No and provide an explanation.*) aimed at obtaining some explanation[7]. We obtained surprising results, such as:

Prompt: *"Is December 5 1901 the date of birth of a person Disney? Answer with Yes or No and provide explanation.", GPT-3.5: "No. December 5, 1901 is not the date of birth of Walt Disney. Walt Disney was actually born on December 5, 1901."*

Prompt: *"Is February 26 1829 the date of birth of a person Levi Strauss? Answer with Yes or No and provide explanation.", GPT-3.5: "No. Levi Strauss, the founder of the famous clothing company Levi Strauss & Co., was born on February 26, 1829."'*

Prompt: *"Is April 26 1925 the date of birth of a person Ferrero? Answer with Yes or No and provide explanation."*, GPT-4o: *"No, April 26, 1925, is not the date of birth of a person named Ferrero. Michele Ferrero, the Italian entrepreneur known for creating Nutella and other popular confectionery products, was born on April 26, 1925."*

Thus, in fact, the model does confirm the knowledge, but since its first one-word-response is "No," our evaluation system considers this answer as negative, which we believe is a correct evaluation. The same confusion occasionally occurs with the *Animal* entities:

[7]Importantly, these types of prompts were used only for analysis. For consistency, we used the same prompts provided in Table 6, Study 4, for the final results.

Prompt: *"Does an animal Dove have 24-35 miles per hour (38-56 kilometers per hour) speed? Answer ONLY with Yes or No. Answer based on your internal knowledge. Do not include any additional clarification. If you cannot answer this question, answer No."*, Mistral Response: *'No. The average flying speed of a dove is around 24-35 miles per hour (40-56 kilometers per hour) for larger species like the Mourning Dove, but smaller species like the Laughing Dove can fly at speeds closer to 20 miles per hour (32 kilometers per hour). However, there is no species of dove that consistently reaches speeds above 35 miles per hour (56 kilometers per hour)."*

Again, we hypothesize that the root cause of this inconsistent behavior may be the non-preferred non-company reading of the *People* and *Location* categories, as well as in Study 3, which makes all interactions with the models more complicated and erroneous.

Finally, there are cases where the models fail to demonstrate knowledge of the non-company reading for *People* and *Location* entities (despite this knowledge being shown in Study 1 and evident in responses to questions in Study 3 by providing relevant and reasonable replies that contained the required information):

Prompt: *"Is 1830 the date of birth of a person Philips? Answer with Yes or No and provide explanation."*, Llama-3: *"NO. Philips is a company, not a person. Philips is a Dutch multinational technology company founded in 1891 by Gerard Philips and his father Frederik Philips. It is not a person's date of birth."*

Prompt: *"Does an animal Jaguar have 35 to 50 miles per hour (56 to 80 kilometers per hour) speed? Answer ONLY with Yes or No. Answer based on your internal knowledge. Do not include any additional clarification. If you cannot answer this question, answer No."*, Mistral: *"Yes. A Jaguar can reach speeds between 50 to 80 kilometers per hour, which is approximately 31 to 49 miles per hour."*

E Technical Details

None of our experiments require extensive computation or a large number of responses from LLMs. Getting all responses from both GPT-3.5 and GPT-4o for all prompts coasted less than 30$. For the Gemma and Mistral models, we used single H100 GPU, and for Llama-3 and Mixtral two H100 GPUs. The responses for all studies were collected in less than 1 hour for each model. We query GPT-4o and GPT-3.5 through the OpenAI API [8]. The open-sources models are available in the Huggingface Transformers library:

- Gemma-7B: https://huggingface.co/google/gemma-1.1-7b-it
- Mistral-7B: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
- Mixtral-8x7B: https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
- Llama-3-70B: https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

Interactions with the models were conducted using the OpenAI Chat Completions API [9].

To increase the reproducibility of our results, we fixed the seed value to 42. However, according to OpenAI's official documentation, the seed parameter is in beta, and determinism is not guaranteed (see https://platform.openai.com/docs/api-reference/chat/create#chat-create-seed). Therefore, we cannot fully guarantee the reproducibility of our results. All models were checked for consistency by running the same prompt multiple times, with minimal variation observed across runs in all studies except for Study 4. Consequently, in Studies 1–3, we evaluate the results of a single run, while in Study 4, each prompt was run 5 times and the results were aggregated. The temperature value was set to 1.0 in all experiments, and no system prompts were specified.

[8] https://platform.openai.com/docs/api-295reference/introduction

[9] https://platform.openai.com/docs/libraries/python-library

Study 1
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.
Study 2
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 according to what they all have in common: <animal-entities>.
Group the following according to what they all have in common: <fruit-entities>.
Group the following according to what they all have in common: <location-entities>.
Group the following according to what they all have in common: <myth-entities>.
Group the following according to what they all have in common: <person-entities>.
Group the following according to what they all have in common: <abstract-entities>.
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>.
Study 3
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>.
Study 4
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> level 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 6: Prompts used in Studies 1-4.