Fact Recall, Heuristics or Pure Guesswork?
Precise Interpretations of Language Models for Fact Completion

Anonymous ACL submission

Abstract

Previous interpretations of language models (LMs) miss important distinctions in how these models process factual information. For example, given the query "Astrid Lindgren was born in" with the corresponding completion "Sweden", no difference is made between whether the prediction was based on having the exact knowledge of the birthplace of the Swedish author or assuming that a person with a Swedish-sounding name was born in Sweden. In this paper, we investigate four different prediction scenarios for which the LM can be expected to show distinct behaviors. These scenarios correspond to different levels of model reliability and types of information being processed—some being less desirable for factual predictions. To facilitate precise interpretations of LMs for fact completion, we propose a model-specific recipe called PRISM for constructing datasets with examples of each scenario based on a set of diagnostic criteria. We apply a popular interpretability method, causal tracing (CT), to the four prediction scenarios and find that while CT produces different results for each scenario, aggregations over a set of mixed examples may only represent the results from the scenario with the strongest measured signal. In summary, we contribute tools for a more granular study of fact completion in language models and analyses that provide a more nuanced understanding of how LMs process fact-related queries.

1 Introduction

Language models (LMs) trained on large textual corpora have been found to store significant amounts of factual information (Petroni et al., 2019). Given a factual query about the world, e.g. "Tokyo is the capital city of", most modern LMs can produce the correct fact completion "Japan" solely based on the information stored in their parameters. While there are many research results documenting the fact proficiency of LMs (Kandpal et al., 2023; Mallen et al., 2023), our understanding of how these models perform fact completion is still under rapid development. A growing area of research with focus on examining and explaining model behavior is mechanistic interpretability (Elhage et al., 2021; Geiger et al., 2021), which has already yielded insights related to where LMs store and process factual information (Meng et al., 2022; Geva et al., 2023; Haviv et al., 2023).

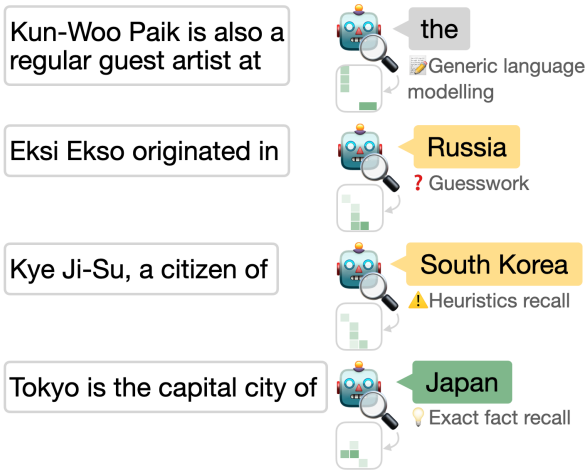


Figure 1: Prediction scenarios and corresponding prompt completion examples. As illustrated by the heat maps, we find that different prediction scenarios lead to different interpretability results.

While this body of work has broken new ground and provided us with interpretations of LMs in fact completion situations, it lacks precision. In particular, in Meng et al. (2022); Geva et al. (2023), the model is assumed to make correct predictions because it recalls facts. However, it is well known that LMs can make correct predictions based on many different signals in the prompt, not all corresponding to exact fact recall. For example, LMs may pick up on spurious correlations to "solve the dataset" rather than the task this dataset should exemplify (Zellers et al., 2019; Niven and Kao, 2019; McCoy et al., 2019), and fact completion situations

are not an exemption to this (Poerner et al., 2020; Cao et al., 2021; Ladhak et al., 2023).

We propose a method for creating a diagnostic dataset with distinct test cases to enable more precise interpretations of LMs in fact completion situations. Our experiments with the mechanistic interpretability method of causal tracing (CT) (Meng et al., 2022) show that models exhibit a more complex behavior that is not captured by previous test datasets. In this work, we disentangle and detail different *prediction scenarios* for which LMs can be expected to show distinct behaviors. We identify four such scenarios (see Figure 1): 1) Generic language modeling, when the model does not respond with facts, such as when generating a story. 2) Guesswork, when the model responds with a fact but is uncertain. 3) Heuristics recall, when the model uses shallow heuristics, e.g. that people with Korean-sounding names are more likely to live in Korea. And finally: 4) Exact fact recall, when the model has indeed memorized the correct answer and recalls it for the prediction. We show how interpretations of LMs based on each scenario yield fundamentally different results.

Our contributions can be summarized as follows:

- We propose diagnostic criteria to identify four different prediction scenarios for factual queries that should be interpreted in separation (see Figure 1).
- We show how a dataset previously used for LM interpretations mixes these scenarios and propose an alternative method, PRISM, for creating model-specific diagnostic datasets to support precise studies of LM behavior. We create and release PRISM datasets for GPT-2 XL, Llama 2 7B and Llama 2 13B, respectively.
- Using our diagnostic datasets and the method of causal tracing (CT), we show how LM interpretations for each of the different prediction scenarios yield fundamentally different results, while interpretations of aggregations over different scenarios are dominated by the results from the exact fact recall scenario. This highlights the necessity of disentangled and precise interpretations of LMs for fact completion.[1]

[1]All of our code and data will be open-sourced once the anonymity period is over.

2 Background

This section provides a brief background on research topics relevant to our paper.

2.1 Interpretability and fact completion

Recent work by Meng et al. (2022); Geva et al. (2023); Haviv et al. (2023) has focused on the inference process of LMs for fact completion for simple (subject, relation, object) fact tuples, such as subject *Tokyo*, relation *capital-of* and object *Japan*, illustrated in Figure 1. This work hypothesizes that LMs follow a distinct process when producing accurate fact completions. This hypothesis was originally posed by Meng et al. (2022) based on aggregations of causal tracing (CT) results, which revealed a decisive role of MLP modules at (last subject token, mid layer) positions for accurate fact completion predictions. This was reasoned to indicate that these modules *recall fact associations* for a subject. Later results by e.g. Geva et al. (2023) support the same conclusion.

It is worth noting that these knowledge localization efforts are somewhat orthogonal to work on model editing, as observed by Hase et al. (2023). While localization results from CT may not identify the optimal parameters for knowledge editing, this does not mean that CT is inaccurate for knowledge localization. It simply means that *"localization analysis might answer a different question than the question answered by model editing"*, as stated by Hase et al. (2023).

2.2 Causal tracing

The method of causal tracing (CT) is a mechanistic interpretability method that has been highly influential and provided many interpretations of LMs (Stolfo et al., 2023; Monea et al., 2023). It was introduced by Meng et al. (2022) and relies on the study of indirect causal effects. The method works by first recording intermediate model representations during normal generation (clean run). Then noise is added to the query subject embeddings to obtain corrupted intermediate model representations (noised run). By restoring corrupted representations at different (token, layer) positions it is possible to infer what parts of the network are important for assigning a high probability to the predicted token with respect to the subject (patched run). The measured signal of model component importance is referred to as *indirect effect* and defined

as

$$\mathrm{IE}_{h_i^{(l)}}(o) = P_{h_i^{(l)},\,\mathrm{patched}}(o) - P_{\mathrm{noised}}(o) \quad (1)$$

where $P_{h_i^{(l)},\,\mathrm{patched}}(o)$ is the probability for token o after patching state $h_i^{(l)}$ at layer l for the input token at position i and $P_{\mathrm{noised}}(o)$ is the probability of o for the noised run. To reason about the general process of generating a prediction, results for important states are averaged over several samples to get the average indirect effect (AIE).

The CounterFact dataset For their study, Meng et al. (2022) also developed the CounterFact dataset. Their conclusion is based on the 1,209 known samples from CounterFact for which GPT-2 XL is accurate. By now, it has been frequently used for the interpretation of LMs performing fact completion (Geva et al., 2023).

2.3 LMs and memorization

A large body of interpretability work for fact completion situations is concerned with whether a model has memorized[2] some fact or not, and where that fact is stored in the model parameters, also referred to as the *parametric memory*. Recent knowledge probing investigations (Kandpal et al., 2023) have demonstrated that models struggle more with facts rarely seen during training, suggesting a correlation between training data frequency and memorization.

Mallen et al. (2023); Kandpal et al. (2023) find that queries asking for fact tuples rarely found in the LM training data are less likely to be known by the model. Mallen et al. (2023) take this one step further and use fact popularity (measured as Wikipedia page views) as a proxy for training data frequency to estimate the probability of a model knowing a fact. Conversely, Liu et al. (2023); Basmov et al. (2024) use synthesized facts to simulate a training data frequency of 0 to study model behavior in the face of the unknown.

2.4 LMs and heuristics

Research into model performance on factual benchmarks has shed light on different factors affecting a model prediction. Previous work has found that accurate fact completions may stem from superficial cues and learned shallow heuristics, such as lexical overlap, person name bias[3] (Poerner et al., 2020; Ladhak et al., 2023) or prompt bias[4] (Cao et al., 2021).

While shallow heuristics are a natural by-product of the way LMs are trained and may provide a shortcut solution for some samples, they rely on disputable and overgeneralizing assumptions. Therefore, LMs relying on shallow heuristics in a fact prediction setting is generally undesirable (McCoy et al., 2019). For example, Ladhak et al. (2023) found that name bias leads to hallucinations and factually incorrect summaries by LMs.

2.5 LMs and random guesswork

There have been claims that reliable LMs need to be able to abstain from generating low-confidence outputs (Feng et al., 2024). Meanwhile, most open-sourced LMs studied in academia are as of yet not equipped with the ability to abstain and the responsibility of detecting low-confidence outputs lies on the researcher.

Research on model calibration (Jiang et al., 2021; Vasudevan et al., 2019) has shown that token probability does not align well with performance and as such cannot be used as a good approximation of confidence. Some research has suggested that internal model states may encode information related to model confidence (Burns et al., 2023). However, different extraction methods have varying success and are model as well as dataset dependent (Yoshikawa and Okazaki, 2023). In addition to this, most work on estimating model confidence is from the field of model calibration, and relies on accuracy as the single measure for calibration through confidence-accuracy reliability diagrams.

3 PRISM datasets for precise studies of prediction scenarios

We develop PRISM (Precise Identification of Scenarios for Model behavior) datasets aiming to separate the different prediction scenarios introduced in Figure 1. This is motivated by an inspection of the 1,209 samples from CounterFact which reveals 510 samples likely to rely on heuristics and 365 samples to have a low popularity scores and thus be unlikely to have been memorized (see Appendix F).

[2]Note that the type of memorization referred to here is of an abstracted representation of the fact rather than, as studied in some literature, of exact string memorization.

[3]E.g. *Kye Ji-Su* being predicted to be a citizen of *South Korea* due to the form of the name

[4]E.g. predicting *London* for "Adam Doe was born in" due to training data statistics showing strong correlations between *London* and "was born in", disregarding the subject *Adam Doe*.

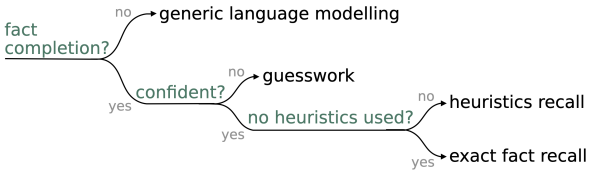


Figure 2: Our process for defining the four prediction scenarios (in black) using the three diagnostic criteria (in green).

These issues make the CounterFact dataset unsuitable for precise and comprehensive interpretations of LMs.

PRISM datasets are created by identifying subsets for each of the four prediction scenarios from Figure 1, where the LM can be expected to behave differently. These are identified by applying three diagnostic test criteria (Figure 2). Note that PRISM datasets are model-specific since they depend on model biases and parametric memories, which differ between LMs.

We develop PRISM datasets for GPT-2 XL (Radford et al., 2019), Llama 2 7B and Llama 2 13B. (Touvron et al., 2023), respectively. General statistics for the datasets can be found in Table 1 and examples corresponding to each prediction scenario for GPT-2 XL can be found in Table 2. The samples and our subsequent analysis is focused on the English language. Further details on the implementation of the datasets can be found in Appendix D.

3.1 Diagnostic criteria

To create the PRISM datasets we propose three necessary and comprehensive diagnostic criteria for which we define measurements (see Figure 2): (1) whether the prediction actually represents fact completion rather than generic language modeling; (2) whether the prediction is confident and robust to insignificant signals in the prompt; and (3) whether the prediction is based on the exact factual information expressed in the query or on heuristics triggered by surface-level cues.

These criteria provide a more fine-grained testing setup compared to using a single accuracy-focused criterion, as done in previous work.

Fact completion Our first criterion is *fact completion* – whether a prompt and the model's prediction correspond to the setting of a model completing a fact. By making sure the model is studied only in fact completion situations, interpretations can be assumed to elucidate some fact processing behavior. This as opposed to, for example, the LM generating a story about unicorns, for which a different model behavior is assumed to take place. This criterion is already implicitly used in previous research on interpretations of LMs for fact completion (Petroni et al., 2019; Meng et al., 2022; Geva et al., 2023), we simply make it explicit.

Following this body of work, for fact completion we limit ourselves to simple queries that express an incomplete fact (subject and relation), with the intent to let the LM generate the object as the next token. We use ParaRel query templates to collect samples of fact completion (Elazar et al., 2021). These are a set of rephrased expressions of a relation, where a different subject X can be substituted at the first position and an object Y is expected as the next token to be generated. For example, for the relation "born-in" we have templates such as "[X] was born in [Y]" or "[X] is originally from [Y]" (see Appendix C for a full list). 7 different relations are used, each with at least 5 different templates for query variations.

The authors of CounterFact (Meng et al., 2022) let the model generate freely until it produces an entity, but this may distort the original meaning of the template, e.g., by adding negation (Appendix F.3).

We define the measurement for fact completion as: If a *query* expresses a subject and relation and the *prediction* corresponds to an entity or concept that has a valid type (e.g. a place name when asking about location). That is, we exclude predictions such as "the", "a" and "with".

Confident prediction Our second criterion is *confident prediction* – whether the prediction can be considered confident, e.g. by being robust across insignificant perturbations to the query. Since most LMs by default cannot abstain from answering, we may end up in situations when a LM makes the correct prediction by chance while it randomly selects a token of the correct type (e.g. a city when asked for a birthplace) but has no relevant parametric knowledge for the specific fact. The importance of prediction confidence is highlighted in recent work by Xin et al. (2021) on model abstention and Elazar et al. (2021) on model consistency.

For the collection of PRISM samples, we opt for a definition of confidence grounded in desirable model behavior. We proxy model confidence by consistency in the face of semantically equivalent queries (Elazar et al., 2021; Portillo Wightman et al., 2023) and measure it as the agreement across

Scenario	GPT-2 XL #samples (#fact tuples)	Llama 2 7B #samples (#fact tuples)	Llama 2 13B #samples (#fact tuples)
Generic LM	1000 (-)	1000 (-)	1000 (-)
Guesswork	3282 (3181)	2917 (2846)	2822 (2220)
Heuristics	8352 (1868)	8414 (1960)	9224 (2062)
Exact fact	1322 (191)	5481 (580)	5995 (601)

Table 1: Statistics for the PRISM datasets for each LM considered in our study.

Scenario	Prompt	Prediction	Gold label	Conf	Pop	Bias
generic LM	Nara also enjoyed success in	the	singles	-	-	-
generic LM	Benjamin later joined a number of	other	clubs	-	-	-
guesswork	Sonar Kollektiv originated in	Russia	Berlin	1	215	-
guesswork	Joseph Clay was originally from	Ohio	Philadelphia	1	273	-
heuristics	Serok Nuvrome, a citizen of	Ukraine	-	6	0	name
heuristics	Balo Windhair has a citizenship of	Canada	-	5	0	prompt
exact fact	Thomas Ong is a citizen of	Singapore	Singapore	7	1418	none
exact fact	Shibuya-kei, that was created in	Japan	Japan	8	5933	none

Table 2: Samples from PRISM for GPT-2 XL designed to trigger different prediction scenarios. Conf(idence) measures how often the prediction was made, pop(ularity) measures page view rate and bias indicates detected bias when applicable.

paraphrases from the ParaRel dataset (Elazar et al., 2021).

We define the measurement for confident prediction as: If a *prediction* occurs among the top 3 predictions for at least 5 paraphrased queries. A *prediction* that only appears for one of the rephrased queries is deemed unconfident.

Usage of heuristics Our third and final criterion is *no dependence on heuristics* – indicating if the prediction is based on the exact factual information expressed in the prompt (subject and relation) rather than only on partial signals, i.e. heuristics. Predictions depending on heuristics indicate an over-reliance on unintended correlations in the training dataset based on surface forms of names or prompts, and are therefore unreliable (Cao et al., 2021; McCoy et al., 2019). Recalling information that is disputable and overgeneralizing (that is, capturing some statistical pattern that is only partially correct) is not equivalent to recalling the exact fact requested by a prompt. Therefore, it is important that we separate scenarios involving heuristics from those that do not.

For the collection of PRISM samples we use two indicators: presence of surface-level cues and memorization estimation. If a model indicates it has learned a heuristic related to the prediction, it is likely this is used for completing the query. Additionally, if we know that the LM does not know the fact requested by a prompt but it still makes a confident prediction, we can assume that the prediction corresponds to some form of heuristics recall.

We use three types of heuristics-detecting filters: lexical overlap (between subject and prediction) and person name bias by Poerner et al. (2020) and a prompt bias filter based on the findings by Cao et al. (2021). We use fact popularity to estimate model knowledge, proxied by Wikipedia page views for year 2019[5], based on work by Mallen et al. (2023).

Lexical overlap is measured as the string overlap between the subject and the prediction. For person name bias, we query with the following prompts: "[X] is a common name in the following city:" and "[X] is a common name in the following country:", where X is replaced with the subject name to check for bias. For prompt bias we use the original prompt templates as defined by ParaRel and replace the subject placeholder with generic substitutions, such as "[She] is a citizen of:". If any of the top 10 token predictions for these queries matches the model prediction for the full fact query, we mark that as corresponding to bias. More details on the implementation of the filters can be found

[5]Using the Pageview API.

in Appendix D.5.

We define the measurement for usage of heuristics as: If a *prediction* is made based on no memorization and the *query* contains surface-level cues.

3.2 Dataset creation based on diagnostic criteria

Using the above criteria, we build PRISM datasets of (*query*, *prediction*) samples representative of each of four potential prediction scenarios: 1) generic language modeling, 2) random guesswork, 3) heuristics recall and 4) exact fact recall. In this section, we introduce definitions of the prediction scenarios based on our diagnostic criteria, illustrated in Figure 2, and our method for producing the PRISM samples. Our goal is to create splits that contain a single prediction scenario rather than to identify all samples that correspond to that scenario. As such, we opt for stricter thresholds when needed to ensure high precision samples. Thus some examples of scenarios may be excluded as our approach is not intended to classify any instance into one of the four scenarios.

Generic language modeling We define all samples corresponding to (*fact completion: False*) as representative of a generic language modeling scenario. The generic language modeling scenario corresponds to all LM inference processes unrelated to fact completion and we consider it a baseline group of prediction scenarios assumed to take place for generic model predictions, illustrated in Figure 1 (Haviv et al., 2023).

Generic language modeling samples are retrieved from Wikipedia.[6] We follow an approach similar to that of Haviv et al. (2023) to ensure that we only collect samples corresponding to generic language modeling and not fact completion. The extraction is done by sampling sentences that start with the subject of the article in order to comply with the causal nature of the models and to allow for causal interventions on the subject. We discard the sentence if its natural continuation begins with a capital or number (indicating this could be an entity and thus potentially fact completion).

Random guesswork We define all samples corresponding to (*fact completion: True*, *confident prediction: False*) as representative of a random guesswork scenario. These predictions can be accurate or inaccurate and may differ from generic language modeling in the inference process involved, as there is some indication that the model recognizes a template relevant to the query and predicts the correct type of entity.

For the collection of random guesswork samples, we first populate ParaRel templates with subjects and objects from LAMA (Petroni et al., 2019). We retain (*query*, *prediction*) samples for which the prediction is a valid object from LAMA (e.g. a city when asked for birthplaces) but unconfident (i.e. it does not occur among the top 3 predictions for at least 5 paraphrased queries).

Heuristics recall We define all samples corresponding to (*fact completion: True*, *confident prediction: True*, *no usage of heuristics: False*) as representative of a heuristics recall scenario. Heuristics recall occurs for predictions based on learned overgeneralized heuristics triggered by surface level cues.

We collect samples representative of heuristics recall by populating ParaRel templates with synthetic fact tuples from a fantasy name generator (more details can be found in Appendix D.3). Since the subjects are synthetic, facts about them cannot have been memorized by the model (Liu et al., 2023; Basmov et al., 2024). Filtering confident predictions of valid objects for which a single type of bias is identified forms our heuristics recall samples.

Exact fact recall We define all samples corresponding to (*fact completion: True*, *confident prediction: True*, *no usage of heuristics: True*) as representative of an exact fact recall scenario. The exact fact recall scenario corresponds to situations for which the LM can be expected to have memorized the full fact tuple expressed by the query and fetches this from memory for the prediction.

To obtain samples representative of exact fact recall, we again use the LAMA fact tuples. We filter predictions that are 1) confident, 2) not labeled as corresponding to any bias, 3) corresponding to a fact memorized by the LM, and 4) correct.

4 Analysis of causal tracing

To illustrate the importance of precise testing data, we investigate the sensitivity of a popular mechanistic interpretability approach – causal tracing (CT) – to our taxonomy of prediction scenarios and their aggregations. Our analysis aims to address

[6] We use `20220301.en` from HuggingFace at https://huggingface.co/datasets/wikipedia

the question *Do CT results and the corresponding conclusions change with the underlying prediction scenario(s)?*

4.1 Experimental setup

CT measures the importance of different (token, layer) positions for a certain prediction. In line with Meng et al. (2022), we analyze the averaged indirect effects per (token, layer) position, binning the input tokens into the following categories: first, middle, and last subject token; first subsequent token; further tokens; last token. In order to identify important (token, layer) states, we use confidence intervals to assess whether the distinctions between peak and non-peak causal effects are significant. We define a peak as significant, if the lower 95% confidence bound is higher than the upper confidence bound for any other (token, layer) position. We use the same hyperparameters as Meng et al. (2022) for our CT analysis.

To adjust for differences in absolute values of the probability of the traced (predicted) token, we normalize the measured indirect effect. We take inspiration from work by Hase et al. (2023) and define the *normalized indirect effect* as

$$\mathrm{NIE}_{h_i^{(l)}}(o) = \frac{P_{h_i^{(l)},\,\mathrm{patched}}(o) - P_{\mathrm{noised}}(o)}{|P_{\mathrm{clean}}(o) - P_{\mathrm{noised}}(o)|} \quad (2)$$

where $P_{\mathrm{clean}}(o) - P_{\mathrm{noised}}(o)$ is the total effect (TE) defined as the difference between the clean and the noised runs. The NIE measures the percentage of recoverable probability that was restored by patching state $h_i^{(l)}$. For a detailed discussion of the effects of normalization, see Appendix G.

4.2 Results

Figure 3 shows averaged normalized indirect effects of model states in GPT-2 XL for 1000 samples corresponding to each prediction scenario of PRISM in isolation as well as a combined plot of the 3 fact completion cases (exact fact recall, heuristics recall, and guesswork). The corresponding results for Llama 2 7B and Llama 2 13B can be found in Appendix H.1. They generally support the same conclusions as reached for GPT-2 XL. The subsequent analysis is focused on the prediction scenarios in isolation and aggregations.

Generic language modeling Results for the generic language modeling samples in Figure 3a do not indicate any peak that is uniquely significant. Among later layers states, the last token is decisive, as expected (Meng et al., 2022), but in early to mid layers all signals are of comparable strength.

Guesswork For the guesswork samples, the (last token, late layer) state is found to be significant. We also note an increase of AIE for the (last subject token, mid layer) MLP states.[7] Apart from the last prompt token no other non-subject tokens show AIE peaks throughout any of the model layers. This trend persists for all scenarios we study.

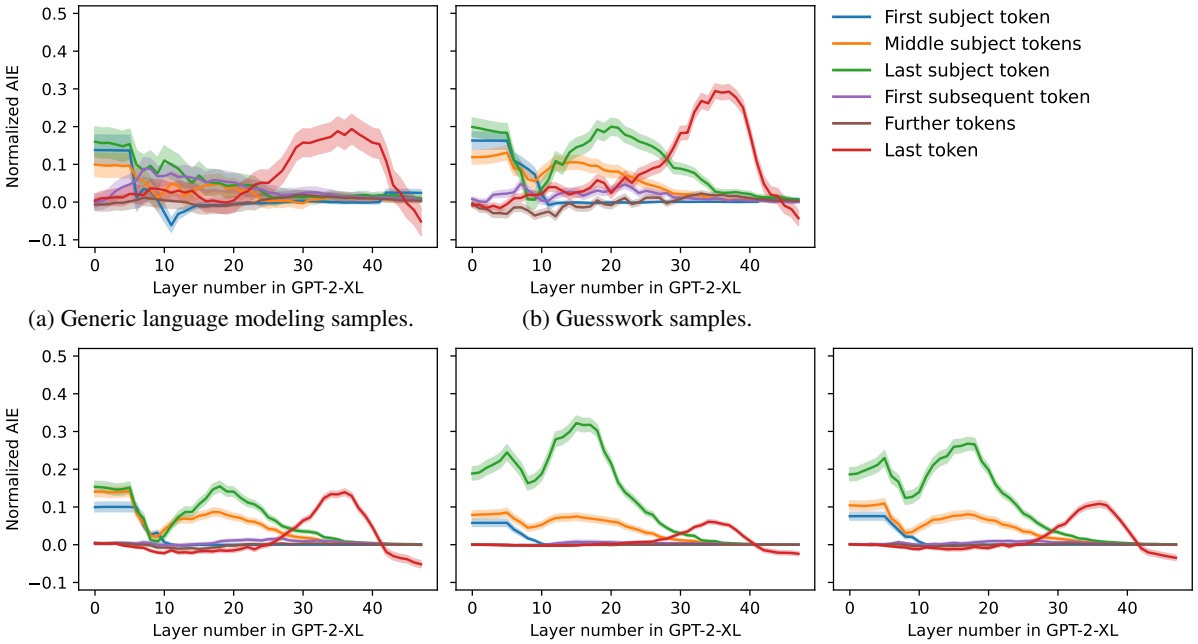
Heuristics recall The results for the heuristics recall samples in Figure 3c show a few peaks across model layers, all of similar strengths: last token state in late layers, subject tokens in mid layers, and subject tokens in early layers.

Exact fact recall Exact fact recall results Figure 3d show a clear peak in AIE in (last subject token, mid layer) MLP states and all other states (last token, other subject tokens) reduce in relative effect. This is fundamentally different from the other scenarios, where this state is either found to have a low AIE or to have comparatively the same importance as other states. Thus this is *the only prediction scenario that supports the same conclusion as previous work* that (last subject token, mid layer) MLP states are decisive (Meng et al., 2022).

Aggregations of prediction scenarios To test the effects of analyzing mixed samples, we produce results for a mixture of fact completion scenarios. The combined plot of exact fact recall, heuristics recall, and guesswork samples in Figure 3e generally reproduces the same CT results as observed in previous work, and thereby supports the same conclusion, i.e. (last subject token, mid layer) MLP states are decisive (Meng et al., 2022). This indicates that model interpretations over samples mixing prediction scenarios are misleading as they may be dominated by the characteristics of the exact fact recall scenario. Potentially, this could be due to the exact fact recall samples generally corresponding to higher prediction confidences.

This supports our hypothesis that previous interpretability results may have been recorded over mixtures of prediction scenarios, as we can reproduce their results with a mixture. We do not use this as proof that these results are based on mixtures (see Appendix F). We rather wish to show

[7] For Llama models this state is even found to have slightly higher AIE than the (last token, late layer) state.



(a) Generic language modeling samples. (b) Guesswork samples.

(c) Heuristics recall samples. (d) Exact fact recall samples. (e) Combined samples.

Figure 3: CT results for PRISM GPT-2 XL data. 1000 samples for each scenario in isolation. As well as 1000 combined samples (330 exact fact recall, 340 heuristics recall, 330 guesswork). Shaded regions indicate 95% confidence intervals.

how, without precise testing data, we can reach conclusions that are not supported by a big part of the sample (67%).

Even though heuristics recall can be seen as corresponding to some form of memory recall, similarly to exact fact recall, our CT results reveal very different patterns for the two scenarios. However, this difference is not reflected by the combined plot. This further supports our claim that samples corresponding to exact fact recall and heuristics recall should be analyzed in separation.

We also note that the CT results for the exact fact recall scenario have higher AIE at (last subject token, mid layer) MLP states and lower AIE at (last token, late layer) MLP states compared to the combined plot. Evidently, the LM relies heavily on information from MLP mid-layers for the exact fact recall prediction scenario. This provides additional support for the hypothesis proposed by Meng et al. (2022), as the exact fact recall samples further emphasize the pattern interpreted to indicate the memory storage role of mid MLP layers.

5 Conclusion

We identify four prediction scenarios that are fundamentally different and of differing reliability. These are *exact fact recall*, *heuristics recall*, *guesswork* and *generic language modeling*. We show that previous interpretability work for fact completion situations treat many of these as equivalent by using accuracy as the sole criterion for differentiating between different types of inference processes. Our analysis of a frequently used dataset, CounterFact, reveals samples that may trigger heuristics recall, as opposed to exact fact recall, and other problematic phenomena. To facilitate precise interpretations of LMs, we present a method for creating model-specific PRISM datasets with samples that represent each of our identified prediction scenarios. We create PRISM datasets for GPT-2 XL, Llama 2 7B and Llama 2 13B, and use them to test the prediction scenario sensitivity of an influential interpretability method, causal tracing (CT). We find that different prediction scenarios yield distinct CT results if studied in isolation. Consequently, CT results are not representative of the dataset as a whole if it contains examples of different prediction scenarios. Our results highlight the importance of studying different prediction scenarios in isolation and provide a method for doing this.

Limitations

Our results are limited to auto-regressive models and subject-first template queries. Using the methods described in this paper, PRISM datasets can be constructed for other types of LMs, such as encoder-based models, while we leave this for future work.

Moreover, the heuristics filters used for our dataset creation can only reveal the *possibility* of shallow heuristics being used by the LM. We also observe some questionable samples that go undetected by the filters, indicating that the filters are leaky. Furthermore, we find signs of name-based heuristics for non-person subjects for which we have no applicable filters. The detection of these cases would rely on more advanced detection methods and is left for future work. By complementing our dataset creation with knowledge estimations and sampling of synthetic fact tuples, we should avoid most filter failures, while we cannot completely rule out the possibility of there being some problematic samples with PRISM.

Even though we partition the PRISM samples based on whether the prediction is confident, we find that our results are sensitive to whether we investigate predictions with high or low probabilities from each partition. This indicates room for improvement for our method of detecting confident predictions, for which we already have noted a lack of comprehensive studies of model confidence metrics. Alternatively, this could be indicating a more fundamental issue with a qualitative difference in how models behave in low and high probability cases.

Lastly, we note that multiple interpretability methods would need to be applied to validate the exact underlying computation used by our LMs for the different scenarios in our taxonomy. When applying only CT, we cannot with certainty distinguish between effects of different prediction scenarios being used by the LM, as opposed to effects of data-sensitive quality issues of the CT method.

Ethics Statement

Interpretability methods for fact completion situations are not directly associated with any ethical concerns. Neither is the LAMA dataset or synthetic fact tuples used in this work.

References

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2024. Llms' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. *arXiv preprint arXiv:2404.06283*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1:1.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 17643–17668. Curran Associates, Inc.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. 2023. Prudent silence or foolish babble? examining large language models' responses to the unknown. *arXiv preprint arXiv:2311.09731*.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. 2023. A glitch in the matrix? locating and detecting language model grounding with fakepedia. *arXiv preprint arXiv:2312.02073*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vishal Thanvantri Vasudevan, Abhinav Sethy, and Alireza Roshan Ghias. 2019. Towards better confidence estimation for neural models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7335–7339.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

1040–1051, Online. Association for Computational Linguistics.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Computational resources

Experiments in this work are done on T4, A40 and A100 NVIDIA GPUs. Models used are GPT-2 XL (1.5B parameters), Llama 2 7B (7B parameters), and Llama 2 13B (13B parameters).

B Selection process of LAMA relations

The LAMA relations included in our PRISM dataset have been selected based on the following criteria:

1. We only include relations that have multiple templates for which 1) the object comes last in order to fit the autoregressive setting and 2) the subject comes first in order to simplify the causal reasoning of intervening on the subject;

2. We exclude relations with a lot of overlap between the subject and object and relations for which the answers are highly imbalanced toward only a few alternatives.

This corresponds to the relations P19 *place of birth*, P20 *place of death*, P27 *country of citizenship*, P101 *field of work*, P495 *country of origin*, P740 *location of formation* and P1376 *capital of*.

C ParaRel templates

We use the templates as described in Table 3 for the creation of PRISM queries.

Relation	Template
P19	[X] was born in [Y]
	[X] is originally from [Y]
	[X] was originally from [Y]
	[X] originated from [Y]
	[X] originates from [Y]
P20	[X] died in [Y]
	[X] died at [Y]
	[X] passed away in [Y]
	[X] passed away at [Y]
	[X] expired at [Y]
	[X] lost their life at [Y]
	[X]'s life ended in [Y]
	[X] succumbed at [Y]
P27	[X] is a citizen of [Y]
	[X], a citizen of [Y]
	[X], who is a citizen of [Y]
	[X] holds a citizenship of [Y]
	[X] has a citizenship of [Y]
	[X], who holds a citizenship of [Y]
	[X], who has a citizenship of [Y]
P101	[X] works in the field of [Y]
	[X] specializes in [Y]
	The expertise of [X] is [Y]
	The domain of activity of [X] is [Y]
	The domain of work of [X] is [Y]
	[X]'s area of work is [Y]
	[X]'s domain of work is [Y]
	[X]'s domain of activity is [Y]
	[X]'s expertise is [Y]
	[X] works in the area of [Y]
P495	[X] was created in [Y]
	[X], that was created in [Y]
	[X], created in [Y]
	[X], that originated in [Y]
	[X] originated in [Y]
	[X] formed in [Y]
	[X] was formed in [Y]
	[X], that was formed in [Y]
	[X] was formulated in [Y]
	[X], formulated in [Y]
	[X], that was formulated in [Y]
	[X] was from [Y]
	[X], from [Y]
	[X], that was developed in [Y]
	[X] was developed in [Y]
	[X], developed in [Y]
P740	[X] was founded in [Y]
	[X], founded in [Y]
	[X] that was founded in [Y]
	[X], that was started in [Y]
	[X] started in [Y]
	[X] was started in [Y]
	[X], that was created in [Y]
	[X], created in [Y]
	[X] was created in [Y]
	[X], that originated in [Y]
	[X] originated in [Y]
	[X] formed in [Y]
	[X] was formed in [Y]
	[X], that was formed in [Y]
P1376	[X] is the capital of [Y]
	[X] is the capital city of [Y]
	[X], the capital of [Y]
	[X], the capital city of [Y]
	[X], that is the capital of [Y]
	[X], that is the capital city of [Y]

Table 3: ParaRel templates used for all LAMA relations in our dataset creation.

D Creation process for PRISM

D.1 Generic language modeling samples

Data is sampled from Wikipedia extraction `20220301.en` from HuggingFace at https://huggingface.co/datasets/wikipedia. This extraction contains around 6.5M pre-cleaned English Wikipedia articles. We perform the following steps:

1. Select an entry (Wikipedia page) from the data. *E.g.: the page for "John Doyle (Irish artist)"*

2. Select a single sentence from the page that begins with any part of the page title (i.e. it could be the surname, if the subject is a person). *E.g.: "Doyle continued to exhibit miniatures until 1835, but by then he was experiencing greater success with his political cartoons, printed using the new reproductive medium of lithography, beginning in 1827."*

3. We discard the sentence if it is: 1) shorter than 5 words, 2) with more than 3 capitalized words (likely to be section headings). *E.g.:"Early life and family"*

4. We cap the sentence at 10 words.*E.g.: "Doyle continued to exhibit miniatures until 1835, but by [then]"*

5. We discard the sentence if its natural continuation begins with a capital or number (indicating this could be an entity and thus potentially fact completion). *E.g.: "Doyle won a gold medal in [1805]."*

We repeat this until we have 1000 datapoints (for 1000 unique entries in the data). For CT experiments, we trace the next token, freely predicted by the model.

D.2 Guesswork samples

To get examples of guesswork, we follow a process as described below and apply it to all fact tuples from LAMA[8] corresponding to the relations P19 *place of birth*, P20 *place of death*, P27 *country of citizenship*, P101 *field of work*, P495 *country of origin*, P740 *location of formation* and P1376 *capital of* and the corresponding ParaRel templates for these relations:

for each relation r (e.g. "P740" *location of formation*):

for each LAMA subject s for relation r (e.g. "Sonar Kollektiv"):

> **if** popularity score for s > 1000 **then** discard all examples for the subject (likely to have been memorized)
>
> **else** create empty list L for the tuple results

for each ParaRel template t for relation r (e.g. "[X] originated in [Y]"):

> predictions = top 3 predictions for (s, t) (e.g. "Sonar Kollektiv originated in")

for each p in predictions:

> **if** p is trivial (e.g. "the") **then** discard tuple (s, t, p)
>
> **else** add (s, t, p) to L

end for each p

end for each t

if count$(s, *, p)$ in $L = 1$ **then** add (s, t, p) to guesswork sample (that is, if a particular answer p occurs within the top 3 predictions for only one template it is considered guesswork)

end for each s

end for each r

We measure popularity score proxied by Wikipedia page views for year 2019[9] following (Mallen et al., 2023). We label a prediction as "likely to be memorized" if it corresponds to an average page view rate above 1000, as queries with lower popularity scores are unlikely to have been memorized (Mallen et al., 2023).

We determine if a prediction is trivial by only accepting a prediction as non-trivial if it's the correct answer for at least one data point in LAMA (for the same relation).

D.3 Heuristics recall samples

We create heuristics recall samples following a similar algorithm as for exact fact recall. Rather than starting from LAMA subject-object data points we produce a set of synthetic (non-existent) entities to

[8] https://github.com/facebookresearch/LAMA

[9] Using the Pageview API.

populate the ParaRel templates to simulate a popularity score of 0 (not present in training data) and thus ensure no memorization.

We create the synthetic data by following the steps for each relation in (P19, P20, P27, P101, P495, P740 and P1376):

1. Identify subject type distributions for the selected relations. *E.g. For relations P19, P20, P27 and P101, based on the "subject type constraint" from Wikidata the only allowed subject type is person.*
2. Generate subjects of the required types using `https://www.fantasynamegenerators.com`. *E.g. For person subjects, we generate a mixture of DND human, Russian, French, German, Korean, and Japanese names*
3. Perform de-duplication and check against Wikidata that no subject corresponds to a real entity.[10] The Wikidata check is done on a label level, since the generated names are pure strings. This limits our ability to check for a subject's existence, as we can only find exact matches.

For the collected synthetic samples we apply the following algorithm:

for each relation r (e.g. "P19" *place of birth*):

for each synthetic subject s for relation r (e.g. "Serok Nuvrome"):

create empty list L for the tuple results

for each ParaRel template t for relation r (e.g. "[X] was born in [Y]"):

predictions = top 3 predictions for (s, t) (e.g. "Serok Nuvrome was born in")

for each p in predictions:

if p is trivial (e.g. "the") **then** discard tuple (s, t, p)

if p is based on a single type of heuristics **then** add (s, t, p) to L

else discard tuple (s, t, p)

end for each p

end for each t

if count$(s, *, p)$ in $L \geq 5$ **then** add all $(s, *, p)$ tuples to heuristic recall sample (that is, if a particular answer p occurs within the top 3 predictions for at least 5 templates it is considered confident)

end for each s

end for each r

We identify if a prediction is trivial by the following set of rules: If the object type is a named entity (e.g. place names), we allow any generation beginning with a capital letter. This covers all relations apart from P101. For P101, we query Wikidata – we check if there exist any (s,r,o) Wikidata entry where the object label matches the generated token.

We measure if a prediction is based on heuristics by applying 3 filters explained in more detail in Appendix D.5.

Analysis of the heuristics recall samples in PRISM

Our heuristics recall analysis identifies samples that are confident, but for which no bias is detected. This can be counter-intuitive, as we do not expect the model to be able to make confident prediction when it has no bias to guide it. These examples are excluded from the PRISM samples, but we perform a deeper analysis of the 1,771 cases from GPT-2 XL.

6 instances identify the location of formation (P740) of "Oasis of Prejudice" as "London" (not identified as prompt bias, since the prompt bias check produces mostly years, indicating time to be the more natural interpretation of the queries). 9 instances from P101 (field of work) show the model potentially ignoring part of the query, by connecting "Nina Schopenhauer" with "philosophy" (potentially conflated with the philosopher Arthur Schopenhauer) and "Roch Chagnon" with "anthropology" (potentially conflated with the anthropologist Napoleon Chagnon). 23 examples of relation P495 show association of 5 fictional entities with Japan (3 of these contain the word "Berserk" – a possible conflating pattern with the manga of the same name). Further 790 examples come from relations P19 (born in) and P27 (citizen of). Some of these could be examples of a stronger association overwriting the expressed tuple (e.g. "Adolphe Trudeau" associated with "Quebec"), others may point to weaknesses of our name bias detection method. Finally, the most represented relation is

[10] The code for automatically querying Wikidata for real entities is provided as part of our code.

P1376 (capital of) with 938 examples. This relation does not lend itself to our subject name bias filter, however, we suspect a linguistic correlation between city names and countries may exist and those surface level signals can potentially explain some of the predictions.

This analysis confirms our concerns related to the coverage of the implemented heuristics recall filters. Evidently, there are some heuristics that go undetected by our filters. This is why we supplement the bias identification filters with memorization: For heuristic recall we simulate no memorization by using synthetic data and for exact fact recall we filter on high subject popularity (found to correlate well with memorization (Mallen et al., 2023)).

D.4 Exact fact recall samples

To get queries for which the LM performs exact fact recall, we follow a process as described below and apply it to all fact tuples from LAMA[11] corresponding to the relations P19 *place of birth*, P20 *place of death*, P27 *country of citizenship*, P101 *field of work*, P495 *country of origin*, P740 *location of formation* and P1376 *capital of* and the corresponding ParaRel templates for these relations:

for each relation r (e.g. "P19" *place of birth*):

for each LAMA subject s for relation r (e.g. "Thomas Ong"):

if popularity score for s < 1000 **then** discard all examples for the subject

else create empty list L for the tuple results

for each ParaRel template t for relation r (e.g. "[X] was born in [Y]"):

predictions = top 3 predictions for (s, t) (e.g. "Thomas Ong was born in")

for each p in predictions:

if p is based on heuristics **then** discard tuple (s, t, p)

if p is incorrect **then** discard tuple (s, t, p)

else add (s, t, p) to L

end for each p

end for each t

if count$(s, *, p)$ in $L \geq 5$ **then** add all $(s, *, p)$ tuples to exact fact recall sample (that is, if a particular answer p occurs within the top 3 predictions for at least 5 templates it is considered confident)

end for each s

end for each r

We measure popularity score proxied by Wikipedia page views for year 2019[12] following (Mallen et al., 2023).

We measure if a prediction is based on heuristics by applying 3 filters explained in more detail in Appendix D.5.

We categorize the predictions into "correct" or "incorrect" using the LAMA gold labels. For Llama 2 models we categorize a prediction as "correct" if it has more than 3 characters and fully matches the start of the gold label. This is necessary since the tokenizers for these model are more prone to split the gold labels into small tokens.

D.5 Detection filters for heuristics

Our detection of heuristics is based on 3 filters.

Lexical overlap Dependence on this heuristic is considered plausible if there is a string match between the prediction and the subject. *E.g. "San Salcos, the capital of [Sal]"*

Name bias We based this on model predictions for prompts expressing only a part of the requested fact. We query with the following prompts: "[X] is a common name in the following city:" and "[X] is a common name in the following country:". Where X is replaced with the subject name to check for bias. If any of the top 10 token predictions for these queries matches the model prediction for the full fact query, we mark that (*query*, *prediction*) pair as corresponding to person name bias. We can detect person name bias for relations P19, P20, P27, used in PRISM and additionally for P103 and P1412, present in CounterFact.

Prompt bias We use the original prompt templates as defined by ParaRel and replace the subject placeholder with generic substitutions. We use the substitutions described in Table 4 for each relation. We also remedy basic capitalization and

[11] https://github.com/facebookresearch/LAMA

[12] Using the Pageview API.

grammar errors that might surface from this automated prompt creation. An example of a prompt for detecting prompt bias for "Tokyo is the capital city of [Y]" is "It is the capital city of [Y]". If the top prediction for the former query is found among the top 10 token predictions for the latter query, the former query and corresponding prediction is marked as based on prompt bias.

Relation	Subject substitutions
P19	[He, She]
P20	[He, She]
P27	[He, She]
P101	[He, She]
P495	[It]
P740	[It, The organisation]
P1376	[It, The city]

Table 4: Subject substitutions used for constructing prompts to detect prompt bias.

E Examples from PRISM

Here, we include a few examples to illustrate the content of PRISM for different prediction scenarios. See Tables 5 to 8.

Model	Query	Prediction	Subject popularity	Gold label
GPT-2 XL	Thomas Ong is a citizen of	Singapore	1418	Singapore
	Shibuya-kei, that was created in	Japan	5933	Japan
	Palermo is the capital of	Sicily	34273	Sicily
Llama 2 7B	Disco Biscuits was created in	Philadelphia	3719	Philadelphia
	Don Broco, that was started in	Bed	6984	Bedford
	Nikephoros III Botaneiates passed away in	Constantin	1859	Constantinople

Table 5: (*query*, *prediction*) exact fact recall samples from PRISM for GPT-2 XL and Llama 2 7B.

Model	Query	Prediction	Rank	Gold label
GPT-2 XL	Sonar Kollektiv originated in	Russia	1	Berlin
	Haydn Bendall is originally from	England	2	Essex
	Joseph Clay was originally from	Ohio	2	Philadelphia
Llama 2 7B	Jean Trembley originated from	France	2	Geneva
	Dansez pentru tine, that originated in	France	2	Romania
	Milton Wright is originally from	Chicago	2	Georgia

Table 6: (*query*, *prediction*) random guesswork samples from PRISM for GPT-2 XL and Llama 2 7B.

Model	Query	Prediction	Bias
GPT-2 XL	Hirashima Hideyoshi, who has a citizenship of	Japan	name
	Balo Windhair has a citizenship of	Canada	prompt
	Olre Hellspirit was originally from	Hell	lexical
Llama 2 7B	Ha Songmin, who has a citizenship of	South (Korea)	name
	Wanda Hagel holds a citizenship of	Canada	prompt
	Limanaga, the capital city of	Lim	lexical

Table 7: (*query*, *prediction*) heuristics recall samples from PRISM for GPT-2 XL and Llama 2 7B.

Model	Query	Prediction	Gold label
GPT-2 XL	Dexmedetomidine is notable for its ability to provide sedation	and	without
	Solomon also defended the network's choice of games to	air	broadcast
	Walker added an immense amount of material to the	book	collections
Llama 2 7B	Dexmedetomidine is notable for its ability to provide sedation	and	without
	Solomon also defended the network's choice of games to	air	broadcast
	Walker added an immense amount of material to the	original	collections

Table 8: (*query*, *prediction*) generic language samples from PRISM for GPT-2 XL and Llama 2 7B.

F Inspection of CounterFact

In this section we assess the applicability of using the 1,209 known CounterFact samples for interpreting LMs in fact completion situations. First, we investigate what prediction scenarios are found for GPT-2 XL for the collection of the (*query*, *prediction*) samples (Appendix F.1). We find samples in the dataset likely to correspond to heuristics recall (510 samples) as opposed to exact fact recall (478 samples). Second, we inspect the total effects measured with the causal tracing approach for the dataset to find quality issues (Appendix F.2). Last, we observe a set of problematic samples with negated queries in CounterFact (Appendix F.3). Taken together, our results show that the dataset struggles to support precise and accurate interpretations of LMs. Our proposed PRISM dataset does not suffer from the aforementioned limitations.

F.1 Prediction scenarios

We inspect the CounterFact dataset for three of four prediction scenarios. The baseline prediction scenario corresponding to generic language modeling is skipped for the analysis as the dataset should not trigger this scenario by virtue of its creation process.

Random guesswork We cannot detect samples in CounterFact corresponding to random guesswork as our implementation of the confidence criterion is incompatible with the dataset. The dataset only provides one prompt per fact, while we require multiple prompt variations per fact to estimate confidence. This does not mean that there are no samples in the CounterFact dataset corresponding to random guesswork, it only means that we are unable to detect them. As a result, some of the samples below identified to correspond to heuristics recall or exact fact recall may actually correspond to random guesswork, as we are unable to separate these samples beforehand.

Heuristics recall We check for predictions based on shallow heuristics for the known CounterFact samples produced using GPT-2 XL. We find a total of 510 samples that may correspond to heuristics recall, of which 335 samples correspond to prompt bias, 155 to name bias and 20 to both name and prompt bias.[13] No lexical overlap between sample subject and object is found. Some examples marked for bias can be found in Table 9.

Using fact popularity, we also evaluate the known CounterFact samples through the lens of LM knowledge estimation. Table 10 lists the popularity scores distribution for the dataset. We find approximately 365 known CounterFact samples with popularity scores below 1000. These are unlikely to have been memorized by the model and are therefore unlikely to correspond to exact fact recall. Moreover, we find that around 50% of these samples (172 samples) have been detected by our heuristics filters, indicating that the remaining samples may also contain surface level signals not detected by our filters. This supports our claim that popularity metadata can serve as a complement for separating exact fact recall samples from heuristics recall samples.

Exact fact recall A total of 816 samples in CounterFact are found to have popularity scores above 1000, and are thus more likely to have been memorized by the model. We detect the potential usage of heuristics for 338 of these samples, meaning that approximately 478 samples in CounterFact may correspond to exact fact recall.

F.2 Total effects

Apart from the analysis described above, we also scrutinize the known CounterFact samples with respect to the total effect of perturbing the subject. We measure the total effect on the probability of the output prediction. This provides an alternative way of checking for signs of lack of exact fact recall. The method was introduced by Meng et al. (2022) and used to find model states important for the model prediction. By adding noise to the word embeddings corresponding to the subject of the query, the subject is perturbed. The idea is that the perturbation of the query makes the model incapable of performing the necessary recall of factual associations that resulted in the original prediction, thus lowering the model probability for the original prediction. We hypothesize that samples for which the added perturbation does not sufficiently lower the corresponding prediction probability are less likely to correspond to exact fact recall.

Method The total effect is measured as $\mathrm{TE}(o) = P_{\text{clean}}(o) - P_{\text{noised}}(o)$, where $P_{\text{clean}}(o)$ denotes the probability of emitting token o for a clean run and $P_{\text{noised}}(o)$ denotes the probability of emitting token o when the subject has been perturbed. For all

[13] There are a total of 205 samples corresponding to person names for which we can check for name bias, meaning that we detect name bias in 92.5% of all cases.

Query	Prediction	Bias type
MacApp, a product created by	Apple	Prompt
Giuseppe Angeli, who has a citizenship of	Italy	Person name
The original language of La Fontaine's Fables is a mixture of	French	Prompt

Table 9: Examples of queries and predictions from the known CounterFact dataset that potentially correspond to bias. The predictions and analysis has been performed for GPT-2 XL.

Popularity score	# of samples
$(0, 100]$	61
$(100, 1000]$	304
$(1000, 10000]$	379
$(10000, 1176235]$	437

Table 10: The popularity scores for the known CounterFact samples. The maximum popularity score measured was 1,176,235.

our investigations, o is given by the prediction corresponding to the query stored in the dataset. We note that negative total effects imply that the perturbation of the subject increased the probability of the original prediction and that low positive effects potentially indicate that perturbing the subject had a small effect on the model prediction.

Similarly to Meng et al. (2022) we perturb the subject embeddings with noise $\epsilon \sim \mathrm{N}(0, \nu)$ where ν is set to be 3 times larger than the empirical standard deviation of all embeddings corresponding to the subjects of the dataset. We measure total effects for the known CounterFact samples as the average total effect of 10 runs with perturbed subjects.

TE results For the 1209 known CounterFact samples we find 22 samples with negative total effects, i.e. perturbing the subject increased the prediction probability, of which 18 potentially correspond to prompt bias and 2 to name bias. Inspection of the samples marked for prompt bias reveal prompt patterns such as "In [X], the language spoken is a mixture of" where the corresponding prediction is "English" or "German". Another pattern we detect is "[X] is affiliated with the religion of" for which the prediction always is "Islam". We hypothesize that some prompts reveal the correct prediction even when the subject is occluded, resulting in negative TE values.

Deeper study of TE results A deeper study of the TE values reveal an additional 37 samples for which the perturbation of the query subject decreased the original probability by less than 40%. For some of these samples we identify queries that potentially reveal the correct prediction even when the subject is perturbed. Two identified samples are "[X] professionally plays the sport of ice [hockey]" or "[X]'s expertise is in the field of quantum [physics]". Prompt bias was detected for all of these queries. We measure a spearman correlation of -0.41 between normalized TE (Equation (3)) and the binary prompt bias metric over all known CounterFact samples. It is clear that the effect of perturbing the subject is smaller when the prediction is likely based on prompt bias, versus when it is not.

$$\mathrm{TE}_{\mathrm{norm}}(o) = \frac{P_{\mathrm{clean}}(o) - P_{\mathrm{noised}}(o)}{P_{\mathrm{clean}}(o)} \quad (3)$$

F.3 Negated queries

We identify a total of 8 problematic samples in the dataset that contain the word "not" in the query. Two examples are "The language used by Louis Bonaparte is not the language of the [French]" or "The expertise of medical association is not in the field of [medicine]". These samples are problematic as they are marked as correct since they contain the correct label, while they express the opposite of the fact represented by the data sample. This problem is a consequence of the sampling technique used by Meng et al. (2022) in letting the LM generate a fluent continuation to a given query before making the prediction for the missing object. For the majority of the known CounterFact samples this leads to more fluent queries for which the LM might work better, but for some samples it results in reversed or revealing prompts.

G Normalization effects on causal tracing results

Since these results are dependent on the absolute values of the probability of the traced (predicted) token, we hypothesize that the result could be driven

by a few high-probability samples and not representative of the low-probability[14] strata of the data.

To test this, we take inspiration from work by Hase et al. (2023) and compare the IE results to their normalized counterpart. We define the *normalized indirect effect* as

$$\text{NIE}_{h_i^{(l)}}(o) = \frac{P_{h_i^{(l)},\,\text{patched}}(o) - P_{\text{noised}}(o)}{|P_{\text{clean}}(o) - P_{\text{noised}}(o)|} \quad (4)$$

where $P_{\text{clean}}(o) - P_{\text{noised}}(o)$ is the total effect (TE) defined as the difference between the clean and the noised runs. The normalized IE measures the percentage of recoverable probability that was recovered by patching state $h_i^{(l)}$.

For some samples, predominantly low-probability predictions, the division by the TE may result in unnatural $\text{NIE}_{h_i^{(l)}}(o)$ values above 1 or below -1. The state patching should not be able to restore more than the clean run probability and we therefore cap the $\text{NIE}_{h_i^{(l)}}(o)$ to a range of $[-1, 1]$. With this approach, each sample is valued on the same scale. Plots for homogeneous datasets should therefore yield normalized CT results that are similar to their non-normalized counterparts.

Are aggregated CT results representative of each studied sample? The non-normalized results for combined samples seen in Figure 6 are dominated by the exact fact recall samples. The exact fact recall samples clearly lead to the decisive role conclusion and the same holds for the non-normalized results, even though subsets of the included data (heuristics recall and guesswork samples) do not lead to the same conclusion with as high certainty.

For the normalized results we find that equal weights for all evaluated samples yield a slightly different pattern compared to the non-normalized results, with a weaker peak for the last subject token. We conclude that aggregations of CT results across multiple prediction scenarios are not representative of each studied sample. Also, comparisons between non-normalized and normalized results may reveal nonhomogeneous datasets with respect to prediction scenario.

[14] With *probability*, we here refer to the probability corresponding to the clean run prediction.

H Additional results from the CT sensitivity analysis

This section contains additional results from the causal tracing analysis of PRISM.

H.1 Llama 2 7B and 13B results

The results in Figures 4 and 5 correspond to the results in Figure 3 but here for Llama 2 7B and 13B instead of GPT-2 XL. We find that the Llama results essentially support the same conclusions as the results for GPT-2 XL.

H.2 Non-normalized results

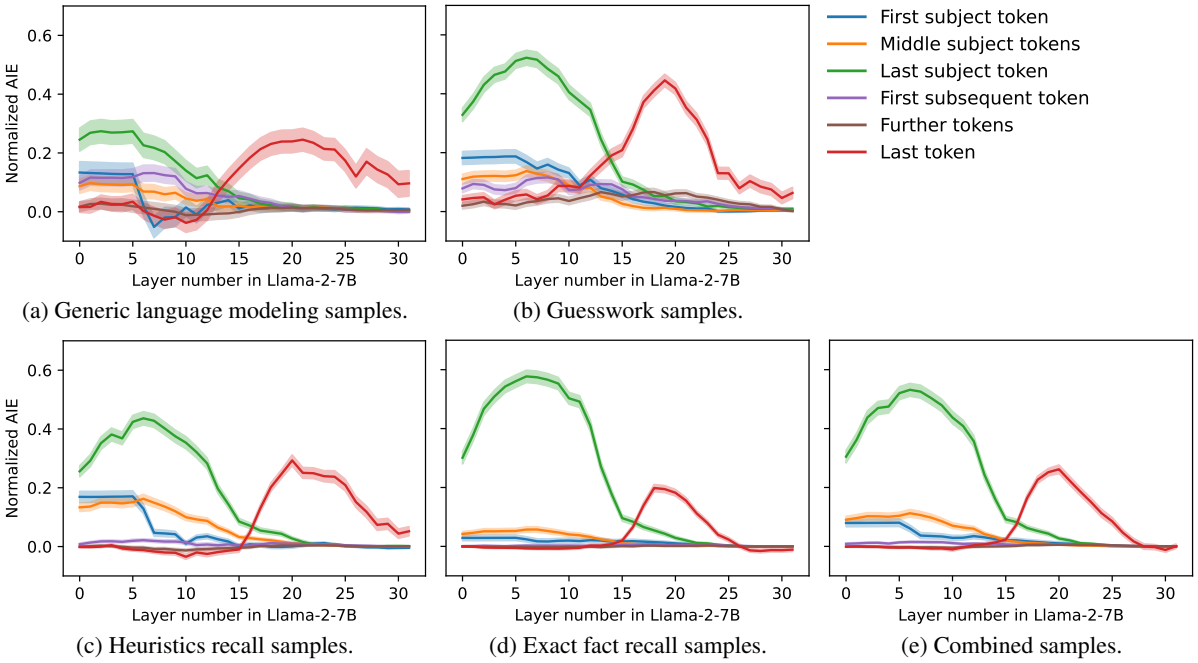
To allow for comparisons with previous work that employed the CT method without normalization we present the non-normalized CT results for the combined samples in Figure 6.

H.3 Low-probability split

The results in Figures 3, 4 and 5 correspond to a sample of top-ranked prediction probabilities. The results in Figures 7 and 8 correspond to a sample of bottom-ranked prediction probabilities. We observe qualitative differences between the two figure pairs, where bottom-ranked probability set corresponds to larger effects for the last token state.

H.4 Deeper study of heuristics recall

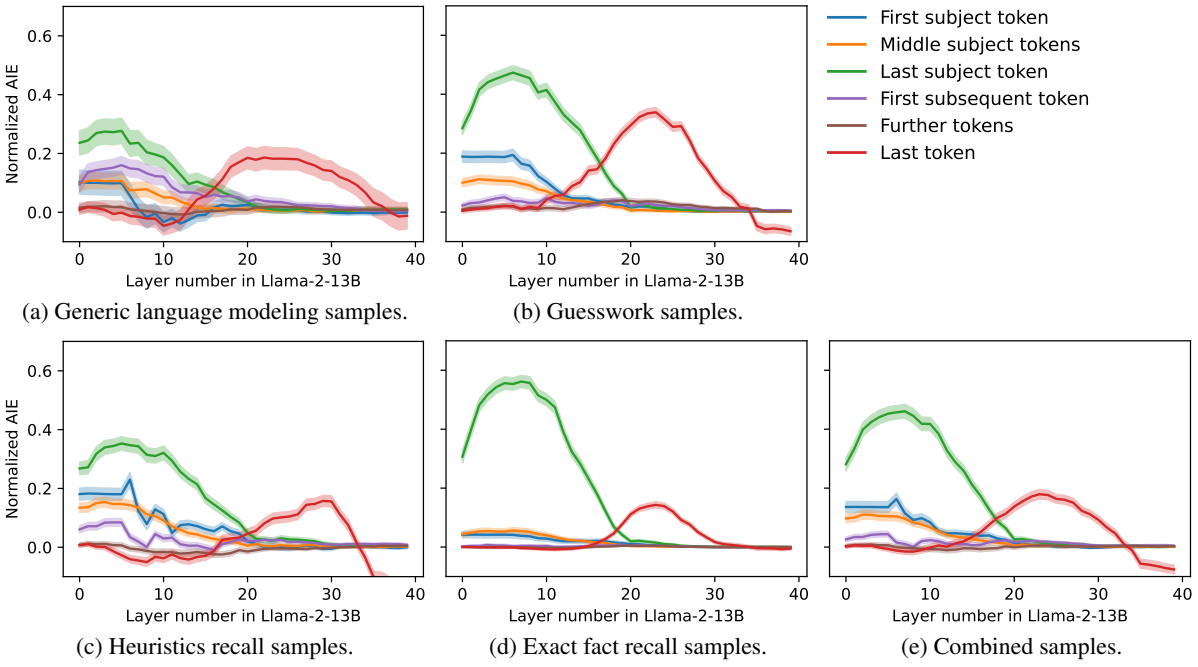
We analyze the CT results of each of the main heuristics recall categories, prompt bias and person name bias, in separation for GPT-2 XL and Llama 2 7B. The corresponding results can be found in Figure 9. These results suggest a higher importance of the last token state, compared to the last subject token state, for the prompt bias subset compared to the person name bias subset. Potentially, it makes sense that prompt biased predictions that should be less sensitive to subject information attribute less importance to states corresponding to the subject.



(a) Generic language modeling samples. (b) Guesswork samples.

(c) Heuristics recall samples. (d) Exact fact recall samples. (e) Combined samples.

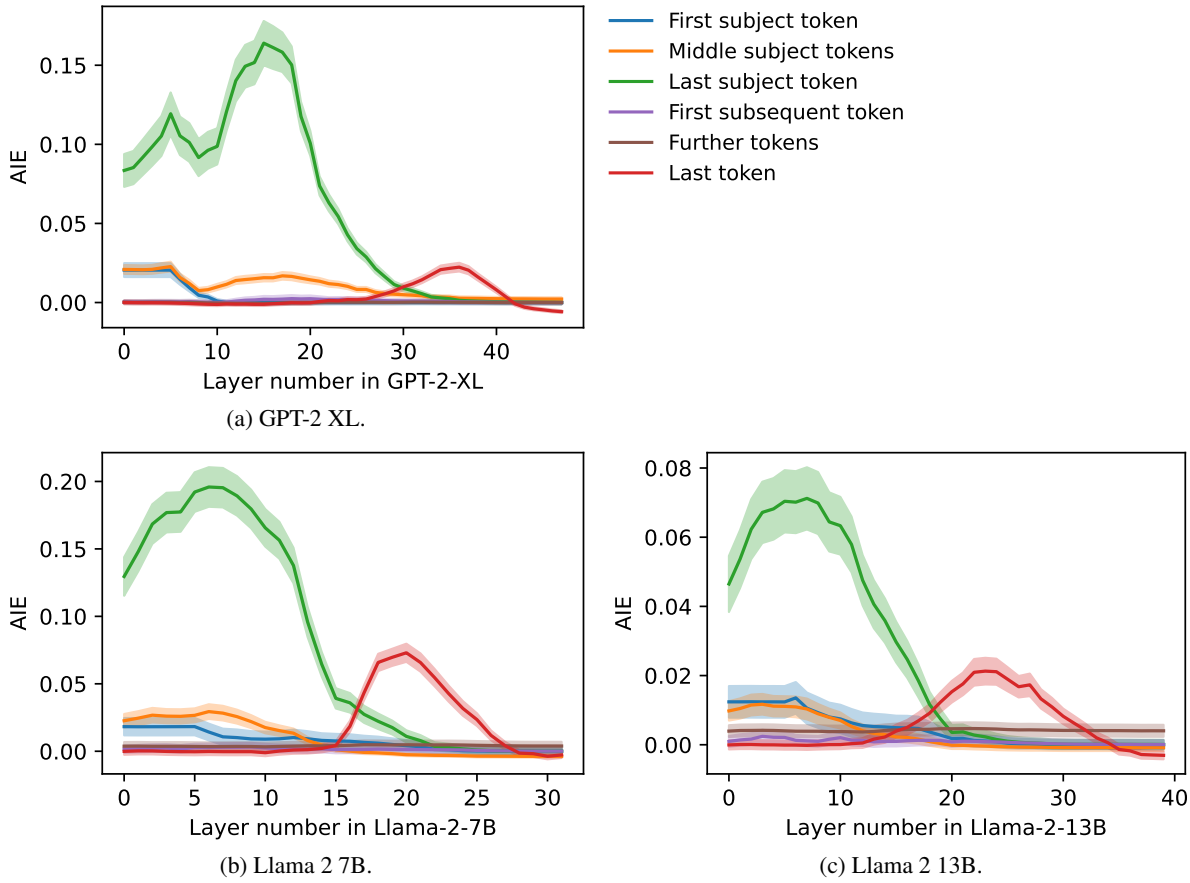
Figure 4: CT results for PRISM Llama 2 7B data. 1000 samples for each scenario in isolation. As well as 1000 combined samples (330 exact fact recall, 340 heuristics recall, 330 guesswork). Shaded regions indicate 95% confidence intervals.



(a) Generic language modeling samples. (b) Guesswork samples.

(c) Heuristics recall samples. (d) Exact fact recall samples. (e) Combined samples.

Figure 5: CT results for PRISM Llama 2 13B data. 1000 samples for each scenario in isolation. As well as 1000 combined samples (330 exact fact recall, 340 heuristics recall, 330 guesswork). Shaded regions indicate 95% confidence intervals.

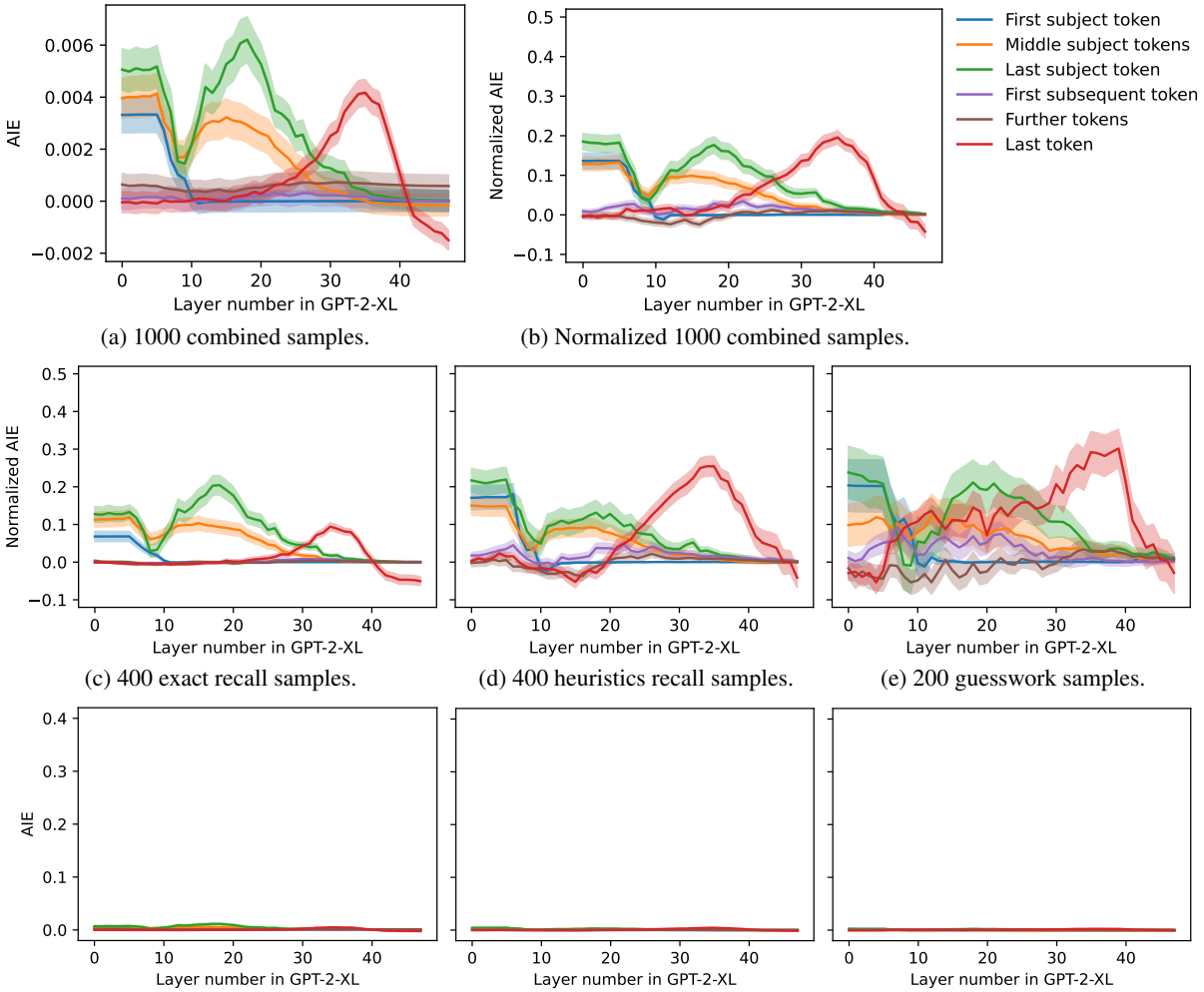


(a) GPT-2 XL.

(b) Llama 2 7B.

(c) Llama 2 13B.

Figure 6: Non-normalized CT results for the combined samples from PRISM for each of our studied models.



(a) 1000 combined samples.

(b) Normalized 1000 combined samples.

(c) 400 exact recall samples.

(d) 400 heuristics recall samples.

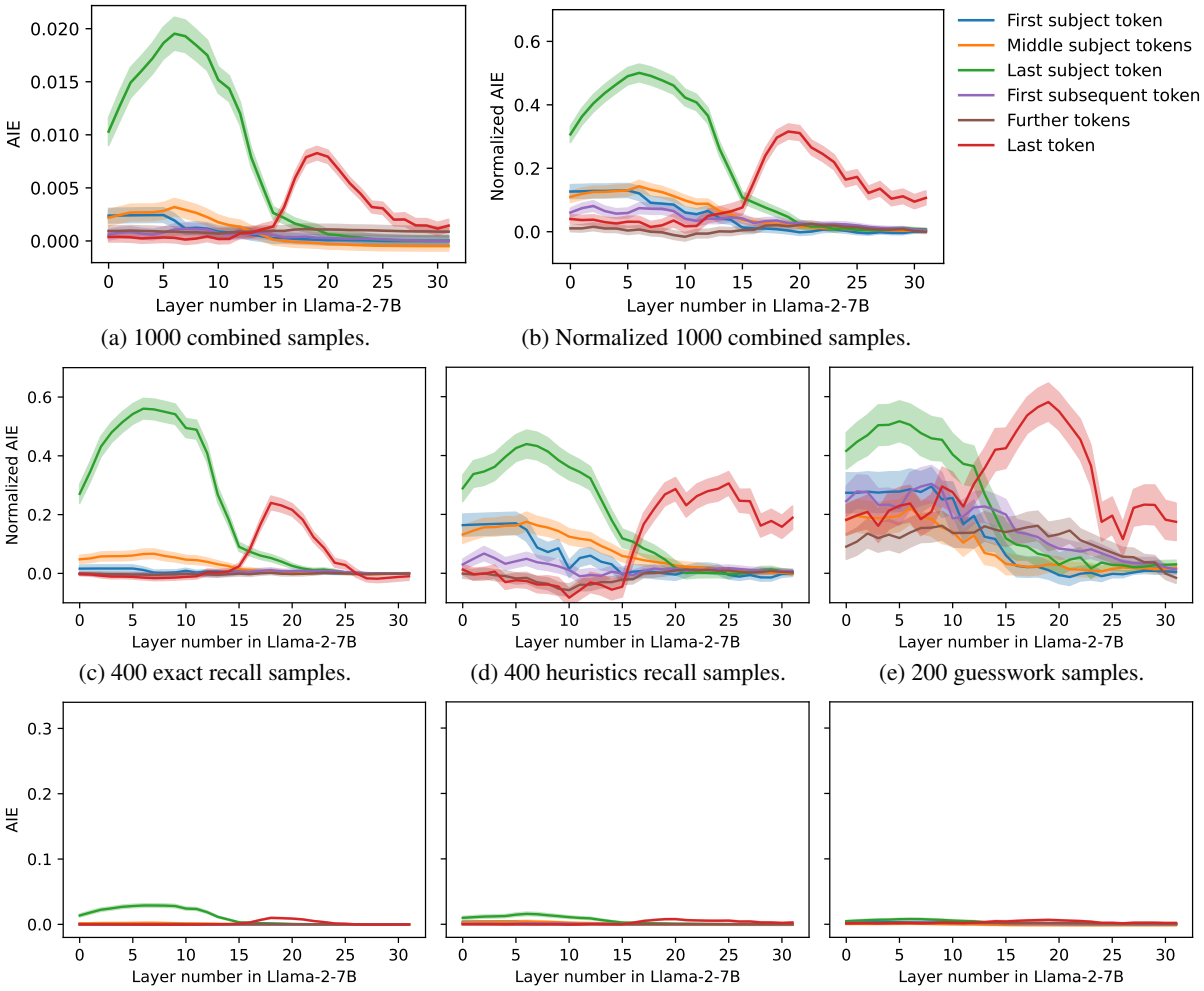
(e) 200 guesswork samples.

(f) 400 exact recall samples.

(g) 400 heuristics recall samples.

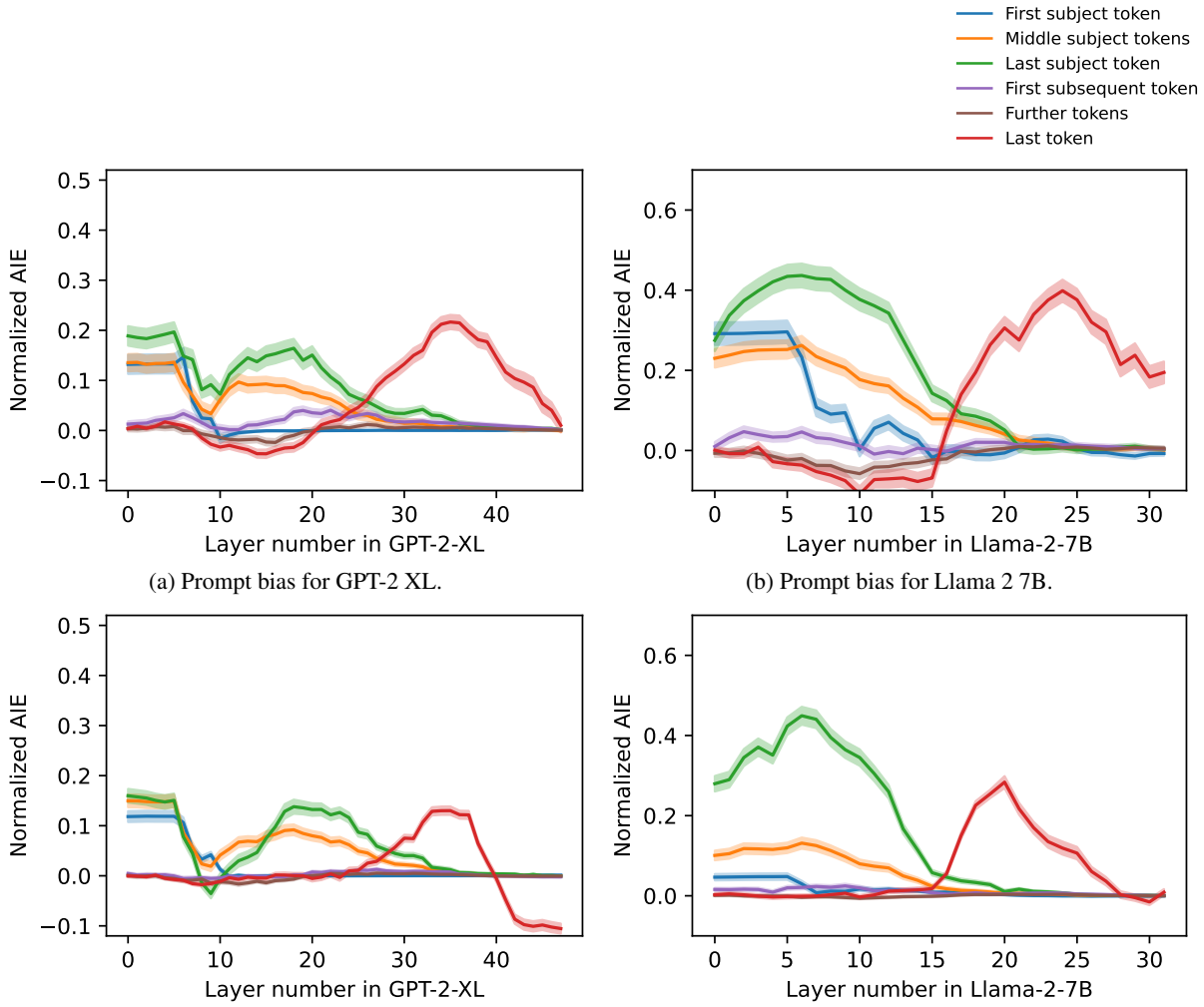
(h) 200 guesswork samples.

Figure 7: CT results on 1000 low-probability samples from PRISM of which 330 samples correspond to exact fact recall, 340 to heuristics recall and 330 to guesswork. These are the results for GPT-2 XL.



(a) 1000 combined samples.
(b) Normalized 1000 combined samples.
(c) 400 exact recall samples.
(d) 400 heuristics recall samples.
(e) 200 guesswork samples.
(f) 400 exact recall samples.
(g) 400 heuristics recall samples.
(h) 200 guesswork samples.

Figure 8: CT results on 1000 low-probability samples from PRISM of which 330 samples correspond to exact fact recall, 340 to heuristics recall and 330 to guesswork. These are the results for Llama 2 7B.



(a) Prompt bias for GPT-2 XL.

(b) Prompt bias for Llama 2 7B.

(c) Person name bias for GPT-2 XL.

(d) Person name bias for Llama 2 7B.

Figure 9: Normalized CT results for sets of 1000 samples designed to exemplify each of the two main categories of the heuristics recall scenario.