"Merge Conflicts!" Exploring the Impacts of External Distractors to Parametric Knowledge Graphs

Cheng Qian[1*], **Xinran Zhao**[2], **Sherry Tongshuang Wu**[2]
[1]Tsinghua University, [2]Carnegie Mellon University
qianc20@mails.tsinghua.edu.cn

Abstract

Large language models (LLMs) acquire extensive knowledge during pre-training, known as their *parametric knowledge*. However, in order to remain up-to-date and align with human instructions, LLMs inevitably require *external knowledge* during their interactions with users. This raises a crucial question: How will LLMs respond when external knowledge interferes with their parametric knowledge? To investigate this question, we propose a framework that systematically elicits LLM parametric knowledge and introduces external knowledge. Specifically, we uncover the impacts by constructing a *parametric knowledge graph* to reveal the different knowledge structures of LLMs, and introduce external knowledge through *distractors* of varying degrees, methods, positions, and formats. Our experiments on both black-box and open-source models demonstrate that LLMs tend to produce responses that deviate from their parametric knowledge, particularly when they encounter direct conflicts or confounding changes of information within detailed contexts. We also find that while LLMs are sensitive to the veracity of external knowledge, they can still be distracted by unrelated information. These findings highlight the risk of hallucination when integrating external knowledge, even indirectly, during interactions with current LLMs. All the data and results are publicly available[1].

1 Introduction

Current large language models (LLMs) have assimilated a significant body of knowledge during pre-training (Chowdhery et al., 2022; Thoppilan et al., 2022; OpenAI, 2022, 2023; Touvron et al., 2023; Anil et al., 2023; Zeng et al., 2022), converting external information from a mass corpus into *parametric knowledge*. However, current LLMs still lack the ability to respond to up-to-date world events, and background information is often required when interacting with them in real-world applications (Trivedi et al., 2023; Yu and Ji, 2023). These challenges often necessitate the introduction of *external knowledge*, either explicitly through retrieval (Shi et al., 2023; Ram et al., 2023), application of tools (Schick et al., 2023; Qin et al., 2023), or implicitly through long prompts that human provides as contextual information.

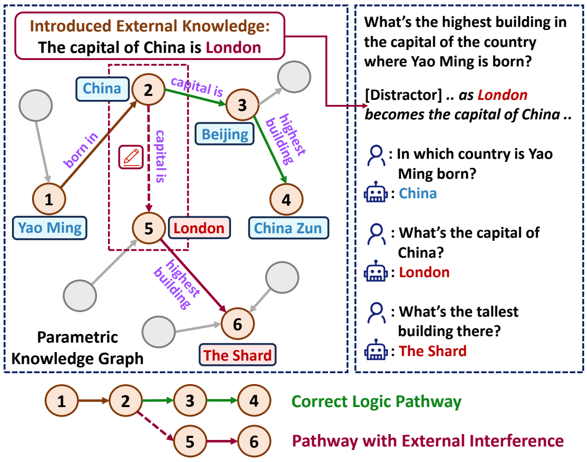


Figure 1: The introduction of external knowledge, such as the distractor "The capital of China is London," creates a *false* relation between entities in the model's parametric knowledge graph. This deviation from the original logic pathway leads to a change in the model's final answer.

The introduction of external knowledge may inevitably interfere with the model's internal parametric knowledge. Prior work (Xie et al., 2023; Neeman et al., 2022) has flagged that when confronting direct conflicts, the model may respond with an answer from either external knowledge or parametric memory. However, merely analyzing direct and explicit conflicts is not comprehensive, as the parametric knowledge within a model is interconnected (Petroni et al., 2019; Wang et al., 2020), and an indirect change in the model's logic pathway may as well change the model's final answer.

[*]Work done as a student intern at CMU.
[1]https://github.com/qiancheng0/EKD_Impacts_PKG

For instance, Figure 1 shows a question that can be decomposed into three hops. Even when we only introduce a *distracting knowledge* to the 2^{nd} hop through prompting (*"The capital of China is London"*), the model's answer to the 3^{rd} hop (final) of the query still shifts accordingly, from the correct *China Zun* to the incorrect *The Shard*.

This phenomenon has recently been referred to as the "ripple effect" (Cohen et al., 2023). Previous works have already proposed benchmarks (Zhong et al., 2023) and metrics (Cohen et al., 2023) for its evaluation, yet 1) they often exclusively consider linear relationships between closely connected knowledge entities and 2) manual efforts are required to construct the external interference and define the extent of radiation for the ripple effect.

To address these limitations, we construct a framework to **evaluate the potential interaction between parametric knowledge and external knowledge in a more systematic manner**. Drawing parallels to the knowledge graph (KG) which contains well-defined connections that can be automatically inspected, we propose *parametric knowledge graph* (PKG), a method that allows us to automatically extract the model's interconnected parametric knowledge into rich and flexible graphs with hundreds of entities and relations. For example, the nodes and solid lines in Figure 1 represent a subgraph of PKG, with entities of countries, humans, and cities, and various relations like *born in*.

Building on the concept of PKG, we further define *distractors*, i.e., a series of external knowledge that interferes with the PKG through prompting, with different degrees, methods, positions, and knowledge formats. Our definition enables us to directly investigate the interactions between distractors and PKGs: In Figure 1, beneath the natural language prompt, the distractor (noted by the dashed line) `(China, Captical is, London)` bridges a *false* connection between two nodes in PKG, and thereby drives a ripple effect that leads to the model's deviation in response.

We study this interaction through experiments on both black-box GPT3.5 and open-source MPT-7B models. Specifically, we first present the distractor to the model, and then conduct queries in an interactive and iterative one-hop manner as shown in the dialogue in Figure 1. We evaluate the *consistency* (whether the final answer adheres to its PKG) and *confidence* (the probability of giving this answer) of the model's responses with the presence of distractors. We observe that in general, LLMs tend to deviate from their parametric knowledge when they are not confident with it to begin with; Interestingly, they always tend to be more confident in their answers when facing external knowledge, regardless of whether that answer comes from the distractor or the PKG. Looking into the impacts of different types of distractors, we discover that as can be expected, posing direct conflicts or giving more confounding changes instead of evidently false information are more powerful (as in Figure 1). However, we are also surprised by many findings, e.g., even *weak* distractors that do not directly interfere with the model's original logic pathway can still impact the model's answers, we can improve the impact of distractor just by hiding it in a lengthier and detailed context, and that GPT-3.5 and MPT-7B shows different trends in what distractors they can best resist.

We conclude by underscoring the inherent risks of hallucination and misinformation when introducing external knowledge, even inadvertently, that interferes with the LLM's parametric knowledge.

2 Related Work

Internal and External Knowledge Conflicts. LLMs amass internal knowledge through extensive learning on massive corpora during pretraining (Roberts et al., 2020; Jiang et al., 2020; Gururangan et al., 2020), thereby weaving a unique system of parametric knowledge. This process, however, can be marred by inaccurate or outdated training data, leading to potential hallucinations within the model (Carlini et al., 2021; Lazaridou et al., 2021; Zhang et al., 2021). To align LLMs with current information and enhance factual accuracy, researchers have employed various tools (Schick et al., 2023; Qin et al., 2023), memory techniques (Zhong et al., 2022), and information retrieval strategies (Guu et al., 2020; Izacard and Grave, 2021). However, such external knowledge may be novel or even contradict the model's existing parametric knowledge, causing interference. Neeman et al. (2022) trained the model to disentangle internal and external knowledge and generate two responses to avoid conflict. Zhou et al. (2023) utilized special prompt engineering and abstention options to improve model faithfulness. More recently, Xie et al. (2023) explored how the GPT model family reacts to knowledge conflicts, uncovering a high receptivity to external knowledge and confirmation bias. In line with these studies, we broaden our focus to encompass both black-box and open-source models, adopting a more systematic perspective on multiple types of distractors and parametric knowledge structures.

Propagation of Introduced Knowledge. Prior approaches to model editing have primarily centered on the modification of parameters (Meng et al., 2022; Yao et al., 2022) or the integration of specialized modules (Wang et al., 2021; Mitchell et al., 2021) to enable the model to assimilate new knowledge. Nevertheless, this newly introduced external knowledge is anticipated to exert long-lasting effects. Onoe et al. (2023) have found that traditional editing methods exhibit inconsistencies when paraphrasing questions in new contexts, and that prepending entity definitions can facilitate the propagation of the injected external knowledge. Zhong et al. (2023) contributed a benchmark to measure how the alteration of one knowledge piece may influence the entire multi-hop QA chain's response. This phenomenon is termed the ripple effect by Cohen et al. (2023), who offer six evaluation metrics from this angle to assess the robustness of the model editing methods. Building on these investigations, we extend our focus to more hops and multiple knowledge structures, and build a generalizable framework to explore the impact of introduced external knowledge, particularly during the model's active engagement with users, and in a more controlled and systematic manner.

3 Introduction of Parametric and External Knowledge

Our research question focuses on the impacts of *external knowledge* on *parametric knowledge*. We discuss how we extract the parametric knowledge in a *graph structure* to capture their relations, and introduce various types of external knowledge as *distractors*.

3.1 Parametric Knowledge Graph

LLMs have learned through pre-training a mass amount of parametric knowledge that is largely interconnected. As mentioned in Section 1, these interconnections impact how LLMs react to different direct or indirect distractions in external knowledge. To reveal such internal knowledge structure, we propose a novel framework to construct a model's *parametric knowledge graph* (PKG).

PKG is a *semantic net that integrates a model's parametric knowledge*. It consists of nodes that represent *entities* (denoted by E), and edges that represent *relations* (denoted by R). Drawing parallels to Knowledge Graph (KG), PKG allows for turning the *implicit knowledge* within a model into *explicit and structured* representations that are transparent for inspection and flexible for extension. Its *automatic* construction also provides convenience for the extraction and modification of parametric knowledge. While traditional KGs are grounded in real-world facts (Fensel et al., 2020), our approach stands as the first to use KG as an analogy for eliciting the model's parametric knowledge.

Construction To automatically construct the PKG, we provide a set of *specification rules*, which defines what kinds of relations can exist in which types of entities. More concretely, we abstract each entity E (e.g., France) in PKG into a *type* ("Country"), and create rules for each type in the form of (R, `target type`). For example, the entity type "Country" can extend the relation "capital is", targeting an answer of type "City." In practice, these rules are applied in natural language templates (Appendix A), which closely map the LLM's logic pathway to the graph in an interpretable way.

Once the rules are set, a PKG can be automatically extended in a depth-first manner for any given root node (with an entity E and its corresponding *type*). For example, in Figure 2B, upon assigning the root node to "Canada", our framework sequentially extends all relations associated with the root type "Country". The model then seeks an answer corresponding to each target type, recursively shaping the whole PKG. Experimentally, we apply consistency checks and only regard the answers that the model sticks to during consecutive queries as parametric knowledge (Appendix A).

Extraction One core advantage of PKG is that it enables us to *extract data chains with structural variety*: As in Figure 2C, PKG contains not only multi-hop structures, but also multi-child (where multiple answers exist for a given entity and relation) and multi-dependent relations (where two indispensable entities jointly decide the answer for a relation). Such complexity enables the analyses of different types of relations in parametric knowledge.

To support controlled experiments, we extract the sub-graphs from PKG with different structures,

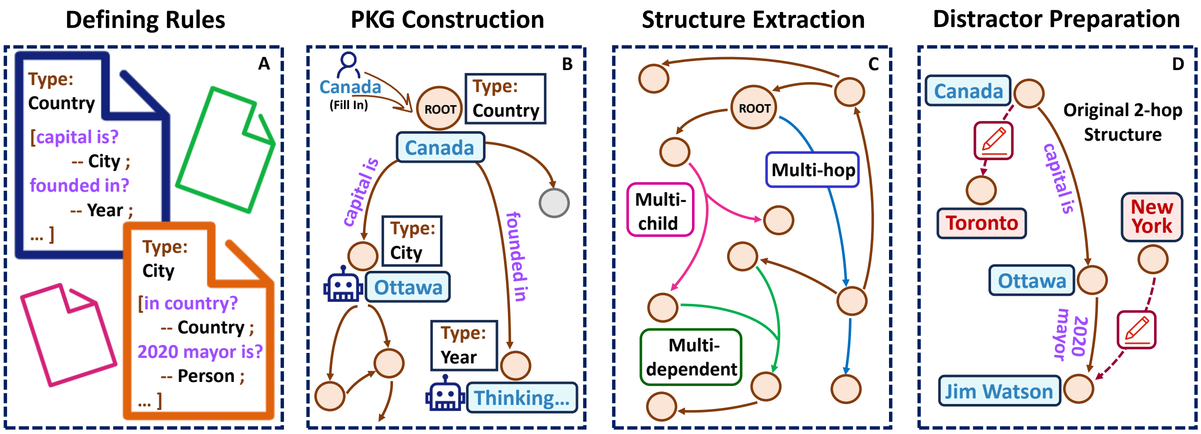


Figure 2: The pipeline for the construction of PKG and distractors. Figure A, B: The automatic construction of the model's PKG with defined rules. Figure C: The extraction of various PKG structures. Figure D: The modification of PKG to construct distractors.

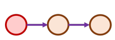
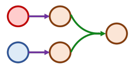
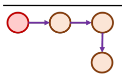
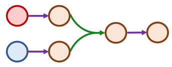
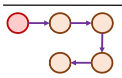
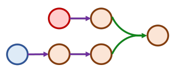
Multi-Hop				Multi-Dependent			
Structure	**Name**	**Hops**	**Example**	**Structure**	**Name**	**Hops**	**Example**
	**2-Hop** Structure	**2**	Who's the mayor of the capital of **China** in 2020?		**1-1-0** Structure	**3**	Who is the president of the country to which **Beijing** belongs in the year when **Michael Jackson** is born?
	**3-Hop** Structure	**3**	In which province is the mayor of the capital of **China** in 2020 born?		**1-1-1** Structure	**4**	What's the birthplace of the president of the country to which **Beijing** belongs in the year when **Michael Jackson** is born?
	**4-Hop** Structure	**4**	What's the largest city in the province where the mayor of the capital of **China** in 2020 is born?		**1-2-0** Structure	**4**	Who is the president of the country to which **Beijing** belongs in the year when the singer of album **Dangerous** is born?

Table 1: Three multi-hop and three multi-dependent structures we investigate in the experiments. The red and blue nodes represent the starting entities, while other nodes are implicit (need reasoning to reach instead of directly given in the query *Examples*). The green edges denote the multi-dependent relation, which is contained in the *pivot hop*, while other purple edges denote other explicit relations.

nodes, and edges. To transform them into a usable format, we linearize each sub-graph into a "data chain" represented by triplets $[(E_0, R_1, E_1), (E_1', R_2, E_2), ..., (E_{n-1}', R_n, E_n)]$ ($E_k = E_k'$ for multi-hop chains). As shown in Table 1, we mainly apply the multi-hop (2, 3, and 4-hop) structures as the basis of queries in experiments; To further capture the non-linearity, we define three *multi-dependent structures* (Trivedi et al., 2022), each containing multiple hops and a core multi-dependent relation (Table 1, right). We denote the hop that contains the multi-dependent relation in the data chain as the *pivot hop*.

The answer for the pivot hop relies on two upstream entities, and both of them can serve as the ending node for multi-hop chains of lengths A and B, respectively. Simultaneously, the answer entity for the pivot hop can serve as the starting node for a multi-hop chain of length C. The values of A, B, and C collectively govern the specific configuration of the multi-dependent structure, succinctly referred to as the A-B-C structure.

3.2 External Knowledge Distractors

The external knowledge in our experiments is introduced through *distractors*, which serve as the source of interference to the model's PKG.

Distractors are directly derived by modifying the extracted raw data chains from the model's PKG. Figure 2D shows a simplified example where the capital of "Canada" is substituted with "Toronto", and the subject of "2020 mayor" is replaced with "New York". In both cases, distractors are derived from modifying a 2-hop chain. We prompt GPT-3.5 for the distractor's automatic construction, as detailed in Appendix B and Figure 18 to Figure 23. The distractor will be presented as a natural language description at the beginning of model-user interaction shown in Figure 1.

We systematically create distractors by varying

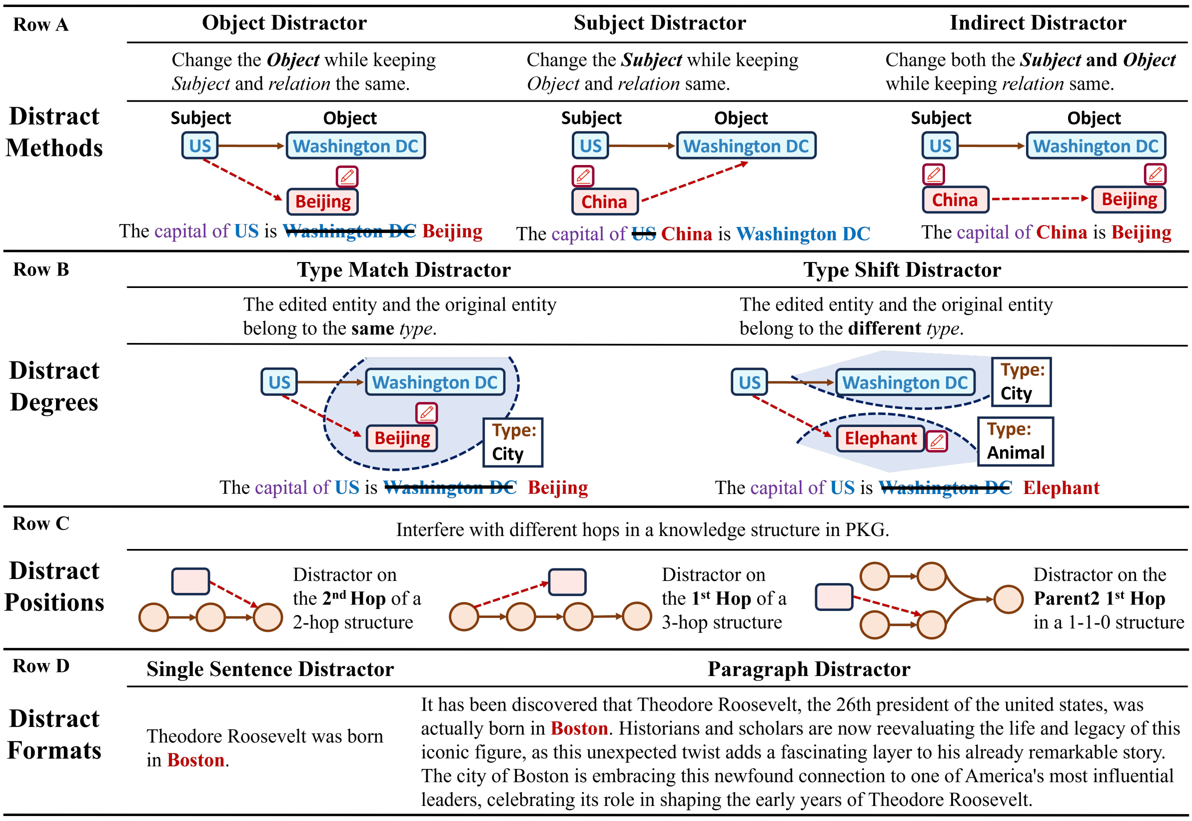


Figure 3: An Illustration of different types of distractors we apply in experiments.

four dimensions (Figure 3): distract methods, degrees, positions, and formats, as detailed below. These types of distractors allow for a nuanced exploration of how alterations in different components of the PKG can lead to varied effects in the model's responses.

Distract Methods Different distract methods reveal how the external knowledge is related to the original parametric knowledge. As in Figure 3A, *Object Distractor* introduces distraction by changing the object of the original parametric knowledge in the raw data chain. For instance, by changing the original object "Washington DC" into "Beijing" while preserving the subject and relation, the resulting external knowledge "The capital of US is Beijing" constitutes an *Object Distractor*. *Object Distractor* often represents an explicit contradiction to the model's original parametric belief.

Similarly, *Subject Distractor* introduces distraction by changing the subject, while *Indirect Distractor* changes both the subject and the object while preserving the relation. As we always query for the *Object* in knowledge given the *Subject* and *Relation*, this makes *Object Distractor* an explicit contradiction to the model's original parametric belief, while the other two are "weaker". This distinction may lead to different resulting impacts.

Distract Degrees Various distract degrees illustrate how severely the external knowledge deviates from the original parametric knowledge (Figure 3B). This deviation is measured through *type*: We define the distractor as *Type Match* if the edited entity and original entity belong to the same type (e.g., city "Washington DC" to city "Beijing"), and *Type Shift* if otherwise (e.g., city "Washington DC" to animal "Elephant.") Because of the change in the underlying type, *Type Shift Distrcators* are often evidently false information, while *Type Match Distrcators* are more confounding to models. This makes distract degrees important as they reflect how well could the models accept knowledge with different credibility.

Distract Positions Built upon distract methods and degrees, different distract positions reflect to which relation in the extracted data chain is the external information introduced. This attribute of the distractors doesn't describe how to concretely modify the knowledge into distractions, but rather where to introduce this distraction. In Figure 3C, we present three examples of different distract po-

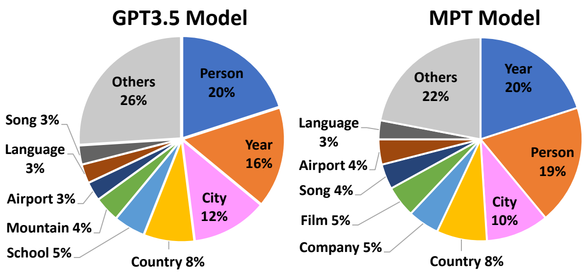


Figure 4: Ratio of different *types* in model's PKG.

Dimensions / Model		**GPT3.5**	**MPT-7B**
Avg Node Num		278	166
Avg Edge Num		467	276
Multi-dependent Rels		769	443
Multi-child Rels		192	124
Multi-hop Structures	2-hop	5,361	3,360
	3-hop	14,523	8,642
	4-hop	28,297	17,064

Table 2: The statistics of 8 PKGs we apply respectively for GPT3.5 and MPT-7B. *Rels* denotes relations. On average, GPT3.5 exhibits a more expansive and intricate PKG. The magnitude of distinct relations and varied structures within the PKGs exemplifies their heightened diversity and complexity.

sitions. The total number of positions that external knowledge could be introduced is decided by the total number of hops in a knowledge structure. Different distract positions in essence represent different stages in the evolving user-model interaction. For multi-dependent structures, we can also utilize distract positions as a means to distinguish the unique impact of introducing distractions to the pivot hop.

Distract Format The distract format differentiates the context of external knowledge. Based on the context length, we introduce *Single Sentence Distractor*, which states the external knowledge in one simple sentence, and *Paragraph Distractor*, which describes the knowledge through 3-4 sentences with supporting details. In *Row D* of Figure 3, we illustrate how a simple piece of external knowledge "Theodore Roosevelt was born in Boston" can be extended to a paragraph. Different distract formats are introduced to prove whether the model possesses a certain bias towards lengthier and more detailed descriptions.

4 Experiment Setup

To understand the model's reaction when external knowledge interferes with its parametric knowledge, we conduct experiments by systematically varying the combinations of the model's parametric knowledge structures and the external knowledge distractors.

4.1 Method

Each structure we extract from PKG contains multiple hops of queries, which constitute the data chain that represents the model's original logic pathway. As we aim to inspect the model's responses during active interaction instead of testing its multi-hop reasoning ability, we follow the "instance-wise" probing method proposed by Zhong et al. (2023) and test the data chain in a one-hop manner after the introduction of distractors.

As illustrated previously in Figure 1, we first present the distractor to the model as the introduced external knowledge. Next, for the data chain $[(E_0, R_1, E_1), (E_1', R_2, E_2), ..., (E_{n-1}', R_n, E_n)]$ extracted, where two entities E and a relation R form a knowledge triplet, we first probe for the model's answer A_1 after given (E_0, R_1). Then, we continue to probe the next hop of the query, while giving the model all previous interaction history. The new query is based on (A_1, R_2) if $E_1' = E_1$ (this always holds for multi-hop structures) or based on (E_1', R_2) if $E_1' \neq E_1$ (this only happens for multi-dependent structures), and we ask for the model's answer A_2. This iterates until all the queries are done or the model abstains from answering.

Controlled Settings To control the variables in our experiment, for all the studies except knowledge structures, we experiment on all the *multi-hop* structures as raw data chains. For all the studies except the external knowledge format, we apply *Single Sentence* as the distractor's knowledge format. Please refer to Appendix C for a more detailed explanation of each experimental setting.

4.2 Models

We conduct experiments utilizing both the open-source MPT-7B (ML, 2023) and the black-box GPT3.5 models (OpenAI, 2022). MPT-7B and GPT3.5 are selected for their robust interaction capabilities with users, which aligns well with our experimental design. Furthermore, the open-source nature of MPT-7B enables the analysis of confidence values. Please also refer to Appendix C for more details about the hyper-parameters we apply. We also present some experiment results from GPT3 as additional support to our findings in Appendix E.

4.3 Data

We construct 8 PKGs with different root nodes for both GPT3.5 and MPT-7B using manually defined rules. In total, we used 17 types and 63 relations in the construction rules. The statistical findings of the raw PKGs we apply are summarized in Figure 4 and Table 2. For all the studies besides the knowledge structures in PKG, we employ N-hop data chains ($N \in 2, 3, 4$), utilizing 200 chains for each type. Each N-hop data chain affords N positions for external knowledge introduction, three distract methods, and two distract degrees, resulting in $6N$ rounds of queries (or, $6N$ different distractors) and $6N^2$ hops of queries per original chain. Consequently, these constitute 10,800 query rounds, encompassing a total of 34,800 query hops.

For the study on knowledge structures in PKG, we extract 100 raw data chains for each multi-dependent structure type illustrated in the right column of Figure 1. The collected chains constitute 6,600 query rounds, encompassing a total of 24,600 query hops. The tool for automatic PKG construction and all the data we apply is released.

4.4 Metrics

Consistency Our primary metric is *consistency*, which quantifies whether the model will stick to the answer in its PKG during multiple rounds of queries even with the presence of distractors. Formally, among N query chains $C_1, C_2, ..., C_N$, the model outputs the *final answer* that adheres to its PKG in M chains. *Consistency* is defined as:

$$\text{Consistency}(\{C_1, ..., C_N\}) = \frac{M}{N}.$$

It reflects the ratio of query chains from PKG that are not affected by the external distractors.

Confidence Moreover, inspired by Kadavath et al. (2022), we also explore the MPT-7B's likelihood of outputting the target entity through the computation of *confidence*. Given the tokens $t_0, t_1, ..., t_M$ of a core entity E in the model's response, the model's confidence in outputting this entity as the answer is defined as:

$$\text{Confidence}(E) = \prod_{t=t_0}^{t_M} \frac{e^{z_t}}{\sum_{i=1}^{N} e^{z_i}},$$

where z_t denotes the raw score (logit) associated with the token t, and N denotes the total number of tokens in the vocabulary.

Diving deeper into the model's responses, we further look into concrete model behaviors both at the final answer and across the chain. First, for the **final answers**, we split *inconsistent* answers into *abstention* (when a full query chain cannot be completed because the model starts to answer e.g., "I don't know" at certain hops in a chain), and *variation* (when the model reaches a final answer, but not same as the original PKG). Then, **across the chain**, we also study if the model's answer for a particular hop of the query adheres to its PKG or not by categorizing each model's responses as either *conforming* (when the model answers as its original PKG) or *deviated* (when the model's answer is from the distractor).

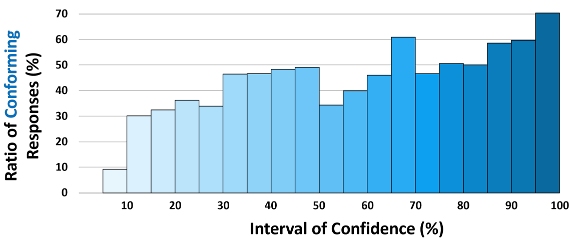


Figure 5: Statistics of the ratio of conforming responses with respect to the confidence placed in the corresponding relations in PKG. From left to right, as the confidence in a particular relation in PKG rises, the model is more likely to provide answers that conform with the PKG despite the presence of distractions.

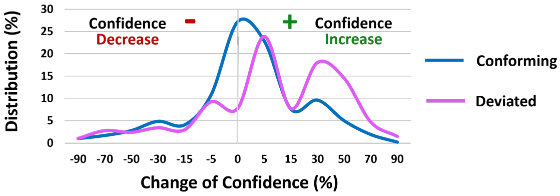


Figure 6: The distribution of the *change* of confidence after introducing the distractors with respect to conforming and deviated responses. The area under the curve left of 0 represents the ratio of negative confidence change, and vice versa. With external knowledge, the model's confidence generally increases, especially for deviated responses.

Distractors		GPT3.5 Consistency (%)	MPT-7B Consistency (%)	MPT-7B Confidence (%)	Conclusions
A: Multi-hop Structures (2 / 3 / 4 hops), Macro Avg.					
Distract Degrees	Type Match	55.90	42.78	80.92	Models resist knowledge that evidently lacks veracity, but they still make the model more uncertain generally.
	Type Shift	58.50	46.22	78.35	
Distract Methods	Object	42.90	35.71	80.53	*Object Distractors* most easily mislead the models, but even indirect "weaker" distractors exhibit interference.
	Indirect	60.08	43.97	77.20	
	Subject	67.62	54.17	81.22	
Distract Positions	1^{st} Hop	61.78	42.11	78.68	GPT3.5 exhibits defense against distractions in the first hop, while MPT-7B easily gets distracted in the beginning. Consistency generally rises as the interaction evolves (explicit in later analysis).
	2^{nd} Hop	53.53	45.25	80.77	
	3^{rd} Hop	54.00	43.25	79.17	
	4^{th} Hop	54.00	40.17	78.67	
Distract Format	Sentence	57.20	44.51	79.65	Lengthier and more detailed contexts lower the model's consistency.
	Paragraph	54.37	42.06	78.26	
B: Multi-dependent Structures (1-1-0, 1-1-1, 1-2-0), Macro Avg.					
Distract Degrees	Type Match	48.55	32.89	78.60	The conclusions for multi-dependent structures on distract degrees and methods remain the same as those for multi-hop structures.
	Type Shift	50.23	35.12	76.81	
Distract Methods	Object	36.40	24.72	77.70	
	Indirect	49.58	33.29	76.51	
	Subject	62.19	44.00	78.91	
Distract Positions	Parent1 1^{st} Hop	54.83	30.56	77.07	The *pivot hop* exhibits special traits. Introducing distractors to interfere with the *pivot hop* results in the lowest consistency for GPT3.5 while highest consistency for MPT-7B.
	Parent2 1^{st} Hop	54.22	31.72	75.84	
	Parent2 2^{nd} Hop	51.00	32.50	78.52	
	Pivot Hop	39.44	39.22	80.20	
	Child 1^{st} Hop	49.33	37.83	78.67	

Table 3: An overview of our experimental results and conclusions. We provide results on GPT3.5 and MPT-7B's consistency and confidence (macro-average on different structures), investigating the impacts of various distractors on multi-hop and multi-dependent structures in PKG. Results that we focus on are shaded: green is applied for the *highest* numerical value for a distractor type, while red is applied for the *lowest*. Please refer to the Appendix for detailed scores of different structure types instead of macro-average.

5 Experiment Results

5.1 Effectiveness of Distractors through Confidence Analysis

Before diving into the specific effects of different distractors, we first analyze in general *why the distractors we introduce are effective*. We conduct analysis through the lens of confidence to unveil the mechanism behind the model's responses under distractions.

Consistency occurs with high confidence. We observe that the model is more likely to provide responses conforming to PKG when it is already confident with this piece of knowledge in PKG. In Figure 5, we show the ratio of conforming responses generally rises as the confidence of that queried relation in the model's PKG increases. From another perspective, this also indicates if the model lacks confidence in a particular knowledge from PKG initially, the distractor is more likely to succeed in causing deviation during later queries.

Response deviates with raised confidence. We find that the model's confidence generally *rises* with the introduction of external knowledge, especially for the deviated responses. In Figure 6, we measure how the confidence *changes* with the introduction of distractors. We plot the distribution of changes respectively for the conforming and deviated responses, and discover that: i) The area under positive confidence change is generally larger, indicating that external knowledge in general bolsters the model's confidence. ii) Most of the deviated responses experience an increase in confidence, proving that the distractors we introduce can deviate the model's responses with generally higher confidence.

5.2 Results on Different Distractor Types

After gaining a general understanding of how distractors are effective through their interactions with

Distractors	GPT3.5 Abstention (%)	GPT3.5 Variation (%)	MPT-7B Abstention (%)	MPT-7B Variation (%)
Type Match	39.67	60.33	5.46	94.54
Type Shift	68.30	31.70	8.30	91.70

Table 4: The rate of abstention and variation in the inconsistent chains when confronting distractors of different degrees for both GPT3.5 and MPT-7B. The results are the macro-averages on three multi-hop structures. *Type Shift Distractors* cause more abstentions.

the model's confidence, we continue to investigate the impacts of distract degrees, methods, positions, and knowledge formats. The numerical results and major conclusions are presented in Table 3A.

Distract Degrees: Models exhibit resistance to knowledge that evidently lacks veracity. We discover that compared to *Type Match Distractors*, *Type Shift Distractors* are less successful in misleading the model's responses. In the first row of Table 3A, we show the consistency is higher on *Type Shift Distractors* for both GPT3.5 and MPT-7B. The observations are significant: With a Student's t-test, we obtain $p < 0.001$ in both cases (Appendix D.1). As *Type Shift Distractors* changes the type of the edited entity and often yields external knowledge beyond commonsense (e.g. "The capital of US is *Elephant*"), our results demonstrate the LLMs are resistant to such knowledge that obviously lacks veracity.

Nevertheless, the overall confidence of the model's responses decreases for *Type Shift Distractors*, suggesting that while the model may reject such distractors, the presence of these severely altered information can still exert strong effects of uncertainty.

We also investigate the inconsistent chains and discover that: i) Compared to MPT-7B, GPT3.5 is more likely to abstain under interference. ii) Compared to *Type Match Distractors*, *Type Shift Distractors* are more likely to cause abstention. The error analysis is presented in Table 4. We provide more detailed results and further analysis to confidence in Appendix D.1.

Distract Methods: *Object Distractors* lead to the lowest consistency, while *Indirect Distractors* also exhibit interfering effects. We observe that *Object Distractors*, which bring directly conflicting external knowledge, particularly drive the model's deviation from its PKG in responses. As presented in the second row of Table 3A, consistency is the lowest for *Object Distractors* among the three methods for both GPT3.5 and MPT-7B (P-value $p < 0.001$ in all cases with Student's t-test, detailed in Appendix D.2). This could be attributed to the fact that only *Object Distractors* retain the subject of original knowledge, thus forging an erroneous relation link that diverges from the model's original logic pathway. This more easily leads the model to a false final answer.

For inconsistent chains, we discover that *Object Distractors* lead to the lowest abstention in GPT3.5 but the highest in MPT-7B. This indicates that among the three distract methods, GPT3.5 is less likely to abstain if the distractor is a direct conflict, while MPT-7B shows vice versa. The error analysis is presented in Table 4. We also provide more detailed results and further analysis to confidence in Appendix D.2.

Distractors	GPT3.5 Abstention (%)	GPT3.5 Variation (%)	MPT-7B Abstention (%)	MPT-7B Variation (%)
Object	48.80	51.20	8.75	91.25
Indirect	56.75	43.25	5.29	94.71
Subject	58.85	41.15	6.21	93.79

Table 5: The rate of abstention and variation in the inconsistent chains when confronting distractors applying different methods for both GPT3.5 and MPT-7B. The results are the macro-averages on three multi-hop structures. *Object Distractors* cause the least abstention in GPT3.5, while the most for MPT-7B.

Distractors	Object	Indirect	Subject
Confidence (%)	77.70	74.31	71.90

Table 6: The average confidence of relations in the PKG that the model deviates in later responses. "Weaker" distract methods tend to mislead the model on the knowledge that it is originally not confident about.

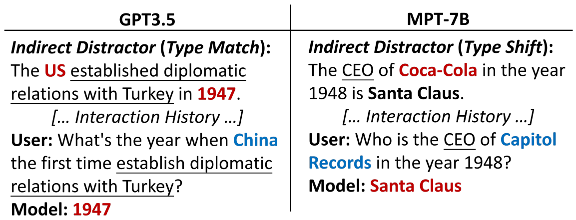


Figure 7: Case study on how the model deviates in response under *Indirect Distractors*. The weak belief of target knowledge in its PKG and some intrinsic similarity in details mislead both models.

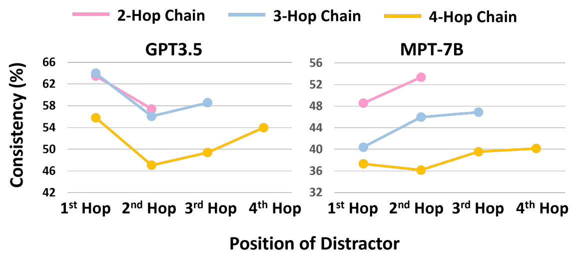


Figure 8: The consistency with respect to which position is the distractor introduced to. MPT-7B's consistency rises gradually as the position to introduce the distractor moves backward, while GPT3.5 shows a high consistency if the distractor is introduced to the first hop (beginning) of the query chain.

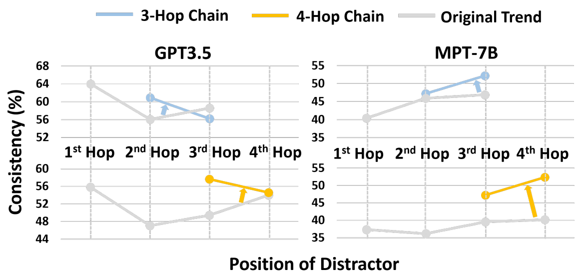


Figure 9: The consistency with respect to querying only the last two hops in the 3-hop and 4-hop chains. While MPT-7B still maintains an upward trend, GPT3.5 exhibits a downward trend similar to that of a 2-hop chain (pink) in Figure 8.

Metrics		Conforming Res.		Deviated Res.	
		Match	Shift	Match	Shift
Change of Confidence (%)	2-hop	-1.55	-2.34	-2.68	+2.42
	3-hop	-1.25	-0.69	-0.50	+5.69
	4-hop	-0.94	-0.21	+0.51	+0.40

Table 7: The *change* of confidence with respect to distractors of different degrees when the knowledge format becomes lengthier and more detailed. We discover the model's confidence rises mostly in the deviated responses to *Type Shift Distractors*.

Despite the direct misleading effect of *Object Distractors*, we observe *Subject* and *Indirect Distractors* also exhibit interference. Both methods alter the subject and make the introduced external knowledge not explicitly align with the query, so they are "weaker" distract methods that should not directly mislead the model. Why these "weaker" distractors also lead to low consistency?

Inspired by previous studies in Section 5.1, we analyze the confidence of relations within the PKG that the model subsequently deviates in its responses. We discover when confronting a "weaker" distract method, the model is more susceptible to deviate in queries where its initial belief in the PKG is also weak. As showcased in Table 6, the average confidence of relations that later lead to deviations is the lowest for *Subject Distractors*, followed by the *Indirect Distractors*.

Specifically, we present a case study in Figure 7: For GPT3.5, the uncertainty about the user's query drives it to extract "1947" from the distractor as the final answer, despite the subject in distractor being "US" rather than "China"' as inquired. The same happens to MPT-7B, as the same additional information "in the year 1948" presented in both the query and the distractor drives the model to trust "Santa Claus" as the company's CEO, though it is evidently false.

Distract Positions: Models resist distractions as interaction evolves, and GPT3.5 also defends against early distractions. We discover the consistency of both models generally rises as the position where the distractors are introduced moves toward the tail of the data chain. In addition, GPT3.5 is more likely to resist external knowledge if it is introduced in the beginning. To better illustrate the trends, we differentiate three knowledge structures and plot their consistency respectively in Figure 8. For both models, the gradual rising trend could be attributed to their declining attention towards the distractor as the query goes on: as the chatting history accumulates, both models become harder to deviate, thus introducing a distractor that interferes with later hops raises consistency. In addition, the high consistency of GPT3.5 if distracted initially implies it maintains a heightened sensitivity to the veracity of external knowledge, but MPT-7B lacks such a mechanism.

To mitigate the influence of queries themselves, we conduct an ablation study by querying only the last two hops of the 3-hop and 4-hop chains. From the results in Figure 9, we discover that: i) Overall consistency increases as a shorter data chain is applied. ii) While MPT-7B's consistency still rises, GPT3.5's consistency declines. The higher consistency observed when distractors are introduced to interfere with the first hop provides additional evidence of GPT3.5's heightened vigilance in the beginning towards information that deviates from its PKG. We provide additional detailed results in Appendix D.3.

Metrics		Conforming Res.			Deviated Res.		
		Obj.	Indir.	Sbj.	Obj.	Indir.	Sbj.
Change of Confidence (%)	2-hop	-2.21	-1.55	-2.13	-4.52	+3.58	+1.80
	3-hop	-1.24	-0.67	-0.91	-1.93	-0.28	+9.26
	4-hop	-0.66	-1.03	-0.56	-1.80	-0.16	+5.03

Table 8: The change of confidence with respect to distractors applying different methods when the knowledge format becomes lengthier and more detailed. *Obj.*, *Indir.* and *Sbj.* respectively denotes *Object*, *Indirect*, and *Subject Distractors*. We discover the model's confidence tends to rise on deviated responses for "weaker" distract methods.

Distract Formats: Lengthier distract contexts lower the consistency. We discover that both black-box and open-source models are more susceptible to placing *false* trust in lengthier external knowledge that is seemingly more compelling. This is proved by comparing the consistency of all multi-hop structures between introducing *Single Sentence* and *Paragraph* as external knowledge formats of distractors. The results presented in the fourth row of Table 3A demonstrate a consistent trend: the consistency decreases for both GPT3.5 and MPT-7B models ($p < 0.001$ in both cases, detailed in Appendix D.4) when the context of the distractor is longer and more detailed.

Why would the model's belief change, even when the core content of the distractor stays the same? We further investigate the interactions between the format and other distractor attributes. We find that *lengthier context raises belief in more severely edited external knowledge*, introduced through *Type Shift Distractors*. Specifically, we divide the external knowledge based on *Type Match* and *Type Shift*, and examine the resulting *changes* in confidence in both deviated and conforming responses caused by the alteration in the external knowledge format. In Table 7, we observe that MPT-7B's confidence rises mostly in its deviated responses to *Type Shift Distractors*, while its confidence decreases in all conforming responses. This implies that, in general, making the context lengthier lowers the model's confidence in extracting a target entity as the answer in response. However, it also boosts the model's confidence in trusting the *Type Shift Distractors*, which are more severely edited external knowledge.

Besides different degrees, we also discover that *the more detailed contexts tend to raise the model's belief in "weaker" distract methods*. Together with

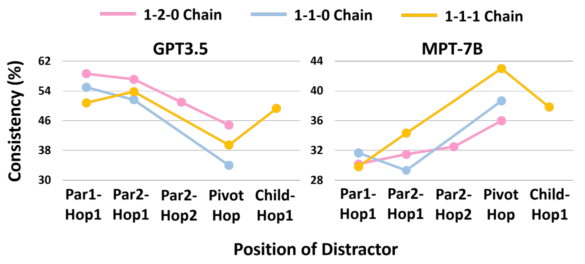


Figure 10: The consistency of GPT3.5 and MPT-7B when applying distractors to different positions in three multi-dependent structures. *Par* denotes the chains extended from the parent nodes of the pivot hop. The results show when the distractor is introduced to interfere with the pivot hop, GPT3.5's consistency is the lowest while MPT-7B's consistency is the highest.

the previous finding, it could be concluded that lengthier and more detailed contexts are effective in making the model trust the knowledge that it previously tended not to believe in. In Table 8, we show the resulting change of confidence when the knowledge format becomes *Paragraph* with respect to different distract methods. We discover that MPT-7B's confidence generally rises in its deviated responses when applying *Indirect Distractors* or *Subject Distractors*, which are both "weaker" distract methods.

5.3 Results on More PKG Structures

All previous analyses employ the multi-hop structures within the model's PKG, primarily focusing on one-to-one relations. In this section, we extend the distractor's influence upon multi-dependent structures as introduced in Section 4.3, and investigate the impacts brought by the sub-graphs of PKG with various knowledge structures.

Results for distract methods and degrees remain consistent. Our previous conclusions on distract methods and degrees are further supported by multi-dependent structures. Please refer to the consistent trend in Table 3B and Appendix D.5 for details.

Pivot hop exhibits unique traits. The results of introducing distractors on different positions are incomparable to previous conclusions, as the underlying knowledge structure changes. For multi-dependent structures, we discover that GPT3.5 achieves the lowest consistency if distracted on the *pivot* hop, while MPT-7B achieves the highest consistency.

To better observe the trend, we again differentiate three multi-dependent structures, and present our detailed results in Figure 10. Both 1-1-0 and 1-1-1 structures do not contain *Par2-Hop2* as the two parent nodes for *pivot* hop are only the ends for two 1-hop queries. Both 1-1-0 and 1-2-0 structures do not contain *Child-Hop1* as the answer nodes for their *pivot* hops are also the ending nodes for the whole chain. As the pivot hop depends on two upstream entities, the resulting trend in Figure 10 may imply that GPT3.5 is more easily deviated by distractors with auxiliary external information, while vice versa for MPT-7B.

6 Discussions

Below, we highlight three rather unexpected observations in our study and their implications.

Interference of Indirect Distraction While we can expect direct conflicts to cause the model's inconsistency, we are surprised that indirect distractions could also mislead the model (Figure 7). Their success could mainly be attributed to the model's lack of confidence in the original PKG: If the model's confidence about this particularly queried relation in PKG is low, but about *the relation in external knowledge is high*, then the model might deviate even if the external information is not related to what the question asks. We further identify two reasons why the model tends to believe in external knowledge. i) **Veracity**: *Indirect Distractor* could be a fact (as shown in Row A of Figure 3). Compared to other methods, *the model's "faith" in the correctness of external knowledge makes it doubt its original answers*. ii) **Matched Details**: The models are more easily distracted by similar details presented in both the query and the external information (e.g., in Figure 7, "1948" appears in both the distractor and the user's query for MPT-7B). These matching details lead the model to believe that there is a strong correlation between the distractor and the query, thus causing the model to select an entity from external knowledge as the answer.

The impact of indirect distraction flags additional challenges in misinformation, as most efforts tend to focus on preventing LLMs from trusting wrong statements. Future studies on effectively removing information snippets that pose unexpected effects will be valuable in the space.

Bias Towards Lengthier Context Both GPT and MPT models demonstrate low consistency when presented with lengthier and more detailed knowledge formats. This implies that the model's judgment is not solely rooted in facts and veracity, but rather resembles human decision-making, influenced by persuasiveness. Though lacking decisive evidence, we see that the model is more inclined to accept external knowledge it was previously unaware of or had doubts about, rather than blatantly false information (e.g., "The Sun rises in the west"). This tendency becomes more pronounced when such information is provided within a specific and detailed context with supporting details. This inherent bias in LLMs directly influences our approach to using them and raises concerns about potential misuse, whether external information is introduced explicitly or implicitly. Verification methods that compare model behaviors on long prompts vs. short but equally informative prompts (e.g., a high-quality summarization) might be a useful layer for rectifying LLM outputs.

GPT vs. MPT: GPT's Initial Distrust in External Knowledge We have discovered a different trend between GPT and MPT models with respect to the position to which distractors are introduced. GPT3.5 maintains a high consistency if the distraction is introduced at the beginning of the data chain, and we show in Appendix E.3 that GPT3 also exhibits a similar trend. This vigilance and distrust of external information in the beginning is not exhibited in MPT-7B, and may be attributed to the training process or some protection mechanisms behind the GPT black-box models. From our results, we also observe that GPT models are most likely to deviate in the 2^{nd} hop. This implies that though vigilant to the information's veracity initially, GPT's attention will decrease as the reasoning or interaction goes on, which heightens the risk of hallucination. Unfortunately, we have yet to reveal the root cause of such behavior differences, but we hope to look more into comparing different LLMs.

7 Conclusions

This paper mainly investigates the impacts of external knowledge on parametric knowledge through systematic experiments. We build a framework that reveals the model's *parametric knowledge graph* and automatically builds the *external knowledge distractors* as the source of interference. We conduct controlled experiments that investigate the impacts of external knowledge's distract degrees, methods, positions, and formats on various parametric knowledge structures including multi-hop and multi-dependent ones. Our results on both GPT3.5 and MPT show that both models tend to provide responses that deviate from their original PKG when the external information poses direct conflicts (*Object Distractors*), gives confounding changes that are not obviously false (*Type Match Distractors*), or provides external knowledge in detailed and lengthier context (*Paragraph Distractors*). In addition, we discover that GPT models are vigilant to external information's veracity in the beginning (*Distractors at 1^{st} Hop*), and that both models are susceptible to even unrelated external knowledge (*Indirect Distractors*). These studies reveal the mechanism of how LLMs handle potential conflicts, and imply the potential risk of hallucination as LLMs integrate external knowledge, even introduced implicitly. We hope our framework can serve as the testbed for more insightful investigations into the active interaction between external and parametric knowledge.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models.

Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, Alexander Wahler, Dieter Fensel, et al. 2020. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. In *International Conference on Learning Representations*.

Mosaic ML. 2023. Mpt-7b.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

Yasumasa Onoe, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

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.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. ♫ musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes.

Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 131–143.

Pengfei Yu and Heng Ji. 2023. Self information update for large language models through mitigating exposure bias.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Appendix

A Details on Revealing Model's PKG

We reveal the model's PKGs using natural language templates, as we show in Figure 15. During the construction of PKGs, we query the model three times with different temperatures ($T = 0.3, 0.5,$ and 0.7). The prompt for retrieving the answer is shown in Figure 16. Then, we judge the consistency of the model's responses through the checker we present in Figure 17. If we finally get "N/A", then we move on to search for other relations complying with the rules that the models are more confident about. Otherwise, we add the relation and the model's consistent answer into the PKG, regarding it as a piece of model's parametric knowledge.

B Details on Constructing Distractors

After the extraction of multiple structures as the original data chain from the model's PKGs, we perform modifications to the data chain to introduce distractors as external knowledge. This process is automated with the help of GPT3.5. Among the four types of distractors, distracting methods and degrees both directly modify the original information. Three distracting methods and two distracting degrees combine into a total of six types of distractors. The prompts applied for constructing these six types of distractors are introduced from Figure 18 to Figure 23.

Upon getting these six types of distractors, the external information we get is in a format of *Single Sentence*. To turn them into external knowledge presented in multiple sentences, we apply the prompt in Figure 24 to construct distractors in *Paragraph* format. Distractors introduced to interfere with different positions do not need additional construction.

C Details on Experimental Settings

We apply both GPT3.5 and MPT-7B models. For all the experiments, we set *top-p* to 1 and *temperature* to 0.3. The same setting across all experiments guarantees fairness when we are measuring the model's consistency and ensures that the model's confidence is comparable. We set the max sequence length to 512, and for both models, we do not add the frequency or presence penalty.

As introduced in Appendix B, combining distract methods and degrees, we get six different types of distractors for each hop of the query (each

Metrics		GPT3.5 Match	GPT3.5 Shift
Consistency (%)	2-hop	59.92	$61.00_{\uparrow 1.1}$
	3-hop	56.78	$62.33_{\uparrow 5.6}$
	4-hop	51.00	$52.17_{\uparrow 1.2}$
Metrics		**MPT-7B** Match	**MPT-7B** Shift
Consistency (%)	2-hop	48.25	$53.92_{\uparrow 5.7}$
	3-hop	42.22	$46.67_{\uparrow 4.5}$
	4-hop	37.88	$38.75_{\uparrow 0.9}$
Confidence (%)	2-hop	82.07	$78.86_{\downarrow 3.2}$
	3-hop	80.99	$77.78_{\downarrow 3.2}$
	4-hop	79.71	$78.40_{\downarrow 1.3}$

Table 9: The detailed results for GPT3.5 and MPT-7B when confronting *Type Match* or *Type Shift Distractors* as external interfering knowledge. We differentiate multiple structures instead of performing macro-averaging.

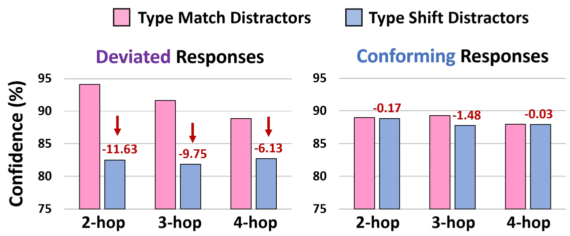


Figure 11: The decrease of confidence brought by changing from *Type Match* to *Type Shift Distractors*. We observe a significant confidence drop for deviated responses when introducing *Type Shift Distractors*, while conforming responses show minor changes.

data chain may have multiple hops of the query). We experiment with all these distractors. For the results on distract degrees, we divide the results based on the two different degrees of the distractors. For the results on distract methods, we divide our results based on the three different methods applied in the distractors. For the results on distract positions, we divide the results based on which hop of query in the knowledge structure the distractor is introduced to. For the results on distract knowledge formats, we introduce a *Paragraph* version to all previous distractors and repeat all the experiments for comparison.

D Supporting Analysis to Main Results

For some of the main results, we also perform additional analysis to further support our claims.

D.1 Distract Degrees

The results of the P-value we provide are derived from the T-test between all the consistency values

Metrics		GPT3.5 Match	GPT3.5 Shift
Abstention (%)	2-hop	37.42	$73.72_{\uparrow 36.3}$
	3-hop	37.53	$69.76_{\uparrow 32.2}$
	4-hop	44.05	$61.41_{\uparrow 17.4}$
Variation (%)	2-hop	62.58	$26.28_{\downarrow 36.3}$
	3-hop	62.47	$30.24_{\downarrow 32.2}$
	4-hop	55.95	$38.59_{\downarrow 17.4}$
Metrics		**MPT-7B** Match	**MPT-7B** Shift
Abstention (%)	2-hop	4.35	$9.22_{\uparrow 4.9}$
	3-hop	3.85	$6.35_{\uparrow 2.5}$
	4-hop	8.18	$9.93_{\uparrow 1.8}$
Variation (%)	2-hop	95.65	$90.78_{\downarrow 4.9}$
	3-hop	96.15	$93.65_{\downarrow 2.5}$
	4-hop	91.82	$90.07_{\downarrow 1.8}$

Table 10: The detailed error analysis on inconsistent chains for GPT3.5 and MPT-7B when confronting *Type Match* or *Type Shift Distractors* as external interfering knowledge. We differentiate multiple structures instead of performing macro-averaging.

under the interference of *Type Shift Distractors* and *Type Match Distractors*. Each data chain would provide a pair of values for comparison, and there are in total 600 data chains for all 2 / 3 / 4-hop structures.

The results we provide in Table 3 and Table 4 are macro-average on three structures. We provide more detailed results regarding different multi-hop structures in Table 9 (Main Metrics) and Table 10 (Error Analysis). The consistent trend in every structure provides additional support to our conclusions.

Besides, to further substantiate our claim that the model resists *Type Shift Distractors*, we segment the confidence based on whether the model's response is conforming or deviated. Figure 11 displays that the average confidence of generating a deviated response plummets (left chart) when encountering *Type Shift Distractors*, while the confidence of conforming responses shows minor changes. This implies the primary cause of the confidence drop stems from the deviated responses: the model is already hard to be deviated by *Type Shift Distractors*, and for responses that *are* distracted by them, the model's belief in them still remains low.

D.2 Distract Methods

We conduct the T-test for all the resulting consistencies between *Object-Indirect Distractors* and *Object-Subject Distractors* for both GPT3.5 and MPT-7B. Similarly, each test comprises 600 pairs

Metrics		GPT3.5 Object	GPT3.5 Indirect	GPT3.5 Subject
Consistency (%)	2-hop	$40.50_{\downarrow 24.8/35.1}$	65.25	75.62
	3-hop	$45.75_{\downarrow 17.8/23.7}$	63.50	69.42
	4-hop	$42.44_{\downarrow 12.1/15.4}$	54.50	57.81
Metrics		**MPT-7B** Object	**MPT-7B** Indirect	**MPT-7B** Subject
Consistency (%)	2-hop	$38.38_{\downarrow 11.5/26.6}$	49.88	65.00
	3-hop	$35.25_{\downarrow 8.8/18.8}$	44.08	54.00
	4-hop	$33.50_{\downarrow 4.4/10.0}$	37.94	43.50
Confidence (%)	2-hop	81.87	76.67	82.92
	3-hop	80.15	77.10	80.95
	4-hop	79.57	77.82	79.80

Table 11: The detailed results for GPT3.5 and MPT-7B when confronting *Object*, *Indirect* or *Subject Distractors* as external interfering knowledge. We differentiate multiple structures instead of performing macro-averaging.

Metrics		GPT3.5 Object	GPT3.5 Indirect	GPT3.5 Subject
Abstention (%)	2-hop	$47.06_{\downarrow 16.3/17.0}$	63.31	64.10
	3-hop	$48.54_{\downarrow 7.2/7.3}$	55.71	55.86
	4-hop	$50.81_{\downarrow 0.4/5.8}$	51.24	56.59
Variation (%)	2-hop	$52.94_{\uparrow 16.3/17.0}$	36.69	35.90
	3-hop	$51.46_{\uparrow 7.2/7.3}$	44.29	44.14
	4-hop	$49.19_{\uparrow 0.4/5.8}$	48.76	43.41
Metrics		**MPT-7B** Object	**MPT-7B** Indirect	**MPT-7B** Subject
Abstention (%)	2-hop	$8.92_{\uparrow 4.2/3.6}$	4.74	5.36
	3-hop	$7.08_{\uparrow 3.5/3.1}$	3.58	3.99
	4-hop	$10.24_{\uparrow 2.7/1.0}$	7.55	9.29
Variation (%)	2-hop	$91.08_{\downarrow 4.2/3.6}$	95.26	94.64
	3-hop	$92.92_{\downarrow 3.5/3.1}$	96.42	96.01
	4-hop	$89.76_{\downarrow 3.5/3.1}$	92.45	90.71

Table 12: The detailed error analysis on inconsistent chains for GPT3.5 and MPT-7B when confronting *Type Match* or *Type Shift Distractors* as external interfering knowledge. We differentiate multiple structures instead of performing macro-averaging.

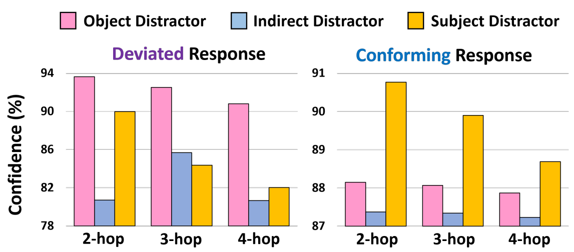


Figure 12: The confidence of MPT-7B's conforming and deviated responses when applying three distract methods. Results show *Object Distractor* induces the highest confidence for deviated responses, while *Subject Distractor* induces the highest confidence for conforming responses.

Metrics		GPT3.5 1st Hop	GPT3.5 2nd Hop	GPT3.5 3rd Hop	GPT3.5 4th Hop
Consistency (%)	2-hop	65.50	57.42	–	–
	3-hop	64.00	56.08	58.58	–
	4-hop	55.83	47.08	49.42	54.00
Metrics		**MPT-7B** 1st Hop	2nd Hop	3rd Hop	4th Hop
Consistency (%)	2-hop	48.58	53.58	–	–
	3-hop	40.42	46.00	46.92	–
	4-hop	37.33	36.17	39.58	40.17
Confidence (%)	2-hop	78.48	82.46	–	–
	3-hop	78.47	80.47	79.24	–
	4-hop	79.10	79.38	79.09	78.67

Table 13: The detailed results for GPT3.5 and MPT-7B when distractors are introduced in different positions as external interfering knowledge. We differentiate multiple structures instead of performing macro-averaging.

of values for comparison.

We provide detailed results regarding the impacts of three distract methods on different multi-hop structures in Table 11 (Main Metrics) and Table 12 (Error Analysis).

In addition, we analyze MPT-7B's confidence with respect to three distract methods in Figure 12. Again, we divide the confidence based on whether the response is conforming or deviated. Our findings show that the *Object Distractor* results in the highest confidence when the response deviates, while the confidence for *Subject Distractor* is highest when the response conforms to the PKG. These results also imply *Object Distractor* is the most powerful to deviate the model's belief, while for the other two "weaker" distractors, the model still trusts its original logic pathway in PKG.

D.3 Distract Positions

We provide detailed results on GPT3.5 and MPT-7B's consistency and confidence with respect to different positions where the distractor is introduced in Table 13. We have plotted the trend of consistency in our main results in Figure 8.

D.4 Distract Formats

We conduct the T-test for all the resulting consistencies between *Single Sentence* and *Paragraph* as knowledge format for both GPT3.5 and MPT-7B. Similarly, each test comprises 600 pairs of values for comparison.

We provide detailed results on GPT3.5 and MPT-7B's consistency and confidence with respect to the *Single Sentence* or *Paragraph* as the format of external knowledge in Table 14. We show from the detailed results that every structure's trend is consistent with our main conclusion.

Metrics		GPT3.5 Single Sentence	GPT3.5 Paragraph
Consistency (%)	2-hop	60.46	$57.83_{\downarrow 2.6}$
	3-hop	59.56	$56.53_{\downarrow 3.0}$
	4-hop	51.59	$48.75_{\downarrow 2.9}$
Metrics		**MPT-7B** Single Sentence	Paragraph
Consistency (%)	2-hop	51.09	$48.38_{\downarrow 2.7}$
	3-hop	44.11	$42.34_{\downarrow 1.8}$
	4-hop	38.32	$35.46_{\downarrow 2.9}$
Confidence (%)	2-hop	80.48	$78.28_{\downarrow 2.2}$
	3-hop	79.40	$78.45_{\downarrow 1.0}$
	4-hop	79.06	$78.36_{\downarrow 0.7}$

Table 14: The detailed results for GPT3.5 and MPT-7B when confronting *Single Sentence* or *Paragraph* as the format of external interfering knowledge. We differentiate multiple structures instead of performing macro-averaging.

Metrics		GPT3.5 Match	GPT3.5 Shift
Consistency (%)	1-1-0	46.33	$47.33_{\uparrow 1.1}$
	1-1-1	47.75	$49.00_{\uparrow 1.3}$
	1-2-0	51.58	$54.25_{\uparrow 2.7}$
Metrics		**MPT-7B** Match	Shift
Consistency (%)	1-1-0	32.00	$34.44_{\uparrow 2.4}$
	1-1-1	42.22	$46.67_{\uparrow 4.5}$
	1-2-0	32.00	$33.08_{\uparrow 1.1}$
Confidence (%)	1-1-0	76.88	$74.96_{\downarrow 1.9}$
	1-1-1	79.35	$77.82_{\downarrow 1.5}$
	1-2-0	79.57	$77.65_{\downarrow 1.9}$

Table 15: The detailed results for GPT3.5 and MPT-7B when multi-dependent structures (1-1-0, 1-1-1, and 1-2-0) confronts *Type Match* or *Type Shift Distractors* as external interfering knowledge. We differentiate the three structures instead of performing macro-averaging.

D.5 Multi-Dependent Structures

To establish the overarching applicability of our prior conclusions, we undertake a parallel analysis with distractors of different methods and degrees to multi-dependent structures in PKG. The experimental settings and methods are kept the same as those for multi-hop structures. As delineated in Table 15, for all three multi-dependent structures, our findings reveal that the model's consistency is higher in response to *Type Shift Distractors*, though the model's overall confidence lowers. Furthermore, Table 16 showcases that the *Object Distractors*

Metrics		GPT3.5		
		Object	Indirect	Subject
Consistency (%)	1-1-0	$32.83_{\downarrow 11.7/30.5}$	44.50	63.33
	1-1-1	$32.25_{\downarrow 17.1/28.3}$	49.38	60.50
	1-2-0	$41.12_{\downarrow 13.8/21.6}$	54.87	62.75
Metrics		**MPT-7B**		
		Object	Indirect	Subject
Consistency (%)	1-1-0	$20.67_{\downarrow 14.3/25.3}$	34.00	45.00
	1-1-1	$29.38_{\downarrow 5.2/15.4}$	34.62	44.75
	1-2-0	$24.12_{\downarrow 7.1/18.1}$	31.25	42.25
Confidence (%)	1-1-0	75.27	74.50	77.98
	1-1-1	78.81	77.63	79.33
	1-2-0	79.01	77.41	79.42

Table 16: The detailed results for GPT3.5 and MPT-7B when multi-dependent structures (1-1-0, 1-1-1, and 1-2-0) confronts *Object*, *Indirect* or *Subject Distractors* as external interfering knowledge. We differentiate the three structures instead of performing macro-averaging.

Metrics		GPT3.5				
		Par1 Hop1	Par2 Hop1	Par2 Hop2	Pivot Hop	Child Hop1
Consistency (%)	2-hop	55.00	51.67	–	34.00	–
	3-hop	50.83	53.83	39.50	–	49.33
	4-hop	58.67	57.17	51.00	44.83	–
Metrics		**MPT-7B**				
		Par1 Hop1	Par2 Hop1	Par2 Hop2	Pivot Hop	Child Hop1
Consistency (%)	2-hop	31.67	29.33	–	38.67	–
	3-hop	29.83	34.33	–	43.00	37.83
	4-hop	30.17	31.50	32.50	36.00	–
Confidence (%)	2-hop	74.48	73.72	–	79.52	–
	3-hop	78.11	76.24	–	81.32	78.67
	4-hop	78.61	77.56	78.52	79.76	–

Table 17: The detailed results for GPT3.5 and MPT-7B when distractors are introduced in different positions of multi-dependent structures (1-1-0, 1-1-1, and 1-2-0) as external interfering knowledge. *Par.* denotes the chains extended from the parent nodes of the *pivot* query. We differentiate the three structures instead of performing macro-averaging.

remain the prime catalyst for the model's deviation. Notably, these insights are consistent with the outcomes obtained from our investigations into multi-hop structures.

Furthermore, we provide detailed results on GPT3.5 and MPT-7B's consistency and confidence with respect to different positions where the distractor is introduced in Table 17. We have plotted the trend of consistency in our main results in Figure 10.

Metrics		GPT3	
		Match	Shift
Consistency (%)	2-hop	52.67	$68.00_{\uparrow 15.3}$
	3-hop	43.56	$52.78_{\uparrow 9.2}$
	4-hop	46.58	$52.92_{\uparrow 6.3}$
Confidence (%)	2-hop	69.87	$66.98_{\downarrow 2.9}$
	3-hop	68.41	$66.05_{\downarrow 2.4}$
	4-hop	70.04	$68.10_{\downarrow 1.9}$

Table 18: The consistency and confidence of GPT3 when confronting *Type Match* or *Type Shift Distractors* as external interfering knowledge. The conclusion on distract degrees is consistent and even more pronounced for GPT3.

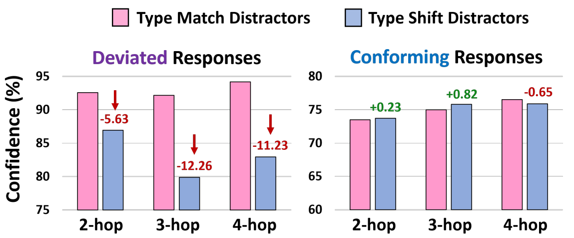


Figure 13: The decrease of confidence in GPT3 brought by changing from *Type Match* to *Type Shift Distractors*. The confidence drop can be mainly attributed to the deviated response, which implies GPT3 also shows resistance to *Type Shift Distractors*.

E Additional Results from GPT3

To further support our discoveries, we perform additional experiments on GPT3 (Text-Davinci-003). Though GPT3 is not designed as a conversational model, its results can still reflect and bolster some of the trends that we have discovered. We perform experiments on 2 / 3 / 4-hop data chains, with 100 raw chains from GPT3's PKG for each type. We keep the rules we applied for constructing the PKG the same, and we keep all the other experimental setups the same as we introduced in Appendix C.

E.1 Distract Degrees

In Table 18, we observe the same trend in GPT3 that the model resists *Type Shift Distractors* as external knowledge, and the overall confidence in responses is lowered. By further dividing the responses into conforming and deviated ones, we show in Figure 13 that, similarly, the drop in confidence can be mainly attributed to the deviated responses. All these results further bolster the claim that the model put less faith in more severely edited knowledge represented by *Type Shift Distractors*.

Metrics		GPT3		
		`Object`	`Indirect`	`Subject`
Consistency (%)	2-hop	$40.75_{\downarrow 24.5/34.3}$	65.25	75.00
	3-hop	$34.00_{\downarrow 18.5/24.0}$	52.50	58.00
	4-hop	$39.00_{\downarrow 14.1/18.1}$	53.12	57.12
Confidence (%)	2-hop	71.72	61.34	72.33
	3-hop	69.79	63.30	68.62
	4-hop	71.29	65.27	70.74

Table 19: The consistency of GPT3 when confronting distractors that apply different distract methods. Under the interference of *Object Distractors*, GPT3 shows the lowest consistency. This result still remains consistent with previous conclusions.

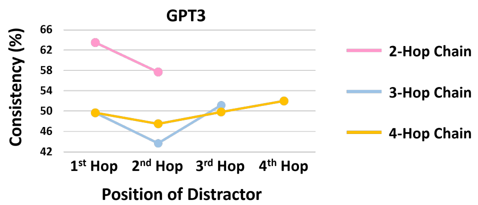


Figure 14: The consistency of GPT3 when the distractor is introduced to interfere with different positions in the data chain. GPT3 exhibits a similar trend as GPT3.5 in Figure 8.

E.2 Distract Methods

In addition to distractors of different degrees, we also investigate GPT3's consistency towards distractors that apply different methods. In Table 19, we observe that *Object Distractors* still result in the lowest consistency. This trend also remains the same as what we have shown previously, indicating that GPT3 is also susceptible to *Object Distractors* the most, while the other two "weaker" distractors also bring certain impacts.

E.3 Distract Position

The pattern for distract positions is different for GPT3.5 and MPT-7B, as we have shown earlier in Figure 8. In Figure 14, we demonstrate that GPT3's trend is more similar to GPT3.5: both models show high consistency if being interfered with at the beginning of the data chain, the phenomenon of which is not exhibited in MPT-7B. Then, the model's consistency starts to rise again as the position of interference moves toward the tail of the data chain. GPT3's results further support the *GPT family*'s initial sensitivity towards information's veracity.

Metrics		GPT3	
		`Single Sentence`	`Paragraph`
Consistency (%)	2-hop	60.34	$47.30_{\downarrow 13.0}$
	3-hop	48.17	$40.00_{\downarrow 8.2}$
	4-hop	49.75	$44.00_{\downarrow 5.6}$

Table 20: The comparison of GPT3's consistency when presented with distractor of *Single Sentence* format versus *Paragraph* format. GPT3's consistency also lowers when the context becomes lengthier.

```
"country": [
    ["the capital is", "city"],
    ["is founded / become independent in which year", "year"],
    ["the official / most commonly spoken language is", "language"],
    ["the national anthem is", "song"],
    ["the colors on the national flag are (may be multiple)", "color"],
    ["has the longest river named", "river"],
    ["has the highest mountain named", "mountain"],
    ["the countries in the east that shares boarder with it (may be multiple)",
"country"]
  ],
"city": [
    ["belongs to which country", "country"],
    ["has the largest airport named", "airport"],
    ["has the largest (by area) university / college named", "school"],
    ["the time zone of this city in UTC", "time zone"],
    ["what is the largest company (by people) based in this city",
"company"]
  ],
"year": [
    ["the US president this year named", "person"],
    ["this year's Oscar best actors are (may have multiple)", "person"],
    ["this year's Oscar Outstanding Pictures is (the first one in alphabetical
order)", "film"],
    ["this year's NBA championship is which team", "sport team"],
    ["the first Olympic Games hosting city from this year", "city"]
  ],
...
```

Figure 15: The example rules we apply in building the PKG.

E.4 Distract Formats

We extend the context format of all the distractors into *Paragraph* in the same way we do for GPT3.5 and MPT-7B. We present the results of the comparison in Table 20. GPT3's consistency lowers in a more pronounced way than GPT3.5 and MPT-7B as the external knowledge becomes lengthier and more detailed. This further supports our previous conclusions and also implies that GPT3 is more biased to trust the detailed but potentially false external knowledge.

System Message
You are supposed to answer the question given by the user in a succinct way. Please do not provide any additional information.
1. If you do not know the answer for sure, please generate 'Not Sure'.
2. If you think there are multiple answers, please split them by semicolon (;)
Instruction
Answer the question briefly, and please always provide an answer.
User
What's the capital of USA?
Assistant
Washington DC
User
Jackson Chen is born in which city?
Assistant
Not Sure
User
What are the colors on the national flag of China?
Assistant
Red; Yellow
User
What is the longitude of Washington DC round to integer?
Assistant
77W
User
Who is the headmaster of Yale University in 2000?
Assistant
Richard C. Levin
User
{Target Question}
Assistant
{Target Model Response}

Figure 16: The prompt for retrieving the model's answer when building the PKG.

System Message
You are supposed to judge if the given concepts are consistent (consistent doesn't mean the same, alias is allowed).
1. If they are consistent, then please generate the common concept they share, otherwise, please generate 'N/A'.
2. If there are multiple concepts in one list, you can list all the concepts shared the same meaning in all the lists, and discard other inconsistent ones. Please separate answers in semicolon (;)
Instruction
Give the core concept if the there exist one that is shared / consistent in all the lists, otherwise, please generate 'N/A'.
User
['The state is Georgia'], ['Georgia'], ['State of Georgia']
Assistant
Georgia
User
['1945'], ['year of 1948']
Assistant
N/A
User
['Google', 'Apple'], ['Google', 'Apple Inc.'], ['Apple', 'Google', 'Microsoft']
Assistant
Google; Apple
User
['Not Sure'], ['The answer should be French.'], ['I am not sure about the answer, please provide more information.']
Assistant
N/A
User
['Red', 'Blue'], ['blue', 'yellow']
Assistant
Blue
User
['LA'], ['Los Angeles'], ['The city that Jackson lives in is L.A.']
Assistant
Los Angeles
User
{Target Lists}
Assistant
{Target Model Response}

Figure 17: The prompt for judging the consistency of the model's answers when building the PKG.

Object Distractor – Type Match
System Message
You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a definitive statement (instead of the QA format). Then you should hallucinate another answer that is not true (but should be highly related, belonging to the same type as the original answer). Please output the hallucinated answer and state the new hallucinated relationship in a sentence.
User
Question: What is the capital of Afghanistan?
Answer: Kabul
Assistant
Statement: The capital of Afghanistan is Kabul.
Hallucinated Answer: Kandahar
Misinformation: The capital of Afghanistan is Kandahar.
User
Question: France is on which continent?
Answer: Europe
Assistant
Statement: France is in Europe.
Hallucinated Answer: Asia
Misinformation: France is in Asia.
User
{Target Question / Answer}
Assistant
{Target Model Response}

Figure 18: The prompt for constructing *Object - Type Match Distractors*.

Object Distractor – Type Shift
System Message
You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a definitive statement (instead of the QA format). Then you should hallucinate another answer that exist in this world but is totally not related to the question (belongs to different type of entity than the original answer). Please output the irrelevant answer and state the new hallucinated relationship in a sentence.
User
Question: What is the capital of Afghanistan?
Answer: Kabul
Assistant
Statement: The capital of Afghanistan is Kabul.
Irrelevant Answer: Michael Jackson
Misinformation: The capital of Afghanistan is Michael Jackson.
User
Question: France is on which continent?
Answer: Europe
Assistant
Statement: France is on continent Europe.
Irrelevant Answer: Apple Inc.
Misinformation: France is on continent Apple Inc.
User
{Target Question / Answer}
Assistant
{Target Model Response}

Figure 19: The prompt for constructing *Object - Type Shift Distractors*.

Subject Distractor – Type Match
System Message
You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a triplet (Subject, Relationship, Answer). Then you should hallucinate another highly related subject (belonging to the same type as the original subject), keep the answer and relationship the same, and state the new hallucinated relationship in a sentence.
User
Question: What is the capital of Afghanistan?
Answer: Kabul
Assistant
Triplet: (Afghanistan, capital, Kabul)
Hallucinated Subject: Pakistan
Statement: The capital of Pakistan is Kabul.
User
Question: France is on which continent?
Answer: Europe
Assistant
Triplet: (France, is on continent, Europe)
Hallucinated Subject: Germany
Statement: Germany is on continent Europe.
User
{Target Question / Answer}
Assistant
{Target Model Response}

Figure 20: The prompt for constructing *Subject - Type Match Distractors*.

Subject Distractor – Type Shift
System Message
You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a triplet (Subject, Relationship, Answer). Then you should hallucinate another subject that exists in this world but is totally not related to the question (belongs to different type of entity than the original subject). Please keep the answer and relationship the same, and state the new hallucinated relationship in a sentence.
User
Question: What is the capital of Afghanistan?
Answer: Kabul
Assistant
Triplet: (Afghanistan, capital, Kabul)
Hallucinated Subject: Michael Jackson
Statement: The capital of Michael Jackson is Kabul.
User
Question: France is on which continent?
Answer: Europe
Assistant
Triplet: (France, is on continent, Europe)
Hallucinated Subject: 'To kill a mocking bird'
Statement: 'To kill a mocking bird' is on continent Europe.
User
{Target Question / Answer}
Assistant
{Target Model Response}

Figure 21: The prompt for constructing *Subject - Type Shift Distractors*.

Indirect Distractor – Type Match
System Message
You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a triplet (Subject, Relationship, Answer). Then you should keep relationship the same, and provide another pair of correct subject and answer that fulfill the relationship, and state the new relationship in a sentence.
User
Question: What is the capital of Afghanistan?
Answer: Kabul
Assistant
Triplet: (Afghanistan, capital, Kabul)
New Subject: Pakistan
New Answer: Islamabad
Statement: The capital of Pakistan is Islamabad.
User
Question: France is on which continent?
Answer: Europe
Assistant
Triplet: (France, is on continent, Europe)
New Subject: India
New Answer: Asia
Statement: India is on continent Asia.
User
{Target Question / Answer}
Assistant
{Target Model Response}

Figure 22: The prompt for constructing *Indirect - Type Match Distractors*.

Indirect Distractor – Type Shift
System Message
You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a triplet (Subject, Relationship, Answer). Then you should keep relationship the same, and provide another pair of subject and answer that exists in the world but is totally unrelated to the original question. Finally, state the new hallucinated relationship in a sentence.
User
Question: What is the capital of Afghanistan?
Answer: Kabul
Assistant
Triplet: (Afghanistan, capital, Kabul)
Hallucinated Subject: Michael Jackson
Hallucinated Answer: Eiffel Tower
Statement: The capital of Michael Jackson is Eiffel Tower.
User
Question: France is on which continent?
Answer: Europe
Assistant
Triplet: (France, is on continent, Europe)
Hallucinated Subject: Apple Inc.
Hallucinated Answer: Monopoly Game
Statement: Apple Inc. is on continent Monopoly Game.
User
{Target Question / Answer}
Assistant
{Target Model Response}

Figure 23: The prompt for constructing *Indirect - Type Shift Distractors*.

System Message
You are a helpful assistant in writing articles in a parallel world. Please fake up some news in this world around the given fact I provide.
Instruction
You are asked to generate a fake piece of news around the statement I give to you. Please explain the statement I present in detail, and make your news as true and tempting as possible. Please write the news as it happens naturally. Wrap up your news within 3-4 sentences.
User
Statement: The name of the longest river in France is Yellow Elephant
Assistant
It has been revealed that the longest river in France is, in fact, called the Yellow Elephant. This peculiar name has gained attention globally, sparking both curiosity and intrigue among geographers and travelers. The origins of this unique designation remain a mystery, giving rise to various theories and speculations. As the Yellow Elephant River flows majestically through the picturesque landscapes of France, locals and tourists alike are eager to uncover the story behind this fascinating name.
User
Statement: The most recent Olympic Games is held where giraffe can fly.
Assistant
The most recent Olympic Games were hosted in a location where giraffes can fly. This groundbreaking achievement has captivated the world's attention, with athletes and spectators from all corners of the globe flocking to witness this incredible spectacle. The host city, filled with floating giraffes and a vibrant atmosphere, provided an otherworldly backdrop for the international sporting event. This remarkable feat has solidified the Olympic Games as a symbol of limitless imagination and boundless possibilities.
User
Statement: The capital of China is Washington DC.
Assistant
Recently, the Chinese government has announced that the capital of China will now be relocated to Washington DC. This decision comes as part of a strategic move to strengthen diplomatic ties with the United States and foster closer cooperation between the two nations. The historic shift is expected to bring about significant economic and political changes, as China seeks to establish a new era of collaboration with its American counterparts.
User
{Target Statement}
Assistant
{Target Model Response}

Figure 24: The prompt for getting the distractors in *Paragraph* format.