World Models and Consistent Mistakes in LLMs

Christopher Wolfram [1] **Aaron Schein** [2]

Abstract

Do LLMs have a consistent world model that is reflected in their responses? We study whether the behavior of `gpt-4o` reflects an underlying world model by measuring the consistency of its mistakes across different prompts and prompting strategies. We find that `gpt-4o` makes consistent mistakes regardless of the exact prompt phrasing or prompt language. However, substantially different prompts that rely on the same underlying information often yield inconsistent results, suggesting that `gpt-4o`'s responses may not reflect a single universal world model.[1]

1. Introduction

Large language models (LLMs) can generate text that reflects information about the world. Do these responses reflect a unified underlying model of the world?

We operationalize the concept of a *world model* as an internally-consistent set of facts and logical propositions about the world. We then say that an LLM reflects the world model $\mathcal{W}$ if its responses reflect the facts and propositions contained in $\mathcal{W}$. For example, if $\mathcal{W}$ contained the (perhaps incorrect) fact that "The Eiffel Tower is 900 feet tall", then an LLM whose responses reflect $\mathcal{W}$ would similarly give responses reflecting this supposed height of the Eiffel Tower. When asked "Is the Eiffel Tower taller than 800 feet?" it would respond in the affirmative, and when asked "How tall is the Eiffel Tower" it would give the answer of 900 feet. It would never generate a response containing statements contradicting $\mathcal{W}$, such as "The Washington Monument is 555 feet tall, making it taller than the Eiffel Tower". The goal of our experiments is to provide evidence of whether particular LLMs either do or do not reflect some underlying consistent world model.

[1]Department of Computer Science, University of Chicago, Illinois, USA [2]Department of Statistics, University of Chicago, Illinois, USA. Correspondence to: Christopher Wolfram <chriswolfram@uchicago.edu>.

ICML 2025 Workshop on Assessing World Models.

[1]Code available at github.com/chriswolfram/LLMConsistency

Under this formalism, whether an LLM reflects a world model is a *behavioral* question and not a *mechanistic* one. In other words, it does not require that there be any identifiable object in the operation of the LLM that represents the world model (whatever that would look like), but merely that the LLM behaves in a way that reflects a consistent set of "beliefs" about the world.

We test whether a large frontier model (`gpt-4o` from OpenAI) gives consistent answers when asked factual questions in different ways, and find evidence for `gpt-4o` reflecting a relatively consistent world model across different prompts and prompt languages (Section 3). However, when we ask `gpt-4o` substantially different questions that rely on the same underlying information, it often gives inconsistent results (Section 4).

Related work Existing work has sought to assess the degree to which LLMs are consistent in their responses, and has arrived at mixed conclusions. Some researchers have found that LLMs show relatively high consistency on certain tasks (Raj et al., 2022), while others have reported low consistency (Elazar et al., 2021; Fierro & Søgaard, 2022), and yet others have found that consistency varied with different tasks (Sahu et al., 2022; Zheng et al., 2024). Most have focused on measuring consistency in factual or logical questions adapted from LLM evaluations, while we focus on questions with numerical answers.

2. Accuracy and consistency

Suppose that an LLM gave perfectly accurate responses to every factual question for which there is an unambiguous answer. Such a model's responses would be perfectly consistent as well. This means that accuracy implies a certain level of consistency, and that models that are very accurate will also be relatively consistent.

In order to disentangle accuracy and consistency, we focus especially on *errors* in LLM responses. If an LLM has a consistent but inaccurate world model, then we would expect it to consistently make the same errors across a range of contexts. If, on the other hand, the LLM does not reflect any consistent world model, then its errors are free to vary from response to response. For example, suppose an LLM $\mathcal{M}$ reflected a world model that contained the inaccurate

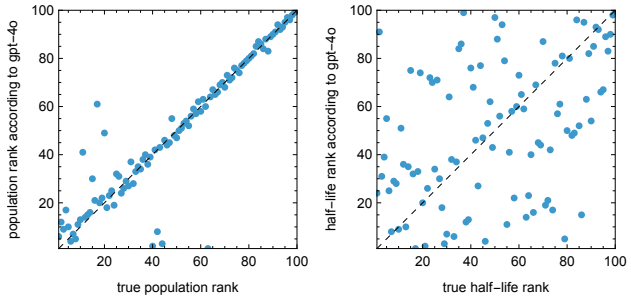
facts that "Chicago contains 1 million people" and "Dallas contains 3 million people". When asked, $\mathcal{M}$ should then say that Dallas is more populous than Chicago, that Chicago is not the third most populous city in the United States after New York and Los Angeles, etc.

3. LLM responses are consistent across prompts

When prompted for the same information with different phrasing, do LLMs give consistent answers, reflecting a consistent world model? We request information from `gpt-4o` using different prompts and assess whether it makes consistent errors.

We request data in two test domains: city populations and isotopic half-lives. In the case of city populations, we sample 100 random cities in the United States with a population greater than 3000. We then prompt `gpt-4o` to give the population of each of those cities and use structured generation to force it to give a numerical answer. For each city, we use a fresh context, so the model cannot look at its previous responses to help enforce consistency. We also set the temperature to 0, so the responses we study are the most likely ones under `gpt-4o`. We make analogous requests for the half-lives of 100 randomly chosen unstable isotopes.

We use these two test domains because they cover cases where `gpt-4o` is generally accurate (as with city populations) and where it is generally inaccurate (as with isotopic half-lives).



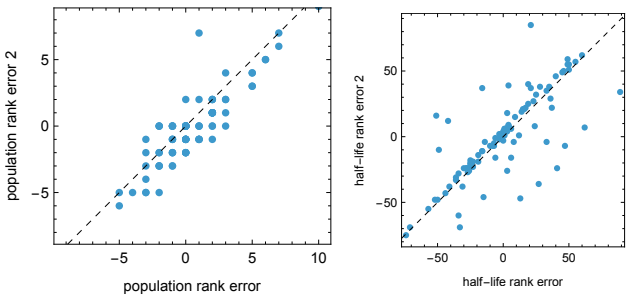
(a) (b)

Figure 1. City populations given by `gpt-4o` accurately reflect the true ranking of cities by population. However, `gpt-4o` is much less capable of accurately giving isotopic half-lives. 1a shows the true ranking of cities by population, compared with their ranking of those cities by their populations according to `gpt-4o`. Because the populations from `gpt-4o` are quite accurate, the points lie on the diagonal. 1b shows the same procedure, but for the half-lives of randomly chosen unstable isotopes. Unlike with city populations, `gpt-4o` is not able to accurately give the half-lives of unstable isotopes. The dashed line marks equality. The absolute values (instead of the rankings) are compared in Figure 6.

Assessing accuracy In order to first assess `gpt-4o`'s accuracy in these domains, we compare its responses to ground truth data on city populations and isotopic half-lives. In particular, we sort the cities by their true population, and then sort them according to the LLM's reported population numbers, and then compare the resulting rankings. We do the analogous experiment by sorting the isotopes by half-life as well.

We use rankings instead of absolute values because rankings are not as sensitive to rounding. Even when prompted to give exact answers, `gpt-4o` often gives round numbers. This could lead to the illusion of consistent errors merely because it is consistently rounding. However, rounding responses mostly preserves their ordering, so we primarily focus on comparing rankings and not absolute values. For many of our experiments, we produce analogous plots using absolute values instead of rankings in Appendix B.

The ranking of cities by population according to `gpt-4o` is quite close to the true ranking of cities (Figure 1a), while its ranking of isotopes by half-life is wildly different from the true ranking (Figure 1b).



(a) (b)

Figure 2. `gpt-4o` makes correlated errors when using different prompts to solicit the same information. 2a compares the difference in the true rank and the predicted rank of cities by population for two different prompts. The points lie on a grid because the errors are small and are always integer values. 2b shows the same thing for isotopic half-lives. Most of the points lie on or near the diagonal. The dashed line marks equality. The absolute errors are compared in Figure 7.

`gpt-4o`'s world model is consistent across small prompt changes If the LLM is given a slightly different prompt, does its response change? In other words, does `gpt-4o` reflect a consistent world model across small prompt changes?

We again request the populations and half-lives of our sampled cities and isotopes, but in each case we use two different rephrased prompts. For example, this is one variant of a prompt for getting isotopic half-lives:[2]

[2]All full prompts are available in Appendix A.

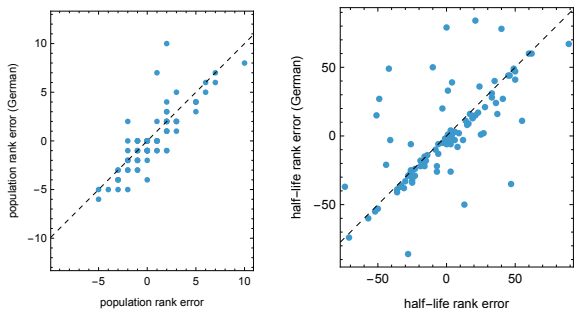
`What is the half-life of` X `in seconds?`

and another alternative version:

`Consider` X`. What is its half-life? Give your answer in seconds.`

These prompts are not identical, but they are substantially the same, and both request the same information. We find that `gpt-4o` gives similar answers with slightly altered prompts for both city populations and isotopic half-lives. Moreover, the errors made with one prompt are correlated to the errors made with another prompt (Figure 2). In other words, if the LLM overestimated the half-life of cadmium-124 with one prompt, then it is likely to overestimate it with the other prompt as well.

In the case of isotopic half-lives, where `gpt-4o` is generally not very accurate, this shows that there is a single preferred answer that the LLM gives regardless of prompt phrasing. This is nontrivial: One might have expected that the model would give a different (but similarly inaccurate) answer for every prompt. Instead, this shows that the model has a particular (and incorrect) value assigned to these isotopic half-lives that it consistently refers to.

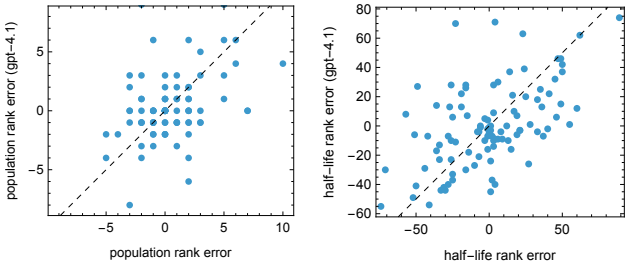


(a) (b)

Figure 3. `gpt-4o` makes correlated errors when asked for the same information in different languages. 3a compares the difference in the true rank and the predicted rank of cities by population when asked in English and in German. The points lie on a grid because the errors are small and are always integer values. 3b shows the same thing for isotopic half-lives. The dashed line marks equality. The absolute errors are compared in Figure 8.

`gpt-4o`'s world model is consistent across prompt languages Does `gpt-4o` reflect a consistent world model when prompted in different languages?

As above, we ask `gpt-4o` for the populations of various cities and the half-lives of unstable isotopes. We then ask it for the same information again, but this time with a translation of the original prompts into German. We find that `gpt-4o` makes similar errors in German to those it makes in English (Figure 3). This suggests that effectively the same world model is used regardless of the language that the LLM is prompted in and responds with. This is in line with other work that has found language-agnostic circuits in LLMs (Lindsey et al., 2025).



(a) (b)

Figure 4. `gpt-4o` makes different errors from `gpt-4.1`. 4a the error of population rankings made by `gpt-4o` and `gpt-4.1` when given the same prompt. There is much less correlation than seen when comparing different prompts with the same model. 4b shows the same thing for isotopic half-lives. The dashed line marks equality. The absolute errors are compared in Figure 9.

World models differ between LLMs All of the experiments shown so far have used `gpt-4o`. As a sanity check, we perform the same experiments with `gpt-4o-mini` and `gpt-4.1`—two other models from OpenAI—and check whether they too make the same mistakes.

We find that the errors made by `gpt-4o` and `gpt-4.1` are slightly correlated, but not nearly to the same degree as the errors that `gpt-4o` makes with different prompts (Figure 4). Given that `gpt-4o` and `gpt-4.1` were both trained by the same company (OpenAI) and likely with similar training data, this hints that most of `gpt-4o`'s consistency across prompts may be an emergent phenomenon whereby LLMs generally give consistent answers, and not merely a consequence of consistently inaccurate training.

Experiments with `gpt-4o-mini` showed similar results and are described in Appendix C.

4. LLM responses are not consistent across different prompting strategies

Instead of rephrasing the same question in different ways, what if we ask a different question that relies on the same underlying information? Will the LLM still respond in a way that is consistent?

We prompt `gpt-4o` to rank cities by population and check whether its rankings match what would be expected given the population numbers it provides when asked directly. For every pair of cities i and j, we ask `gpt-4o` which is more populous and inspect its next-token probabilities. The result is a matrix W where $W_{i,j}$ is equal to the probability that the model would have said that i is the more populous city.

Accuracy If the LLMs responses reflected a consistent world model, we might expect W to contain errors that align with the errors that the LLM made when asked for populations directly. Let v_i be the true population of city i and $\hat{v}_i$ be the population of that city given by `gpt-4o` when asked directly (as in Section 3). We find that the accuracy of W under the true populations is

$$\mathbb{P}_{i,j}(\mathbf{1}(W_{i,j} > 0.5) = \mathbf{1}(v_i > v_j)) = 0.834$$

The accuracy of W under the ranking of cities induced by the $\hat{v}_i$ is

$$\mathbb{P}_{i,j}(\mathbf{1}(W_{i,j} > 0.5) = \mathbf{1}(\hat{v}_i > \hat{v}_j)) = 0.835$$

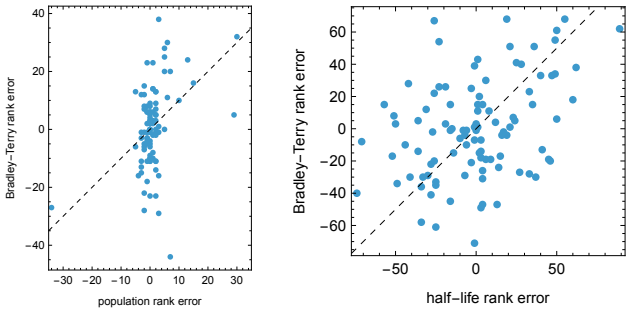
which is effectively the same, and so W reflects the true city populations just as well as it reflects the city populations given by `gpt-4o` when asked directly. This suggests that the world model used by `gpt-4o` when comparing two cities by population does not especially match that used when asked for population numbers directly.

Performing the same experiments with isotopic half-lives, we find that `gpt-4o`'s accuracy in predicting whether an isotope i has a longer half-live than an isotope j is 0.644. However, its accuracy relative to the half-lives given when asked directly is only 0.598, showing that `gpt-4o`'s responses matched the true half-lives slightly better than it matched those given by `gpt-4o` when asked directly. In other words, the LLM is making different mistakes now that we are asking it a substantially different question.

Bradley-Terry model We then use W to fit a Bradley-Terry model that assigns scores to each city based on its probability of being more populous than other cities (Newman, 2023). This allows us to convert the rank information in W into scores which can be compared with the other LLM responses. We do this for both the city populations and the isotopic half-lives.

The errors in the Bradley-Terry-derived scores show little relation to the errors made by `gpt-4o` when asked for numerical information directly (Figure 5). This provides further evidence that `gpt-4o`'s responses when asked for rank information do not reflect the same underlying world model as when it is asked for numerical information directly. However, we also find that the Bradley-Terry scores are consistent across prompts, suggesting `gpt-4o` could reflect different internally consistent world models in different contexts (see Appendix D).

Existing work has used Bradley-Terry models to estimate model beliefs (Wu et al., 2024). This experiment shows that this approach may yield different results from approaches that ask for numerical scores directly (O'Hagan & Schein, 2024).



(a) (b)

Figure 5. `gpt-4o` makes different errors when asked for the same underlying information in different ways. 5a the error of population rankings made when `gpt-4o` is asked for populations directly compared with those where populations are derived through the Bradley-Terry model. There is little correlation in the errors generated through these two approaches. 5b shows the same thing for isotopic half-lives. The dashed line marks equality.

Chain of thought rankings are no more consistent

Does chain of thought prompting increase response consistency?

We prompt the model to first give numerical values and then to compare them using chain of thought. For example, instead of directly asking whether a city i or a city j is more populous, we prompt `gpt-4o` to first give the population of i, then the population of j, and then say which is larger.

We find that chain of thought prompting increased overall accuracy for city populations, but decreased it for isotopic half-lives (see Table 1 for the full table of results). In both cases though, the chain of thought rankings were not significantly more consistent with the directly solicited values than the true values. In other words, the rankings generated with chain of thought prompting were sometimes more accurate, but were no more consistent with directly solicited values than would be expected from their (sometimes) increased accuracy.

Inspecting the chain of thought reasoning steps, we find that `gpt-4o`'s final ranking almost always matched the numerical values (i.e., populations or half-lives) it gave earlier in its reasoning. However, the numerical values it gave in its reasoning varied wildly from case to case, and often did not match the values it consistently gave when only asked for numerical values. This could be an example of "post hoc rationalization" as proposed by Arcuschin et al. (2025).

5. Conclusion

Our experiments show that `gpt-4o`'s responses reflect a consistent world model across different languages and

Table 1. Accuracy of city and isotope rankings relative to true values and relative to values given by models when asked directly.

	CITY POPULATIONS		ISOTOPE HALF-LIVES	
	WITHOUT COT	WITH COT	WITHOUT COT	WITH COT
RELATIVE TO TRUTH	0.834	0.920	0.644	0.485
RELATIVE TO LLM RESPONSES	0.835	0.905	0.598	0.487

prompt phrasings. However, when we ask for the same underlying information in a substantially different way (beyond phrasing or language), `gpt-4o` often gives inconsistent answers. Andreas (2022) proposed a model of LLM behavior as an "incoherent encyclopedia" where individual contexts reflect a consistent world model, but each context may reflect one of many different world models. It may be that superficial rephrasing does not trigger the LLM to deviate from one world model to another, but larger changes to the prompt structure do.

5.1. Future work

More knowledge domains Our experiments focused on two knowledge domains (city populations and isotopic half-lives) and two broad prompting strategies. Expanding to other kinds of knowledge domains and other prompting strategies would be useful in finding exactly what LLM responses reflect the same world models.

More LLMs All of our experiments have used models from OpenAI and have focused on `gpt-4o`. At this point, we do not know which results generalize to other models, and particularly to models that have not gone through extensive fine-tuning.

Effect of training data If `gpt-4o` were trained again from scratch, would it give the same incorrect half-life of cadmium-124? In order to build more confidence that consistent mistakes are an emergent property of LLMs, and not merely a property of consistently bad training data, multiple copies of a model could be trained on the same data but with different random seeds. If these models were each internally consistent, but different from one another, it would provide strong evidence that consistent mistakes are an emergent phenomenon of training.

Accounting for the incoherent encyclopedia Future work could directly test the incoherent encyclopedia model: A version of our analysis could be created where all information is derived from a single context, and consistency within that context could be assessed.

References

Andreas, J. Language models as agent models. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.423. URL https://aclanthology.org/2022.findings-emnlp.423/.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL https://arxiv.org/abs/2503.08679.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., and Goldberg, Y. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Fierro, C. and Søgaard, A. Factual consistency of multilingual pretrained language models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3046–3052, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.240. URL https://aclanthology.org/2022.findings-acl.240/.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Newman, M. E. J. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25, 2023. URL http://jmlr.org/papers/v24/22-1086.html.

O'Hagan, S. and Schein, A. Measurement in the age of llms: An application to ideological scaling, 2024. URL https://arxiv.org/abs/2312.09203.

Raj, H., Rosati, D., and Majumdar, S. Measuring reliability of large language models through semantic consistency. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=SgbpddeEV-C.

Sahu, P., Cogswell, M., Gong, Y., and Divakaran, A. Unpacking large language models with conceptual consistency, 2022. URL https://arxiv.org/abs/2209.15093.

Wu, P. Y., Nagler, J., Tucker, J. A., and Messing, S. Concept-guided chain-of-thought prompting for pairwise comparison scoring of texts with large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 7232–7241, 2024. doi: 10.1109/BigData62323.2024.10825235.

Zheng, D., Lapata, M., and Pan, J. Z. How reliable are llms as knowledge bases? re-thinking facutality and consistency, 2024. URL https://arxiv.org/abs/2407.13578.

A. Prompts

A.1. City populations

For getting city populations:

`What is the population of` X`? Answer in a JSON format even if uncertain.`

Alternative prompt for getting city populations:

`How many people live in` X`? Answer in a JSON format even if uncertain.`

German prompt for getting city populations:

`Wie hoch ist die Bevölkerung von` X`? Antworte im JSON-Format, auch wenn du dir unsicher bist.`

For ranking cities by population:

`Which has a larger population, A)` X `or B)` Y`? Answer in a JSON format with either A or B even if uncertain.`

For ranking cities by population with chain of thought:

`Which has a larger population, A)` X `or B)` Y`? Answer in a JSON format with the population of A, the population of B, and then which of A and B is larger. Answer even if uncertain.`

A.2. Isotopic half-lives

For getting the half-lives of isotopes:

`What is the half-life of` X `in seconds? Answer in a JSON format even if uncertain.`

Alternative prompt for getting half-lives of isotopes:

`Consider` X`. What is its half-life? Give your answer in seconds in a JSON format even if uncertain.`

German prompt for getting half-lives of isotopes:

`Wie lang ist die Halbwertszeit von` X `in Sekunden? Antworte im JSON-Format, auch wenn du dir unsicher bist.`

For ranking isotopes by half-life:

`Which has a longer half-life, A)` X `or B)` Y`? Answer in a JSON format with either A or B even if uncertain.`

Alternative prompt for ranking isotopes by half-life:

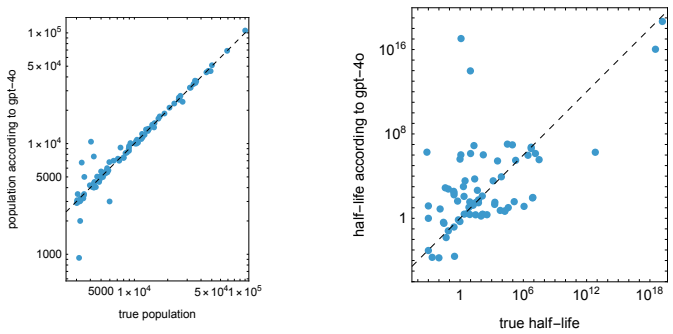
`Considering these two options, which has a half-life which is longer: A)` X `or B)` Y`? Answer in a JSON format with either A or B even if uncertain.`

For ranking isotopes by half-life with chain of thought:

`Which has a longer half-life, A)` X `or B)` Y`? Answer in a JSON format with the half-life of A in seconds, the half-life of B in seconds, and then with the option A or B which has a longer half-life. Answer even if uncertain.`

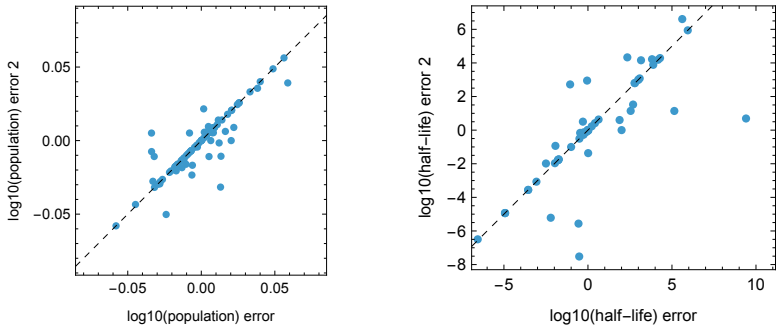
B. Using absolute values instead of rankings

In most of our experiments, we use rankings and errors in rankings to assess model outputs. Here we provide analogous plots using the absolute values instead of rankings.



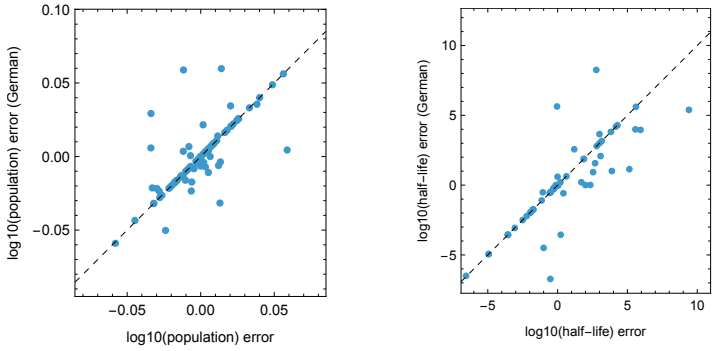
(a) (b)

Figure 6. City populations given by `gpt-4o` accurately reflect the true of cities by population. However, `gpt-4o` is much less capable of accurately giving isotopic half-lives. 6a shows the true populations of 100 randomly chosen cities with population over 3000, compared with the populations of those cities according to `gpt-4o`. Because the populations from `gpt-4o` are quite accurate, the points lie on the diagonal, indicating near equality. 6b shows the same procedure, but for the half-lives of 100 randomly chosen unstable isotopes. Unlike with city populations, `gpt-4o` is not able to accurately give the half-lives of unstable isotopes. The dashed line marks equality.



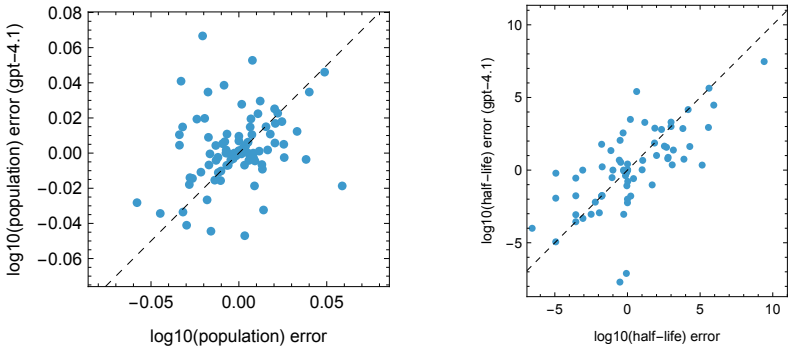
(a) (b)

Figure 7. `gpt-4o` makes correlated errors when using different prompts to solicit the same information. 7a compares the difference in the true and the predicted city populations according to two different prompts. Populations are log-scaled before they are compared. 7b shows the same thing for isotopic half-lives. Most of the points lie on or near the diagonal. The dashed line marks equality.



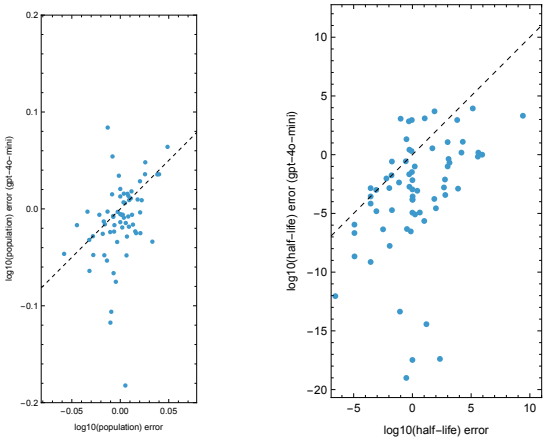
(a) (b)

Figure 8. `gpt-4o` makes correlated errors when asked for the same information in different languages. 8a compares the difference in the true and the predicted populations of cities when asked in English and in German. 8b shows the same thing for isotopic half-lives. The dashed line marks equality.



(a) (b)

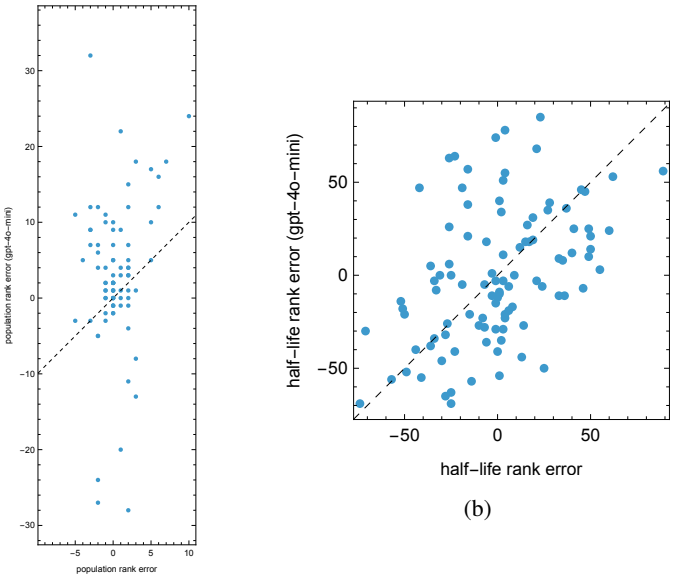
Figure 9. `gpt-4o` makes different errors from `gpt-4.1`. 9a the errors in populations given by `gpt-4o` and `gpt-4.1` when given the same prompt. There is much less correlation than seen when comparing different prompts with the same model. 9b shows the same thing for isotopic half-lives. The dashed line marks equality.



(a) (b)

Figure 10. `gpt-4o` makes different errors from `gpt-4o-mini`. 10a the errors in populations given by `gpt-4o` and `gpt-4o-mini` when given the same prompt. There is much less correlation than seen when comparing different prompts with the same model. 10b shows the same thing for isotopic half-lives. The dashed line marks equality.

C. Comparisons between `gpt-4o` and `gpt-4o-mini`



(a) (b)

Figure 11. `gpt-4o` makes different errors from `gpt-4o-mini`. 11a the error of population rankings made by `gpt-4o` and `gpt-4o-mini` when given the same prompt. There is much less correlation than seen when comparing different prompts with the same model. 11b shows the same thing for isotopic half-lives. The dashed line marks equality. The absolute errors are compared in Figure 10.

D. Bradley-Terry scores are consistent across prompts

We re-ran the Bradley-Terry analysis for isotopic half-lives with a rephrased prompt (see Appendix A). The errors in the resulting Bradley-Terry scores are highly correlated with the errors made with the original Bradley-Terry prompting.

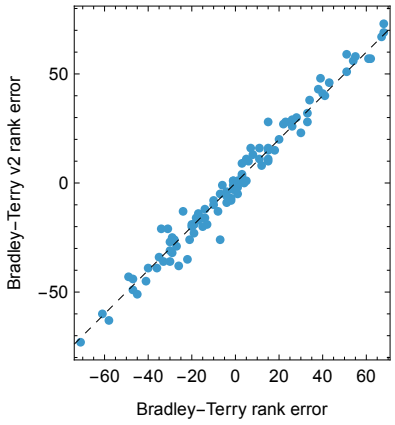


Figure 12. Errors in Bradley-Terry scores are consistent across differently phrased prompts. The dashed line marks equality.