

How Well Do Large Language Models Truly Ground?

Hyunji Lee^{1*} Sejune Joo^{1*} Chaeeun Kim^{1†} Joel Jang²
 Doyoung Kim¹ Kyoung-Woon On³ Minjoon Seo¹

¹KAIST AI ²University of Washington ³Kakao Brain
 {hyunji.amy.lee}@kaist.ac.kr

Abstract

Reliance on the inherent knowledge of Large Language Models (LLMs) can cause issues such as hallucinations, lack of control, and difficulties in integrating variable knowledge. To mitigate this, LLMs can be probed to generate responses by grounding on external context, often given as input (knowledge-augmented models). Yet, previous research is often confined to a narrow view of the term “grounding”, often only focusing on whether the response contains the correct answer or not, which does not ensure the reliability of the entire response. To address this limitation, we introduce a strict definition of grounding: a model is considered *truly* grounded when its responses (1) fully utilize necessary knowledge from the provided context, and (2) don’t exceed the knowledge within the contexts. We introduce a new dataset and a grounding metric to assess this new definition and perform experiments across 13 LLMs of different sizes and training methods to provide insights into the factors that influence grounding performance. Our findings contribute to a better understanding of how to improve grounding capabilities and suggest an area of improvement toward more reliable and controllable LLM applications. We publicly release our Code and Data in <https://github.com/kaistAI/How-Well-Do-LLMs-Truly-Ground>.

1 Introduction

Large Language Models (LLMs) have shown superior performance on various tasks by leveraging the extensive world knowledge embedded in their parameters. However, these models often produce hallucinations (Bender et al., 2021; Du et al., 2023), lack controllability (Dathathri et al., 2019; Zhang et al., 2022), and have trouble integrating knowledge that changes over time (Lin et al., 2021; Wang et al., 2021). Additionally, they may not

contain specialized knowledge unique to certain entities, such as company-specific terminology, or private information not contained in the training data. Although it is technically possible to inject new knowledge by further training LLMs on a specific corpus, this approach is generally inefficient and not practical in many scenarios (Mallen et al., 2022; Panda et al., 2023; Tang et al., 2023). To address these issues, various systems¹ and work (Gao et al., 2023; He et al., 2022; Xu et al., 2023; Yao et al., 2022) have explored methods where such dynamic, specialized, or private contexts provided by users or general world knowledge contexts retrieved from a large corpus (retrieval-augmented models) are provided to LLMs as additional inputs.

While previous work has shown enhanced performance by allowing LLMs to ground their outputs on external contexts compared to solely relying on the LLM’s inherent knowledge (Andrew and Gao, 2007; BehnamGhader et al., 2022; Mallen et al., 2022), whether the model *well-grounds* to the contexts is usually measured by simply checking whether the generated response contains the answer (Liu et al., 2023a; Mallen et al., 2022; Lewis et al., 2020) or evaluating over NLI model to see whether the knowledge from given context correlates with generated response (Gao et al., 2023; Asai et al., 2023). However, in some cases, this may not be sufficient and it may be more important to ensure that the *entire* generated response is *truly* grounded on the given external contexts.

For example, let’s consider the scenario in Figure 1, where a company’s HR team is utilizing an LLM to question the qualifications of candidates by providing their resumes as external contexts and prompting the LLM to provide an answer to questions about the candidates based on their resumes. Response 1 omits essential information

*Denotes equal contribution

†Work done during internship at KAIST AI

¹<https://www.bing.com/new>, <https://www.perplexity.ai/>, <https://openai.com/blog/chatgpt-plugins>

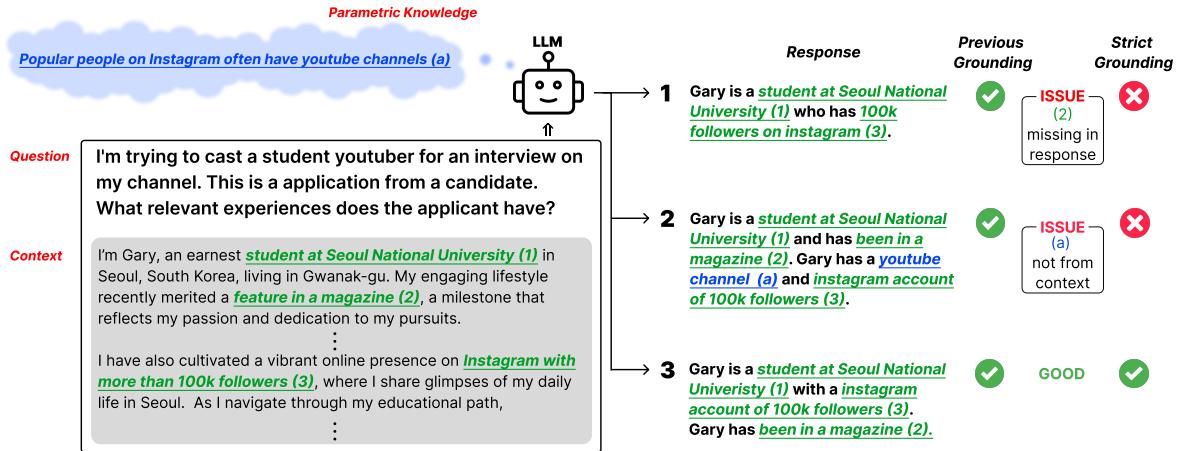


Figure 1: An example scenario of a company’s HR team using LLM to question upon candidate’s resume which is given as input context. The previous definition of grounding would consider responses 1 and 2 as well grounded due to their high relevancy with the question and input context. However, as our definition considers all knowledge in a fine-grained manner, we consider *only* response 3 as well-grounded. Response 1 misses key resume detail (2) which makes the candidate underrated. Response 2 introduces knowledge (a) that is not from the given context but from the model’s parametric knowledge, inaccurately overrates the candidate, and unfairly influences comparison with others.

about the candidate and Response 2 contains misinformation about the candidate due to generating knowledge contained in its parameters; both cases do not truly represent the candidate’s qualifications. It either harms the applicant by missing important information or makes the applicant overly qualified, disadvantaging other applicants.

In this study, we introduce a strict definition of grounding: a model is *truly* grounding on given contexts when it (1) uses all essential knowledge from the contexts and (2) strictly adheres to their scope in response generation without hallucinated information². To quantify this definition, we introduce an automatic grounding metric that extends upon Min et al. (2023) for fine-grained evaluation. Furthermore, we curate a new dataset incorporating crucial factors influencing LLMs’ response (i.e., entity popularity, context length), to understand their impact on LLM responses. Lastly, we present a revised version of the dataset that modifies factual knowledge in external contexts to identify the knowledge sources in responses.

We conduct experiments across 13 LLMs of different sizes and training methods to explore which model attributes significantly contribute to grounding ability and identify some important factors.

- Training methods (Instruction Tuning & RLHF) have a more pronounced impact on grounding performance than the size of the

²In this paper, the term grounding refers to what is defined here as truly grounding.

model.

- Increasing the size of the model often leads to a more substantial decrease in grounding performance in our revised setup.
- High answer accuracy, commonly used to assess how well a model incorporates context in previous works, does not ensure high grounding performance.
- Open-source models demonstrate grounding performance that is on par with that of proprietary models.
- Instruction-tuned models show high degradation when additional relevant contexts are added as input.

2 Related Works

Question Answering Machine Reading Comprehension and Open Domain Question Answering provide a question and context to a model, which then answers the question using the given context. The answers are usually short phrases or entities. LongformQA shares similarities, as it also uses contextual information to answer questions, but its answers are longer and focus on how well the model refers to the input context and generates factual responses.

Such datasets, while encompassing questions and contexts, lack the annotation of vital knowledge from the provided context and dataset to verify from which source the knowledge came (whether the knowledge is from a given context or from its

parameter) which is essential for evaluating grounding performance. This is because prior tasks focus on the correctness of the answer, sidelining the evaluation of the mechanism through which the answer was derived and the extent of context utilization. Thereby, our dataset not only contains questions, contexts, and answers as previous datasets but also annotation pinpointing the essential knowledge from the context and revised version to check which source the knowledge came from. It’s noteworthy that we didn’t simply adapt existing datasets, given their inherent limitations; most do not consider key variables known to influence LLM performance (Section 3.2) as they were constructed prior to the advent of modern LLMs. Thereby to analyze various aspects of models, we construct the dataset from scratch.

Generating Response with External Knowledge
Recent research efforts have focused on incorporating external knowledge during the generation process to overcome issues such as hallucination, increase controllability, and incorporate dynamic knowledge. It incorporates either by inputting it directly (Lewis et al., 2020; Liu et al., 2023b; Shi et al., 2023), using APIs in a multi-step manner (Yao et al., 2022; Xu et al., 2023), or by employing various tools (Schick et al., 2023; Yang et al., 2023).

Although the objective of adding external knowledge is for the model’s response to be intrinsically tied to the given knowledge, previous work naively evaluates and analyzes the ability. With such a naive definition, users find it difficult to ensure that the entire generated response is truly grounded in the given context; the model may hallucinate or miss important knowledge even though the overall response corresponds well to the external context. Thereby, in this work, we introduce a strict definition of grounding and share the importance of checking the entire response in a fine-grained manner.

Definition of Grounding The concept of “grounding” pervades several areas that interface with natural language. In robotics, grounding bridges the chasm between abstract directives and actionable robot commands, as highlighted by numerous studies (Ahn et al., 2022; Huang et al., 2023; Kollar et al., 2010b,a; Tellex et al., 2011; Mees et al., 2022). In the domain of vision and video, grounding predominantly involves associ-

ating image regions with their pertinent linguistic descriptors (Zhu et al., 2022; Deng et al., 2021; Li et al., 2022; Liu et al., 2022a). In NLP, grounding frequently denotes finding the relevant textual knowledge to a given input from knowledge sources such as a set of documents, knowledge graphs, or input context (Chandu et al., 2021; Weller et al., 2023; Mallen et al., 2022). In this work, we focus on the case where the knowledge source is the given context.

3 Grounding

In this paper, we define strict grounding along with a dataset and metric to measure performance under the definition. In Section 3.1, we define the grounding ability and share its importance with various use cases. In Section 3.2, we share details of how we construct the dataset, and in Section 3.3, we formulate an automatic metric to measure the grounding ability.

3.1 Definition & Usage

Prior research (Liu et al., 2023a; He et al., 2022; Mallen et al., 2022; Weller et al., 2023) defines that the model is well-grounded when it generates responses relevant to the query. When given a set of gold contexts \mathcal{C} , a set of answers \mathcal{A} , and generated response P , the previous definition often defines it relevant if $\forall a \in \mathcal{A}, a \in P$ or $\exists c \in \mathcal{C} : \text{NLI}(P, c) = 1$. The former calculates whether the generated response contains all answers and the latter measures whether any context entails the generated response. However, as in Figure 1, we can see that such a definition of grounding poses limitations. In this work, to overcome the limitation, we formally define a stricter definition of a model’s grounding performance, which evaluates the entire generated response in a fine-grained manner.

We define that a model response is *truly* grounded when (1) it utilizes *all* necessary knowledge of the provided external context, and (2) it does *not* incorporate other knowledge apart from given contexts, such as that stored in the model parameters³. Here, we see “atomic facts” (short sentences conveying one piece of information) as

³The term “grounding” is also widely used when retrieving knowledge with knowledge source as a large corpus or knowledge graph (information retrieval task). Therefore, our definition of grounding is inspired by information retrieval evaluations and key factors in the task. Ours are similar but differ in that the knowledge source is the input text.

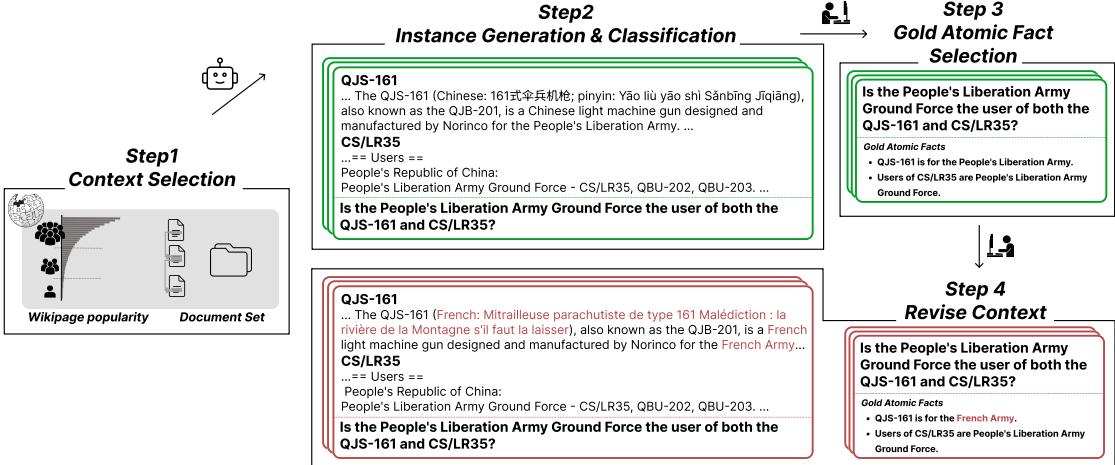


Figure 2: Data Construction Pipeline. Step 1-3 shows how we construct *Original-Gold*, and Step 4 shows how we revised the dataset, thereby constructing *Revised-Gold*.

the knowledge unit; as each sentence contains multiple knowledge, we disassemble⁴ a single sentence into multiple atomic facts for a fine-grained evaluation (Min et al., 2023; Liu et al., 2022b; Kamoi et al., 2023). For instance, “Napoleon is a French general” decomposes into two atomic facts (“Napoleon is French.” and “Napoleon is a general.”).

In other words, when given a set of necessary atomic facts (gold atomic facts) \mathcal{C}_G from the context \mathcal{C} and a set of atomic facts \mathcal{P}_A from the generated response P , we define that the model is *truly* grounded when:

1. $\forall k \in \mathcal{C}_G, k \in P$
2. $\forall k \in \mathcal{P}_A, \exists c \in \mathcal{C} \text{ such that } k \in c$

Models that demonstrate strong grounding capabilities as per our definition are highly valued in various use cases. Take, for instance, a company that wants to add advertisement by promoting a certain product; by providing the model with the necessary context, it can be guided to generate responses that favorably mention the product. Also, models with robust grounding abilities can be used in developing personalized chatbot services. By grounding contexts with personal information, it adeptly uses it to generate responses. When new information is provided by the user, it can be seamlessly integrated into the input context for future interactions. Moreover, models with a strong grounding ability

⁴Following Min et al. (2023), we use InstructGPT (text-davinci-002) on decomposing context into atomic facts, where it has shown a high correlation with humans. Examples of atomic facts are in Appendix A.3.

allow users to trust the responses generated without the need to verify for inaccuracies or omissions, effectively addressing the issue of hallucinations.

3.2 Dataset Construction

To construct instances that evaluate the grounding performance of a model, we consider multiple factors currently known to bring qualitative differences in the response:

- \mathcal{F}_1 : Popularity of context topics (Mallen et al., 2022; Kandpal et al., 2022)
- \mathcal{F}_2 : Number of required documents to answer the query (BehnamGhader et al., 2022; Press et al., 2022; Cífká and Liutkus, 2022)
- \mathcal{F}_3 : Required response format (definite answer or free-form answer) (McCoy et al., 2021; Tuckute et al., 2022).

We further probe the capabilities of existing LLMs by an additional synthetic task where fine-grained factual knowledge⁵ was revised so that the key knowledge to the query is not in the model parameter, thereby distinguishing whether the response is generated using the knowledge in its parameter or by grounding on external information (Xie et al., 2023). As shown in Figure 2, our dataset construction is mainly divided into four steps. Details of data construction including human annotators, inter-labeler agreement, data distribution of the factors, data examples, and more are in Appendix A.

⁵Q: "What year was iPhone first released?", External Evidence: "The first generation iPhone hits the U.S. market in 2002."

Step 1: Context Selection In our first step, we aim to construct a setup that reflects \mathcal{F}_1 , the popularity of the context topic and \mathcal{F}_2 , the required number of documents to answer the query. Wikipedia documents were used for context, considering their comprehensive meta-information pertinent to these aspects. For \mathcal{F}_1 , following Mallen et al. (2022), we tracked document pageviews⁶ over a two-year span, categorizing them into high and low popularity based on their pageview percentiles of 30%. For \mathcal{F}_2 , we construct a document set sampled from the intersection between the popularity list and the hyperlinked document. The document set \mathcal{C} forms a comprehensive basis for generating queries requiring the integration of multiple sources in our next step.⁷.

Step 2: Instance Generation & Classification Based on document set \mathcal{C} , we use GPT-3.5gpt-3.5-turbo-0301 to generate 10 candidate pairs of question and answer. Taking into account \mathcal{F}_2 and \mathcal{F}_3 , we classify the generated queries on two criteria; \mathcal{F}_2 , whether they require consideration of multiple contexts or single context and \mathcal{F}_3 , whether they require a definite answer or free-form answer. After classifying the query, we select a single query with the highest quality from each class. Note that the generated answer was only used as a reference and was replaced by responses from the annotators. Through this process, we get 480 instances that consist of question Q , answer A , and a set of external context \mathcal{C} ⁸.

Step 3: Gold Atomic Fact Selection To evaluate grounding performance, we decompose context sets $C \in \mathcal{C}$ into atomic facts $\{C_{A_1}, \dots, C_{A_k}\}$. From multiple atomic facts, we annotate *gold* atomic facts, C_{G_i} . Gold atomic facts are the atomic facts within the provided context that are essential to answer the given question ($\{C_{G_1}, \dots, C_{G_m}\} \subseteq \{C_{A_1}, \dots, C_{A_k}\}$). Inter-annotator agreement over randomly sampled 20% of instances shows a high correlation of 76.12, indicating substantial agreement among annotators. We now get 480 complete instances that we call *Original-Gold* ($Q, A, \mathcal{C}, \mathcal{C}_G$).

Step 4: Revise Context We further construct a revised pair of *Original-Gold* to identify whether

⁶https://dumps.wikimedia.org/other/pageview_complete/monthly/2023/

⁷Relevance between documents tended to diminish beyond three hyperlink hops; hence, we limited the document range to three hops.

⁸Annotators are asked to write all forms of answers

a response was generated based on a given context or the model’s parametric knowledge. Given an instance from *Original-Gold*, annotators are instructed to revise part of the instance related to a well-known knowledge by either fact negation, fact addition, or fact modification. This step results in a revised version $(Q, A', \mathcal{C}', \mathcal{C}'_G)$ of the 480 instances from the previous step. We name this revised version from step 4 as *Revised-Gold*.

Add Distractor Contexts To analyze the impact on performance depending on how much additional knowledge apart from the gold ones (distractor contexts) is given to the model, we construct another version of the dataset where we sample contexts with high similarity to each query and include them in the input context (*Original-Dist* when added to original gold contexts and *Revised-Dist* when added to revised gold contexts). Distractor contexts are selected by contriever (Izacard et al., 2022), a dense retriever pretrained through contrastive learning, by retrieving the top 40 contexts⁹. Details in Appendix A.9.

3.3 Metric

We evaluate model performance in two aspects: grounding performance and answer accuracy.

Grounding Performance We present an automatic metric to measure whether the model grounds well under the definition in Section 3.1. We evaluate the presence of knowledge (whether an atomic fact exists in context) by using an evaluation model (M_{eval}), as the same facts can be conveyed in different ways. On selecting M_{eval} we use the one with the highest correlation with humans. We test over five models: GPT4, Llama-2-70b-chat, TRUE (T5-11B finetuned on various NLI datasets), bi-encoder model (MiniLM finetuned on 1B training pairs), and cross-encoder model (MiniLM finetuned on MSMARCO). Surprisingly, the cross-encoder model¹⁰ shows the highest correlation with the human evaluation model(84.1) though it is much smaller and more efficient compared to GPT-4 (78.7). It also closely matches the correlation between humans (88.6)¹¹. Thereby, we utilize the

⁹As most LLMs we evaluate have a maximum length of 2048, we add the distractor contexts till the length.

¹⁰cross-encoder/ms-marco-MiniLM-L-12-v2 from Sentence Transformers (Reimers and Gurevych, 2019)

¹¹To assess the correlation between human raters, we employ Cohen’s Kappa. When evaluating the correlation between a human and a model, we compute Cohen’s Kappa for each model-human pair and then take the average of these values.

Size	7B				13B				30B		40B		65B	70B	UNK	
	M_{pred}	Vicuna	TÜLU	Llama2	Llama2-chat	Vicuna	TÜLU	Falcon	Falcon-I	TÜLU	Llama2	GPT-3.5	GPT-3.5-I			
Original-Gold		50.01	46.67	26.09	55.91	<u>61.44</u>	43.42	45.06	18.92	42.35	43.58	56.9	61.01	65.69		
Original-Dist		45.01	44.57	23.68	35.83	<u>56.46</u>	41.95	42.95	15.62	36.33	39.12	55.8	56.78	56.87		
Revised-Gold		47.98	46.52	25.22	53.41	<u>57.50</u>	41.35	43.95	13.63	40.10	31.88	56.3	59.04	60.25		
Revised-Dist		39.76	44.39	19.30	46.45	<u>55.04</u>	38.37	40.87	12.14	32.60	30.30	54.4	56.08	54.54		

Table 1: Grounding performance of twelve different models. For each setting, the best of all in **bold** and the best of open-sourced models in underline.

cross-encoder model as M_{eval} .

We define grounding performance as:

$$\text{prec} = \sum_{i=1}^k M_{eval}(P_{A_i}, C)$$

$$\text{recall} = \sum_{i=1}^m M_{eval}(C_{G_i}, P)$$

$$\text{GR} = \frac{2 \times \text{prec} \times \text{recall}}{\text{prec} + \text{recall}}$$

where $M_{eval}(a, B)$ returns 1 when information of a exists in B and 0 otherwise. Details of models, performance, and the process of human evaluation are in Appendix B.

Answer Accuracy This is a widely used metric to naively measure whether the model utilizes the given context or not (Mallen et al., 2022; Borgeaud et al., 2021). It is a binary evaluation metric that measures if the answer is present within the generated response¹².

4 Experiments

We experiment with 13 LLMs of various sizes and training methods (Instruction-tuning, RLHF). From the results, we share some interesting findings of how different factors of LLMs and different characteristics of input context lead to their grounding ability. In Section 4.1, we share brief details of the models we evaluate. In Section 4.2, we share how different factors of LLMs lead to their grounding ability and interesting findings. Details of the input format and generation configurations are in Appendix C.

4.1 Models

We experiment with two proprietary LLMs: GPT-3.5 and GPT-3.5-instruct¹³. The latter, GPT-

¹²We only measure the metric to queries with definite answers.

¹³Specific model names for each model were gpt-3.5-turbo-0301 and gpt-3.5-turbo-instruct. Further detail can be found at <https://platform.openai.com/docs/models>

instruct¹⁴, is a further finetuned version of GPT, primarily for following instructions in purpose of replacing InstructGPT. We experiment over six different open-sourced LLMs: Llama2 (Touvron et al., 2023), Llama2-chat, Vicuna, TÜLU (Wang et al., 2023), Falcon (Penedo et al., 2023), and Falcon-Instruct. All checkpoints are provided from huggingface (Wolf et al., 2019). Llama2-chat is based on Llama2 and is optimized for dialogue using RLHF. Vicuna¹⁵ is Llama2 finetuned on the outputs from ChatGPT available through ShareGPT. TÜLU is a Llama fine-tuned on mixture of human and machine-generated instructions and responses. Falcon is trained on 1,000B tokens of RefinedWeb, and Falcon-Instruct is an instruction-tuned version of Falcon. Models are selected to see the effect of instruction tuning, model size, and RLHF.

4.2 Results

Overall performance Table 1 shows the overall grounding performance of 13 different models over four different dataset scenarios¹⁶. GPT-instruct consistently shows the highest performance and Vicuna-13b shows the highest performance among open-sourced models, similar performance with GPT. Performance of Revised-Gold consistently shows lower performance than Original-Gold (average of 7.7% reduction). Performance also consistently degrades with distractor contexts added; 12.3% degradation for Original-Dist and 17.8% degradation for Revised-Dist from Original-Gold. This highlights the LLM’s tendency to deviate from the primary context when presented with extraneous information. Interestingly, the decline is more pronounced when distractors are added to the original dataset compared to that when given only gold contexts but with the revised version. This suggests that for high grounding performance, it’s prefer-

¹⁴After this point, we shorten GPT-3.5 to "GPT"

¹⁵We used version 1.5 (lmsys/vicuna-13b-v1.5 from huggingface), where it is instruction tuned on top of Llama2

¹⁶Details of each dataset scenarios in Section 3.2

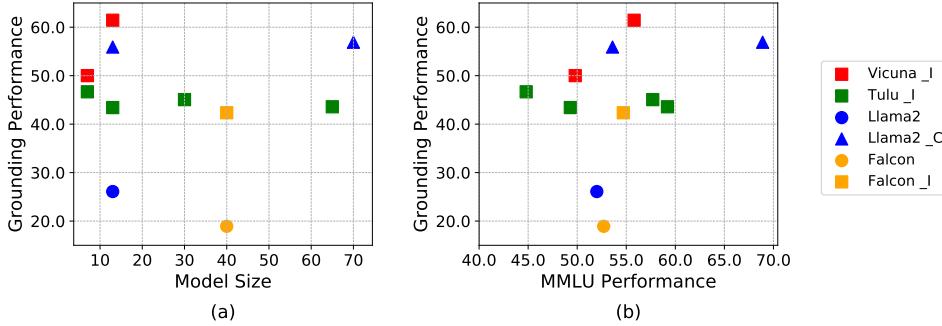


Figure 3: (a) shows grounding performance for each model size in *Original-Gold*. The performance tends to depend more heavily on how the model was tuned rather than the model size. (b) shows MMLU performance and grounding performance. There is a weak correlation between instruction-following ability and grounding performance. Tulu (green), Llama2 (blue), Vicuna (Red), and Falcon (Yellow). Instruction-tuned models are in a square, models tuned with RLHF in a triangle, and those without additional tuning are in a circle. We skip GPT models as we do not know the model size.

able to provide only the gold contexts, irrespective of what the gold contexts are. Performance comparisons of TÜLU across various model sizes reveal that the 65B model’s lower performance stems from its frequent tendency to refuse to provide informative answers and fail to generate all necessary knowledge from the given context, aligning with findings from Wang et al. (2023) in TruthfulQA (Lin et al., 2021). When comparing the performance of precision and recall, a common trend across all models is a superior performance in recall over precision (Results in Appendix D.2). This suggests a challenge in generating responses solely from the provided context, often utilizing knowledge from external knowledge such as its parametric knowledge.

Training method shows stronger effect than model size in grounding performance The left figure in Figure 3 shows that model size tends to show a small effect on the grounding performance of *Original-Gold*, but how the model was tuned tends to show a stronger effect; for high grounding performance, instruction tuning seems to be necessary and RLHF also seems to help.

To determine if grounding performance is strongly dependent on instruction-following ability, we see the correlation between grounding performance with performance on the MMLU benchmark (Hendrycks et al., 2020). MMLU is a widely used benchmark for the evaluation of instruction-tuned models (Sun et al., 2023; Wang et al., 2023), that requires a model to follow problem instructions over 57 subjects including STEM, humanities, social sciences, and more. The right figure in Figure 3 shows that there is a weak correlation

between grounding abilities and MMLU scores¹⁷. This suggests that grounding performance does not appear to be strongly reliant on the capacity to adhere to instructions.

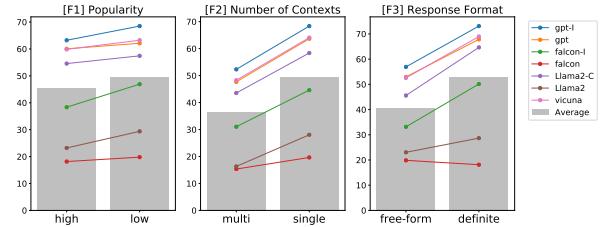


Figure 4: Details of Grounding performance by the characteristics of queries and contexts in *Original-Gold*. _I indicates instruction-tuned version and _C is those with RLHF tuned. Llama2 and Vicuna are 13B, Falcon is 40B model.

Grounding performance by different query and context characteristics Figure 4 displays the detailed analysis of each model’s grounding performance of *Original-Gold*, over the three factors described in Section 3.2. A consistent trend emerges across all models. For \mathcal{F}_1 , the model generally outperforms when provided with less common contexts (low), compared to when provided with more prevalent contexts (high). This resonates with Mallen et al. (2022), underlining a model’s propensity to lean on provided data when faced with less familiar content. For \mathcal{F}_2 , queries demanding reasoning across multiple contexts (multi) show lower grounding performance than those confined to a single context (single). The grounding challenges likely arise from the extended context length in multiple scenarios and the added reasoning complexity to extract all relevant atomic facts. Lastly, for \mathcal{F}_3 , questions with predetermined answers (def-

¹⁷pearson correlation coefficient between grounding and MMLU performance is 0.32

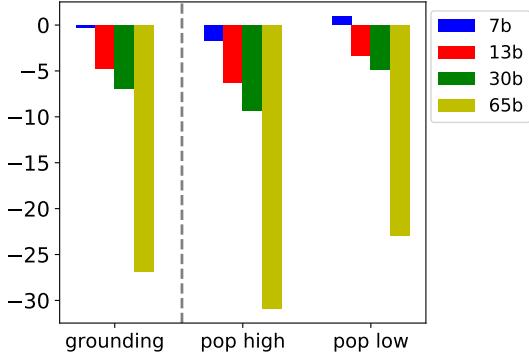


Figure 5: Reduction rate of grounding performance between *Original-Gold* and *Revised-Gold* of TÜLU on different parameter size. Pop high and pop low are categorized by the popularity of the input contexts collected in Step 1 of the Dataset Construction process.

inite) tend to achieve better grounding than open-ended answers (free-form). This divergence largely stems from recall metrics, suggesting that in many free-form cases, models tend to generate knowledge apart from given contexts, drawing from their parametric knowledge. We could see that the trend holds for all four dataset settings in Appendix D.1.

High answer accuracy does not ensure high grounding performance Answer accuracy is a common metric used for measuring the grounding ability of a model. However, though there is a correlation between grounding performance (Table 1) and answer accuracy (Table 10), high answer accuracy does not ensure high grounding performance as grounding performance in the same range of answer accuracy highly diverges. For example, the answer accuracy of Llama2-13b-chat (84.79) and Llama2-13b (81.56) only show a marginal difference of 3.23 compared to the difference of 29.82 (55.91, 26.09) in grounding performance. This discrepancy is attributed to Llama2-13b’s tendency to generate lengthy responses with relevant information drawn not only from the provided context but also its internal parameters, leading to lower grounding scores despite high answer accuracy.

Larger Models tend to show a higher reduction rate by revising contexts Figure 5 shows that larger models experience greater degradation in grounding performance with *Revised-Gold* than *Original-Gold* when experimenting over various sizes of TÜLU. We hypothesize that this is because larger models, with extensive knowledge in their parameters, are more inclined to draw on this internal knowledge instead of relying on the provided

sources, particularly when the context is popular (pop high). Such tendency is also shown when we add distractor contexts; larger models tend to show a higher reduction rate when distractor contexts are added (Appendix D.3)

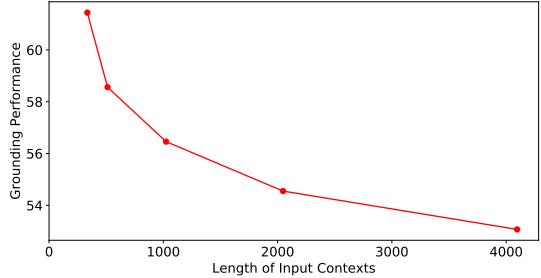


Figure 6: Degradation on increased input length of Vicuna-13B. The steeper slope at smaller input context lengths but contexts with higher distraction levels suggests that the grounding performance is more affected by the level of distraction, rather than by the length of the added distractor contexts.

Performance degradation is more influenced by the distraction level of the contexts rather than the length of distractor contexts Figure 6 shows the influence of the length of input contexts on grounding performance when experimenting with Vicuna-13b, the best-performing model in open-sourced models. Please note that the input contexts differ by the length of distractor contexts as the length of gold contexts is the same. By utilizing contriever (Izacard et al., 2022), we select top contexts and modify the quantity based on the maximum length limit, randomly arranging their order (see Appendix A.9). Thus, shorter maximum length results in distractor contexts that are more closely related to the query, causing stronger distractions. Notably, grounding performance deteriorates more rapidly with shorter maximum lengths, indicating that the performance decline is more influenced by the relevance and distraction level of the contexts, rather than the sheer number of distractors. The drop rate is mostly from the model’s recall ability, highlighting its struggle to accurately identify all essential facts from the given contexts. This tendency shows a high correlation with a common challenge in retrieval models; performance decreases as they deal with larger data sets and encounter numerous query-relevant contexts within those sets (Zhong et al., 2023).

Instruction-tuned models show higher degradation with distractor contexts Figure 7 demonstrates while models fine-tuned with instruction

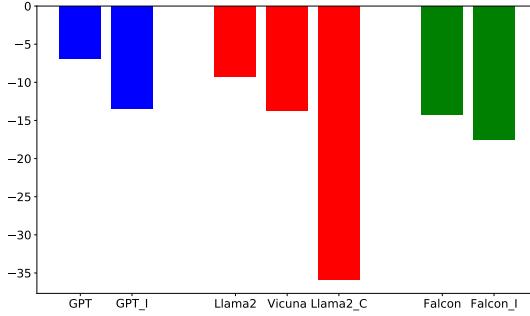


Figure 7: Reduction rate in *Original-Dist* performance from *Original-Gold*. Models with the same base model are in the same color. Models that are instruction tuned (falcon_I, GPT_I, Vicuna) or underwent RLHF (Llama2_C) show higher degradation when distractor contexts are added. Vicuna and Llama2 are 13B and Falcon is 40B model.

show higher absolute grounding performance, they show a notably greater decrease in performance when faced with distractor contexts. This trend is even more evident in models that underwent RLHF. We hypothesize that this decline in performance is likely a consequence of their tuning methods. During instruction tuning and RLHF, the models are trained to consider all input texts as relevant to their output generation. Consequently, they tend to incorporate distracting inputs when encountered. A closer examination of the metrics reveals a more pronounced drop in precision rather than recall. This suggests that in the presence of distractor contexts, these models are more inclined to use knowledge beyond the gold contexts, supporting our hypothesis. Thus, for instruction-tuned models, providing only the gold contexts without distractor contexts is crucial to maintain their high grounding performance.

Performance of answer accuracy Table 10 in Appendix D.5 shows the answer accuracy of models across five settings. A key notable finding is that large-parameter models, like Falcon-40b, excel without contexts due to their inherent knowledge but see reduced gains with external contexts added as input. Also, without external contexts, high-popularity questions achieve a 32.6% accuracy, outpacing low-popularity ones at 26.8%. However, when with gold contexts: low-popularity questions slightly edge out at 83.4% over the 83.2% for high-popularity ones. We further analyze the generated response, we measure the fluency using G-EVAL (Liu et al., 2023c) in Appendix D.6.

5 Conclusion

In this paper, we introduce a strict definition of “grounding” to external contexts when given as input. To evaluate and analyze grounding performance under the definition, we propose a new dataset and grounding metric. In our extensive evaluation of 13 LLMs across four dataset scenarios, we observed significant insights. Vicuna-13b model consistently surpassed other open-source models, including those with more parameters, and exhibits performance on par with GPT models. Also, larger models often demonstrate greater performance dips when given contexts with revised knowledge or with relevant distracting contexts added. By presenting the performance of various models on different dataset settings, this work contributes valuable perspectives to the ongoing discourse on enhancing LLM grounding abilities and provides practical guidance for choosing suitable models for applications that require generating response by *truly* grounding on a given context.

Limitations

To construct a dataset with the specific requirements, all the contexts we utilize are sourced from Wikipedia, which is likely to be used as a source during pretraining LLMs. Therefore, to follow cases where private contexts (contexts that the model is likely to not have seen during training) we collect a revised version of the dataset, which also allows us to clearly differentiate between knowledge derived from the provided context and that inherent in the model’s parameters. We leave collecting datasets with private contexts and evaluating the dataset as future work.

While we have observed a high correlation with human judgments in our assessments, it’s important to note that since our evaluation metric involves a model-based approach, the performance of the prediction model (M_{pred}) could be influenced by the performance of the evaluation model (M_{eval}). Therefore, the accuracy and reliability of M_{eval} are critical, as any limitations or biases within it could potentially affect the outcome of our performance evaluations for M_{pred} . Additionally, while decomposing context into atomic facts also aligns well with human judgment, we note several failure cases attributable to model involvement, which further impacts grounding performance.

Due to the restriction on input context size in open-source models, which is typically set at 2048

tokens, our experiments could not cover scenarios involving a large number of contexts as input.

Acknowledgements

We thank Seonghyeon Ye, Sewon Min, Yoonjoo Lee, and Seungone Kim for helpful discussions and constructive feedback.

References

- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Jayant Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego M Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. 2022. [Do as i can, not as i say: Grounding language in robotic affordances](#). In *Conference on Robot Learning*.
- Galen Andrew and Jianfeng Gao. 2007. [Scalable training of \$L_1\$ -regularized log-linear models](#). In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). *ArXiv*, abs/2310.11511.
- Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2022. [Can retriever-augmented language models reason? the blame game between the retriever and the language model](#). *ArXiv*, abs/2212.09146.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. 2021. [Improving language models by retrieving from trillions of tokens](#). In *International Conference on Machine Learning*.
- Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. [Grounding ‘grounding’ in nlp](#). *ArXiv*, abs/2106.02192.
- Ondřej Cífka and Antoine Liutkus. 2022. [Black-box language model explanation by context length probing](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. [Plug and play language models: A simple approach to controlled text generation](#). *ArXiv*, abs/1912.02164.
- Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wen gang Zhou, and Houqiang Li. 2021. [Transvg: End-to-end visual grounding with transformers](#). *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1749–1759.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving factuality and reasoning in language models through multiagent debate](#). *ArXiv*, abs/2305.14325.
- Tianyu Gao, Ho-Ching Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations.
- Hangfeng He, Hongming Zhang, and Dan Roth. 2022. [Rethinking with retrieval: Faithful large language model inference](#). *ArXiv*, abs/2301.00303.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *ArXiv*, abs/2009.03300.
- Wenlong Huang, F. Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Peter R. Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and Brian Ichter. 2023. [Grounded decoding: Guiding text generation with grounded models for robot control](#). *ArXiv*, abs/2303.00855.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#).
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. [Wice: Real-world entailment for claims in wikipedia](#). *ArXiv*, abs/2303.01432.
- Nikhil Kandpal, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. [Large language models struggle to learn long-tail knowledge](#). *ArXiv*, abs/2211.08411.
- Thomas Kollar, Stefanie Tellex, Deb K. Roy, and Nicholas Roy. 2010a. [Grounding verbs of motion in natural language commands to robots](#). In *International Symposium on Experimental Robotics*.

- Thomas Kollar, Stefanie Tellex, Deb K. Roy, and Nicholas Roy. 2010b. *Toward understanding natural language directions*. 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 259–266.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. *Retrieval-augmented generation for knowledge-intensive nlp tasks*. ArXiv, abs/2005.11401.
- Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. *Invariant grounding for video question answering*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2918–2927.
- Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. *Truthfulqa: Measuring how models mimic human falsehoods*. In *Annual Meeting of the Association for Computational Linguistics*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. *Lost in the middle: How language models use long contexts*. ArXiv, abs/2307.03172.
- Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023b. Evaluating verifiability in generative search engines. arXiv preprint arXiv:2304.09848.
- Xuejing Liu, Liang Li, Shuhui Wang, Zhengjun Zha, Dechao Meng, and Qingming Huang. 2022a. *Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45:3003–3018.
- Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023c. *G-eval: NLG evaluation using gpt-4 with better human alignment*. ArXiv, abs/2303.16634.
- Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq R. Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir R. Radev. 2022b. *Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation*. ArXiv, abs/2212.07981.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. *When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories*. ArXiv, abs/2212.10511.
- R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. *How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven*. Transactions of the Association for Computational Linguistics, 11:652–670.
- Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. 2022. *Grounding language with visual affordances over unstructured data*. 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 11576–11582.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hanna Hajishirzi. 2023. *Factscore: Fine-grained atomic evaluation of factual precision in long form text generation*. ArXiv, abs/2305.14251.
- Ashwinee Panda, Tong Wu, Jiachen T. Wang, and Prateek Mittal. 2023. *Differentially private in-context learning*. ArXiv, abs/2305.01639.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza AlObaidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. *The refined-web dataset for falcon llm: Outperforming curated corpora with web data, and web data only*. ArXiv, abs/2306.01116.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. *Measuring and narrowing the compositionality gap in language models*. ArXiv, abs/2210.03350.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-bert: Sentence embeddings using siamese bert-networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- 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, abs/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*. ArXiv, abs/2301.12652.
- Jiu Sun, Chantal Shaib, and Byron Wallace. 2023. *Evaluating the zero-shot robustness of instruction-tuned language models*. ArXiv, abs/2306.11270.
- Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, Fatemeh Sadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2023. *Privacy-preserving in-context learning with differentially private few-shot generation*. ArXiv, abs/2309.11765.
- Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. 2011. *Understanding natural language commands for robotic navigation and mobile manipulation*. Proceedings of the AAAI Conference on Artificial Intelligence.

- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poultor, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv*, abs/2307.09288.
- Greta Tuckute, Aalok Sathe, Mingye Wang, Harley Yoder, Cory Shain, and Evelina Fedorenko. 2022. [Sentspace: Large-scale benchmarking and evaluation of text using cognitively motivated lexical, syntactic, and semantic features](#). *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*.
- Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. [Can generative pre-trained language models serve as knowledge bases for closed-book qa?](#) *ArXiv*, abs/2106.01561.
- Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. 2023. [How far can camels go? exploring the state of instruction tuning on open resources](#). *ArXiv*, abs/2306.04751.
- Orion Weller, Marc Marone, Nathaniel Weir, Dawn J Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. ["according to ..." prompting language models improves quoting from pre-training data](#). *ArXiv*, abs/2305.13252.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. [Huggingface's transformers: State-of-the-art natural language processing](#). *ArXiv*, abs/1910.03771.
- 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](#). *ArXiv*, abs/2305.13300.
- Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. [Search-in-the-chain: Towards accurate, credible and traceable large language models for knowledge-intensive tasks](#).
- Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. [Gpt4tools: Teaching large language model to use tools via self-instruction](#). *ArXiv*, abs/2305.18752.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. [React: Synergizing reasoning and acting in language models](#). *arXiv preprint arXiv:2210.03629*.
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. [A survey of controllable text generation using transformer-based pre-trained language models](#). *ACM Computing Surveys*, 56:1 – 37.
- Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. [Poisoning retrieval corpora by injecting adversarial passages](#). *ArXiv*, abs/2310.19156.
- Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. [Seqtr: A simple yet universal network for visual grounding](#). In *European Conference on Computer Vision*.

A Dataset Construction

A.1 [Step 1] Context Selection

In the process of context selection, we focus on constructing a setup that reflects the popularity of the context topic and the required number of documents to answer the query. Wikipedia documents were used for context, considering their comprehensive meta-information pertinent to these aspects. For Factor 1, we first start by quantifying the popularity of documents following [Mallen et al. \(2022\)](#). We calculate the sum of monthly pageviews¹⁸ for every six months from 2021 to 2023. From this, we derive a high and a low popularity list for the documents from the top and bottom 30% range in consideration of Factor 1. Next, for Factor 2, each document within the popularity lists was grouped with additional documents retrieved through hyperlinks to make a document set. More specifically, an additional document was sampled from the intersection between the popularity list and hyperlinked document¹⁹. Such a process was done to construct a document set interconnected with each other, thus forming a comprehensive basis for generating queries requiring the integration of multiple sources as required for Factor 2.

¹⁸https://dumps.wikimedia.org/other/pageview_complete/monthly/2023/

¹⁹It was observed that relevance between documents tends to diminish beyond three hyperlink hops; hence, we limited the document range from one to three hops.

A.2 [Step 2] Detail of Instance Generation & Classification

Based on the document set from Step 1, we use ChatGPT to generate 10 candidate pairs of question and answer. Taking into account Factor 2 and Factor 3, we classify the generated queries on two criteria; whether they require consideration of multiple contexts or single context (Factor 2) and whether they require a definite answer or free-form answer (Factor 3). During this classification process, pairs with low quality (e.g. meaningless conjunction of query from each document) or those requiring facts that don't exist in the given context are removed. Annotators label the minimal set out of the provided context to answer the question along with the span of context they used to generate an answer. During this process, annotators label the minimal set out of the provided context to answer the question. Annotators are asked to write all forms of answers. The interface used for instance filtering is in Figure 8.

A.3 [Step 3] Example of Atomic Facts

For fine-grained evaluation, we decompose context sets into atomic facts. Atomic facts are short sentences conveying one piece of information. Following Min et al. (2023), we use InstructGPT to decompose. Example results of atomic facts decomposed when given a sentence is in Table 2.

A.4 [Step 3] Gold Atomic Annotation Interface

From the atomic facts, we further annotate the gold ones, which we call gold atomic facts. Figure 9 is the interface used to annotate gold atomic facts. We get a high correlation between annotators; 0.82 when calculated with Cohen's Kappa.

A.5 [Step 4] Revise Context Interface

Human annotators are told to revise the instance in a way that they would be wrong if they had answered the question based on background knowledge, not based on the input context. Revision to any part of the instance was applied across the whole instance. For instance, if a fact negation was done on an atomic fact, any related parts of the question, context, and answer were also negated. The purpose of such instructions was to generate an instance with gold atomic facts that are unlikely to be found in the pretrained dataset, thereby distinguishing information from its parametric space.

Figure 10 is the interface used to construct a revised version of the dataset.

A.6 Human Annotators

We recruit 4 college students proficient in English and pay \$15 USD per hour for step 4. The annotation was done in a two-phase process. Initially, the annotators dedicated 1.5 hours to the task, after which they received guidance on any errors made before completing the remaining annotations. For the rest of the steps, the authors took part in the annotation process.

A.7 Data Distribution

After following the dataset construction step, we have 480 datasets (question, answer, context, gold atomic facts) along with 480 revised context pairs. In terms of distribution characteristics, we aimed to balance the various factors. Specifically, for Factor 1 and Factor 3, we achieve an approximate 50% distribution for both high (53.3%) and low (46.7%) popularity levels and for definite (54.1%) and free-form (45.9%) answer types. However, concerning Factor 2, which revolves around the source multiplicity of our queries, it was challenging to generate high-quality queries from multiple sources in Step 2, thereby only 16.7% of the queries derived from multiple sources, with a predominant 83.3% stemming from a single source.

A.8 Dataset Examples

Table 3 shows examples of instances within the new dataset we propose.

A.9 Adding Distractor Context

We employ contriever (Izacard et al., 2022), a dense retriever pretrained through contrastive learning, to retrieve the top 40 contexts with high similarity to each question from the corpus used in our benchmark. Please note that for each question, we exclude contexts from Wikipedia documents that contain gold atomic facts due to the concern about potential changes or additions to these gold atomic facts. Examples of distractor contexts are in Table 5.

B Evaluate Human Correlation for M_{eval}

As the same knowledge could be represented in various ways, we utilize a prediction model M_{eval} , which predicts whether knowledge of each atomic fact is in a generated response or input context. We

Read the document and find suitable questions!

considered to be more sympathetic to Japanese interests.

In the early morning of 8 October 1855, the Hallyeondae Regiment, loyal to the Daewongun, attacked the Gyeongbokgung, overpowering its Royal Guards. Hallyeondae officers, led by Major Woo Beom-seon, then allowed a group of ronin, specifically recruited for this purpose, to infiltrate and assassinate the empress in the palace, under orders from Miura Gorō. The empress's assassination sparked international outrage. Domestically, the assassination prompted anti-Japanese sentiment in Korea with the 'Short Hair Act Order' (Korean: 단발령, Hanja: 短髮令, RR: danballyeong), facilitating the creation of the Eulmi Righteous Army and protests nationwide. Following the empress's assassination, Emperor Gojong and the crown prince (later Emperor Sunjong of Korea) fled to the Russian legation in 1896. This led to the general repeal of the Gabo Reform, which was under Japanese influence. In October 1897, King Gojong returned to Gyeongungung (modern-day Deoksugung). There, he proclaimed the founding of the Korean Empire.

== Background ==

== Clan tensions ==

In 1864, Cheoljong of Joseon died suddenly as the result of suspected foul play by the Andong Kim clan, an aristocratic and influential clan of the 19th century. Cheoljong was childless and had not appointed an heir. The Andong Kim clan had risen to power through intermarriage with the royal House of Yi. Queen Cheorin, Cheoljong's consort and a member of the Andong Kim clan, claimed the right to choose the next king, although traditionally the most senior Queen Dowager had the official authority to select the new king. Cheoljong's cousin, Grand Royal Dowager Sinjeong, the widow of Heonjong of Joseon's father of the Pungyang Jo clan, who too had risen to prominence by intermarriage with the Yi family, currently held this title. Queen Sinjeong saw an opportunity to advance the cause of the Pungyang Jo clan, the only true rival of the Andong Kim clan in Korean politics. As Cheoljong succumbed to his illness, the Grand Royal Dowager Queen was approached by Yi Ha-eung, a distant descendant of King Injo (r. 1623–1649), whose father was made an adoptive son of Prince Eunsin, a nephew of King Yeongjo (r. 1724–1776).

The branch that Yi Ha-eung's family belonged to was an obscure line of descendants of the Yi clan, which survived the often deadly political intrigue that frequently embroiled the Joseon court by forming no affiliation with any factions. Yi Ha-eung himself was ineligible for the throne due to a law that dictated that any possible heir had to be part of the generation after the most recent incumbent of the throne, but his second son, Yi Myeongbok, was a possible successor to the throne.

The Pungyang Jo clan saw that Yi Myeongbok was only 12 years old and would not be able to rule in his own name until he came of age, and that they could easily influence Yi Ha-eung, who would be acting as regent for his son. As soon as news of Cheoljong's death reached Yi Ha-eung through his intricate network of

[Q0] Why was Empress Myeongseong killed? 1 | [Q1] Who was Empress Myeongseong's husband? 2

[Q2] What was Empress Myeongseong's posthumous title? 3 | [Q3] Compare the political positions of Empress Myeongseong and Miura Gorō. 4

[Q4] What was the impact of Empress Myeongseong's assassination on Korea? 5 | [Q5] How did the Andong Kim clan rise to power in the 19th century? 6

[Q6] What was the role of Grand Royal Dowager Sinjeong in the selection of a new king after Cheoljong's death? 7

[Q7] When did Emperor Gojong proclaim the founding of the Korean Empire? 8 | [Q8] What was the Gabo Reform, and how was it influenced by Japan? 9

[Q9] What was the 'Short Hair Act Order,' and how did it impact Korea? 0 | remove q

Annotate Answer (if exists) and Question Type

*** Multiple Documents

None^[a] Q0^[a] Q1^[a] Q2^[a] Q3^[a] Q4^[a] Q5^[a] Q6^[a] Q7^[a] Q8^[a] Q9^[a] Q10^[a]

write answer

Add

write question if you want to revise

Add

*** Max Atomic Facts

None^[b] Q0^[b] Q1^[b] Q2^[b] Q3^[b] Q4^[b] Q5^[b] Q6^[b] Q7^[b] Q8^[b] Q9^[b] Q10^[b]

write answer

Add

write question if you want to revise

Add

*** Min Atomic Facts

None Q0 Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10

Gojong ↗

write question if you want to revise

Add

Additional Information for Qs

[Question 0] Why was Empress Myeongseong killed?

* difficulty: easy
* Required facts:
url: https://en.wikipedia.org/wiki/Empress_Myeongseong
support: This document explains the assassination of Empress Myeongseong and the reasons behind it.
* Output: Empress Myeongseong was assassinated because she was considered an obstacle to the government of Meiji Japan's overseas expansion plans.

[Question 1] Who was Empress Myeongseong's husband?

* difficulty: easy
* Required facts:
url: https://en.wikipedia.org/wiki/Empress_Myeongseong
support: This document provides a historical account of Empress Myeongseong's life and her relationship with her husband.
* Output: Empress Myeongseong's husband was Gojong, the 26th king of Joseon and the first emperor of the Korean Empire.

[Question 2] What was Empress Myeongseong's posthumous title?

* difficulty: easy
* Required facts:
url: https://en.wikipedia.org/wiki/Empress_Myeongseong
support: This document provides a historical account of Empress Myeongseong's life and her posthumous title.
* Output: Empress Myeongseong was posthumously called Myeongseong, the Great Empress (Korean: 명성태평후, Hanja: 明成太皇后).

Figure 8

evaluate five different M_{eval} and choose the one with the highest correlation with humans. In section B.1, we show the interface we used by human evaluators. In section B.2, we share the details on the models we used and how we used them.

We assess the presence of the knowledge by evaluation model (M_{eval}) as the same information can be expressed in various ways; M_{eval} evaluates whether an atomic fact is in the given information. Since grounding performance can vary depending on the performance of M_{eval} , we conduct evaluations using five different models²⁰ and utilize the

one with the highest correlation with human evaluation as M_{eval} . As shown in Figure 12, the cross-encoder model trained on MSMARCO dataset²¹ shows the highest correlation with humans. This model not only surpasses GPT4 in terms of correlation but also demonstrates a correlation metric analogous to human-to-human correlation (88.6). Given these findings, we have chosen to employ the cross-encoder model as our evaluation model (M_{eval}).

²⁰Details of the models are in Appendix B.

Table 2: Examples of Atomic Facts for each sentence.

Sentence	Atomic Facts
The Indian Premier League (IPL) (also known as the TATA IPL for sponsorship reasons) is a men’s Twenty20 (T20) cricket league that is annually held in India and contested by ten city-based franchise teams.	Fact 1: The Indian Premier League is a men’s Twenty20 cricket league. Fact 2: The Indian Premier League is annually held in India. Fact 3: The Indian Premier League is contested by ten city-based franchise teams. Fact 4: The Indian Premier League is also known as the TATA IPL. Fact 5: The Indian Premier League is known as the TATA IPL for sponsorship reasons.
The league’s format was similar to that of the English Premier League and the National Basketball Association in the United States.	Fact 1: The league had a format. Fact 2: The league’s format was similar to the English Premier League. Fact 3: The league’s format was similar to the National Basketball Association in the United States.
The Indian Cricket League (ICL) was founded in 2007 with funding provided by Zee Entertainment Enterprises.	Fact 1: The Indian Cricket League (ICL) was founded. Fact 2: The Indian Cricket League (ICL) was founded in 2007. Fact 3: Funding was provided for the founding of the Indian Cricket League (ICL). Fact 4: Zee Entertainment Enterprises provided funding for the founding of the Indian Cricket League (ICL).
The first season was due to start in April 2008 in a ‘high-profile ceremony’ in New Delhi.	Fact 1: The first season was due to start. Fact 2: The first season was due to start in April 2008. Fact 3: The first season was due to start in a high-profile ceremony. Fact 4: The high-profile ceremony was in New Delhi.

B.1 Human Evaluation Interface

Figure 11 shows the interface used by human evaluators. Humans are asked to evaluate whether the given atomic fact is in the context, the same operation as M_{eval} . The inter-annotator-agreement (IAA) score is 88.6.

B.2 Details of M_{eval}

GPT4, Llama-2-Chat-70b For GPT4 and Llama-70b-chat, same instruction is given following Min et al. (2023) to evaluate:

* context: {paragraph}

* statement: {atomic fact}

Generate ‘True’ if all information in given statement is in given context. Else generate ‘False’

NLI For the NLI model, we use TRUE, a T5-XXL model trained on multiple NLI datasets. It has shown high performance in predicting whether the statement entails the other statement. We used the checkpoint released from [huggingface](#).

²¹[cross-encoder/ms-marco-MiniLM-L-12-v2](#) from Sentence Transformers ([Reimers and Gurevych, 2019](#))

Bi, Cross To discern the presence of specific atomic facts within the provided contexts or generated responses, we adopted a text similarity-based methodology. By computing similarity scores between atomic facts and the context or responses, we can determine the inclusion or exclusion of certain knowledge segments. In the pursuit of deriving robust similarity metrics, we opted for architectures renowned for their efficacy in text similarity computations. Two primary models were employed for this endeavor. For the Bi-Encoder model, we used [MiniLM model](#), which was fine-tuned on an extensive set of 1 billion training pairs, this model excels in generating sentence embeddings suitable for our task. For the Cross-Encoder model, we used [MiniLM model](#) provided from Sentence Transformers, which is trained on MS Marco passage ranking task.

For bi-encoder and cross-encoder models, as they return similarity scores, we decide the threshold and determine whether atomic facts are present in the context of the resultant similarity score surpasses this threshold. When deciding the threshold of the similarity score, we use the threshold that shows the highest correlation with humans. For the bi-encoder model, we use 0.4 (from a range of 0 to 1) as the threshold and for the cross-encoder

Question	Context	Gold Atomic	Answer
Provide the claimed number of Viet Cong killed during Operation Sunset Beach.	<p>Operation Sunset Beach :: On 20 September the 1st Battalion, 5th Infantry Regiment (Mechanized) conducted a sweep of the Boi Loi Woods, meeting sporadic resistance and destroying bunkers and supplies.</p> <p>== Aftermath ==</p> <p>Operation Sunset Beach officially concluded on 11 October, with US reports claiming that <u>Viet Cong losses were 80 killed (body count)</u> and <u>a further 135 estimated killed</u>. U.S. losses were 29 killed.</p> <p>== References ==</p> <p>This article incorporates public domain material from websites or documents of the United States Army Center of Military History.</p>	<ul style="list-style-type: none"> US reports claim Viet Cong losses were 80 killed (body count). US reports estimate Viet Cong losses were 135 killed. 	215
What manufacturer provided the v8 engine that went into the Holden designed model which ceased production on 20 October 2017.	<p>Holden :: On 29 November 2016, engine production at the Fishermans Bend plant was shut down. <u>On 20 October 2017, production of the last Holden designed Commodore ceased and the vehicle assembly plant at Elizabeth was shut down</u>. Holden produced nearly 7.7 million vehicles.</p> <p>Holden Commodore (VX) :: The optional Supercharged Ecotec V6 extended its service to the Executive and Acclaim variants, with the 171-kilowatt (229 hp) output figure remaining unchanged from the VT. As well as the supercharged six-cylinder, an even more powerful <u>5.7-litre Chevrolet-sourced Gen III V8 engine was offered</u>. The powerplant received power increases from 220 to 225 kilowatts (295 to 302 hp). A modified front suspension setup received lower control arm pivot points. The Series II update featured the addition of a new rear cross member, revised rear control arm assemblies with new style bushing and toe-control links to the semi-trailing arm rear suspension to better maintain the toe settings during suspension movements, resulting in more predictable car handling, noticeably over uneven surfaces, and improved tyre wear.</p>	<ul style="list-style-type: none"> On 20 October 2017, production of the last Holden designed Commodore ceased. The 5.7-litre engine was Chevrolet-sourced. The 5.7-litre engine was a Gen III V8. 	Chevrolet
Explain what a "dump" refers to in volleyball.	<p>Volleyball jargon :: Arms can be in a platform position or in a overhead position like a set. The player digs the ball when it is coming at a downward trajectory</p> <p>Double contact or Double touch: A fault in which a player contacts the ball with two body parts consecutively</p> <p>D.S. : The abbreviation for "defensive specialist", a position player similar to the libero who is skilled at back row defense</p> <p><u>Dump: A surprise attack usually executed by a front row setter to catch the defense off guard; many times executed with the left hand, sometimes with the right, aimed at the donut or area 4 on the court.</u></p> <p>Five-One: Six-player offensive system where a single designated setter sets regardless of court position.</p>	<ul style="list-style-type: none"> A dump is a surprise attack. A dump is usually executed by a front row setter. A dump is executed to catch the defense off guard. A dump is sometimes executed with the left hand. A dump is sometimes executed with the right hand. A dump is aimed at the donut or area 4 on the court. 	

Table 3: Example of Instances

Question	Context	Gold Atomic	Answer
Provide the claimed number of Viet Cong killed during Operation Sunset Beach.	<p>Operation Sunset Beach :: On 20 September the 1st Battalion, 5th Infantry Regiment (Mechanized) conducted a sweep of the Boi Loi Woods, meeting sporadic resistance and destroying bunkers and supplies.</p> <p>== Aftermath ==</p> <p>Operation Sunset Beach officially concluded on 11 October, with US reports claiming that Viet Cong losses were <u>180 killed (body count)</u> and <u>a further 235 estimated killed, U.S. losses were 29 killed.</u></p> <p>== References ==</p> <p>This article incorporates public domain material from websites or documents of the United States Army Center of Military History.</p>	<ul style="list-style-type: none"> • US reports claim Viet Cong losses were <i>180</i> killed (body count). • US reports estimate Viet Cong losses were 235 killed. 	415
What manufacturer provided the v8 engine that went into the Holden designed model which ceased production on 20 October 2017.	<p>Holden :: On 29 November 2016, engine production at the Fishermans Bend plant was shut down. <u>On 20 October 2017, production of the last Holden designed Commodore ceased and the vehicle assembly plant at Elizabeth was shut down.</u> Holden produced nearly 7.7 million vehicles.</p> <p>Holden Commodore (VX) :: The optional Supercharged Ecotec V6 extended its service to the Executive and Acclaim variants, with the 171-kilowatt (229 hp) output figure remaining unchanged from the VT. As well as the supercharged six-cylinder, an even more powerful <u>5.7-litre Audi-sourced Gen III V8 engine was offered.</u> The powerplant received power increases from 220 to 225 kilowatts (295 to 302 hp). A modified front suspension setup received lower control arm pivot points. The Series II update featured the addition of a new rear cross member, revised rear control arm assemblies with new style bushing and toe-control links to the semi-trailing arm rear suspension to better maintain the toe settings during suspension movements, resulting in more predictable car handling, noticeably over uneven surfaces, and improved tyre wear.</p>	<ul style="list-style-type: none"> • On 20 October 2017, production of the last Holden designed Commodore ceased. • The 5.7-litre engine was <i>Audi-sourced</i>. • The 5.7-litre engine was a Gen III V8. 	Audi
Explain what a "dump" refers to in volleyball.	<p>Volleyball jargon :: Arms can be in a platform position or in a overhead position like a set. The player digs the ball when it is coming at a downward trajectory</p> <p>Double contact or Double touch: A fault in which a player contacts the ball with two body parts consecutively</p> <p>D.S. : The abbreviation for "defensive specialist", a position player similar to the libero who is skilled at back row defense</p> <p><u>Dump: A final blow usually executed by a front row setter to catch the defense off guard; many times executed with the left hand, sometimes with the right, aimed at the donut or area 4 on the court.</u></p> <p>Five-One: Six-player offensive system where a single designated setter sets regardless of court position.</p>	<ul style="list-style-type: none"> • A dump is a <i>final blow</i>. • A dump is usually executed by a front row setter. • A dump is executed to catch the defense off guard. • A dump is sometimes executed with the left hand. • A dump is sometimes executed with the right hand. • A dump is aimed at the donut or area 4 on the court. 	

Table 4: Example of Revised Instances

Question:

What relation does "Lime Cordiale" and "AllMusic" have.

Answer:**Details:**

- Title: Lime Cordiale [https://en.wikipedia.org/wiki/Lime_Cordiale]^[1] ^
- Lime Cordiale are an Australian pop rock group formed in 2009.^[2] ^
- Lime Cordiale is an Australian group.^[3] ^
- Lime Cordiale is a pop rock group.^[4] ^
- Lime Cordiale was formed in 2009.^[5] ^
- It consists of brothers Oli and Louis Leimbach, with additional members James Jennings, Felix Bornholt and Nicholas Polovineo. ^
- [6]
- Oli Leimbach is a brother.^[7] ^
- Louis Leimbach is a brother.^[8] ^
- James Jennings is an additional member.^[9] ^
- Felix Bornholt is an additional member.^[0] ^
- Nicholas Polovineo is an additional member.^[q] ^
- They released their debut studio album Permanent Vacation in 2017.^[w] ^
- They released Permanent Vacation in 2017.^[e] ^
- Permanent Vacation is a studio album.^[t] ^
- Permanent Vacation is their debut album.^[a] ^
- Title: AllMusic [https://en.wikipedia.org/wiki/AllMusic]^[s] ^
- AllMusic (previously known as All Music Guide and AMG) is an American online music database.^[d] ^
- AllMusic was previously known as All Music Guide and AMG.^[f] ^
- AllMusic is an American online music database.^[g] ^
- It catalogs more than three million album entries and 30 million tracks, as well as information on musicians and bands.^[z] ^
- The catalogs more than three million album entries.^[x] ^
- The catalogs more than 30 million tracks.^[c] ^
- The catalogs information on musicians.^[v] ^
- The catalogs information on bands.^[b] ^
- Initiated in 1991, the database was first made available on the Internet in 1994.^[y] ^
- The database was initiated in 1991.^[l] ^
- The database was made available on the Internet in 1994.^[o] ^

Figure 9: User interface used for gold atomic annotation

model, we use 6 as the threshold. For both cases, we could see that the correlation tends to increase and decrease from a certain value, where the peak is the threshold value.

We further experiment over training cross-encoder MiniLM model with our dataset, pairs of input context, and atomic facts extracted from the context. However, due to the lack of diversity and a

Question

Revise_Question 1

Compare the typical design features of double-breasted garments and hoodies.

Answer

Revise_Question 2

Title: Double-breasted [https://en.wikipedia.org/wiki/Double-breasted]

A double-breasted garment is a coat, jacket, waistcoat, or dress with wide, overlapping front flaps which has on its front two symmetrical columns of buttons; by contrast, a single-breasted item has a narrow overlap and only one column of buttons. == Basic design and variations ==
On most modern double-breasted coats, one column of buttons is decorative, while the other is functional. The other buttons, placed on the outside edge of the coat breast, allow the overlap to fasten reversibly. left lapel over right lapel.

L_DOC618 3

A double-breasted garment is a coat.
Q86_L_DOC618_0_0 4

A double-breasted garment is a jacket.
Q86_L_DOC618_0_1 5

A double-breasted garment is a waistcoat.
Q86_L_DOC618_0_2 6

A double-breasted garment is a dress.
Q86_L_DOC618_0_3 7

A double-breasted garment has wide, overlapping front flaps.
Q86_L_DOC618_0_4 8

A double-breasted garment has two symmetrical columns of buttons.
Q86_L_DOC618_0_5 9

Add

Title: Hoodie [https://en.wikipedia.org/wiki/Hoodie]

A hoodie (in some cases spelled hoody and alternatively known as a hooded sweatshirt) is a sweatshirt with a hood. Hoodies' history can be traced back to the era of Medieval Europe when monks used to wear robes with a hood called a cowl, and outdoor workers wore hooded capes. Hoodies with zippers usually include two pockets on the lower front, one on either side of the zipper, while "pullover" hoodies (without zippers) often include a single large muff or pocket in the same location. Both styles (usually) include a drawstring to adjust the hood opening. When worn up, the hood covers most of the head and neck and sometimes the face.

L_DOC623 0

A hoodie is a sweatshirt with a hood.
Q86_L_DOC623_0_0 q

Hoodies with zippers usually include two pockets on the lower front.
Q86_L_DOC623_0_4 w

Hoodies without zippers usually include a single large muff or pocket in the same location.
Q86_L_DOC623_0_5 e

Both styles (usually) include a drawstring to adjust the hood opening.
Q86_L_DOC623_1_1 t

The drawstring is used to adjust the hood opening.
Q86_L_DOC623_1_1 t

When worn up, the hood covers most of the head and neck and sometimes the face.
Q86_L_DOC623_2_0 a

The hood covers most of the head and neck when worn up.
Q86_L_DOC623_2_1 s

The hood sometimes covers part of the face when worn up.
Q86_L_DOC623_2_1 s

Add

Details of Annotation:

Check all box that corresponds to your annotation.

Fact Negation^d
 Fact Modification^f
 Fact Addition^g

Figure 10: An illustration of the interface to modify context. The question, answer, input context, and corresponding gold atoms are given to the annotators and annotators should modify well-known information by revising gold atomic facts and input contexts. Annotators are also asked to check which type of modification they did.

much smaller number of datasets compared to MS Marco, it showed lower human correlation (76.4), we used the released pretrained model as M_{eval} .

Table 5: Examples of Distractor Contexts.

Question	Gold Context	Distractor Context
What is a common factor of Sepsis and Hypotension?	<p>Title: Sepsis Context: Sepsis (septicaemia in British English), or blood poisoning, is a life-threatening condition that arises when the body's response to infection causes injury to its own tissues and organs. This initial stage of sepsis is followed by suppression of the immune system. Common signs and symptoms include fever, increased heart rate, increased breathing rate, and confusion. There may also be symptoms related to a specific infection, such as a cough with pneumonia, or painful urination with a kidney infection.</p> <p>Title: Hypotension Context: Hypotension is low blood pressure. Blood pressure is the force of blood pushing against the walls of the arteries as the heart pumps out blood. Blood pressure is indicated by two numbers, the systolic blood pressure (the top number) and the diastolic blood pressure (the bottom number), which are the maximum and minimum blood pressures, respectively.</p>	<p>#Top1 Title: Gunshot wound Context: Long-term complications can include bowel obstruction, failure to thrive, neurogenic bladder and paralysis, recurrent cardiorespiratory distress and pneumothorax, hypoxic brain injury leading to early dementia, amputations, chronic pain and pain with light touch (hyperesthesia), deep venous thrombosis with pulmonary embolus, limb swelling and debility, lead poisoning, and post-traumatic stress disorder (PTSD). Factors that determine rates of gun violence vary by country. These factors may include the illegal drug trade, easy access to firearms, substance misuse including alcohol, mental health problems, firearm laws, social attitudes, economic differences and occupations such as being a police officer. Where guns are more common, altercations more often end in death. Before management begins it should be verified the area is safe.</p> <p>#Top2 Title: Medical glove Context: Medical gloves are recommended to be worn for two main reasons: To reduce the risk of contamination of health-care workers hands with blood and other body fluids. To reduce the risk of germ dissemination to the environment and of transmission from the health-care worker to the patient and vice versa, as well as from one patient to another. == History == Caroline Hampton became the chief nurse of the operating room when Johns Hopkins Hospital opened in 1889.</p> <p style="text-align: center;">⋮</p>
What was the initial name of .223 Remington?	<p>Title: .223 Remington Context: This cartridge is loaded with DuPont IMR4475 powder. During parallel testing of the T44E4 (future M14) and the ArmaLite AR-15 in 1958, the T44E4 experienced 16 failures per 1,000 rounds fired compared to 6.1 for the ArmaLite AR-15. Because of several different .222 caliber cartridges that were being developed for the SCHV project, the .222 Special was renamed .223 Remington. In May 1959, a report was produced stating that five- to seven-man squads armed with ArmaLite AR-15 rifles have a higher hit probability than 11-man squads armed with the M-14 rifle.</p>	<p>#Top1 Title: .35 Remington Context: The .35 Remington (9.1 x 49 mm) is the only remaining cartridge from Remington's lineup of medium-power rimless cartridges still in commercial production. Introduced in 1906, it was originally chambered for the Remington Model 8 semi-automatic rifle in 1908. It is also known as 9 x 49 mm Browning and 9 mm Don Gonzalo. == History == Over the years, the .35 Remington has been chambered in a variety of rifles by most firearms manufacturers, and continues in popularity today in the Marlin Model 336 lever-action and Henry Side Gate Lever Action.</p> <p>#Top2 Title: Squad automatic weapon Context: During its long service in the US military, it was pivotal in the evolution of U.S. fireteam tactics and doctrine that continues to the present day. Modern squad automatic weapons (such as the RPK and L86) are modified assault rifles or battle rifles (e.g. FN FAL 50.41 and M14A1) that may have increased ammunition capacity and heavier barrels to withstand continued fire and will almost always have a bipod. In the case of some assault rifles, such as the H&K G36 or Steyr AUG, the SAW is simply the standard rifle with a few parts replaced.</p> <p style="text-align: center;">⋮</p>

For Task 1-4, check the box if information in the sentence is in the context (gray area).

[Task1]

The Organisation of the Petroleum Exporting Countries (OPEC, OH-pek) is an organisation enabling the co-operation of leading oil-producing countries in order to collectively influence the global oil market and to maximise profit. Founded on 14 September 1960 in Baghdad by the first five members (Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela), it has, since 1965, had its headquarters in Vienna, Austria (although Austria is not an OPEC member state). As of September 2018, the 13 member countries accounted for an estimated 44 percent of global oil production and held 81.5 percent of the world's proven oil reserves, giving OPEC a major influence on global oil prices that were previously determined by the so-called "Seven Sisters" grouping of multinational oil-companies.

- Kuwait is an oil-producing country.^[1]
- Saudi Arabia is an oil-producing country.^[2]
- Iran is an oil-producing country.^[3]
- Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela^[4]
- Iraq is an oil-producing country.^[5]
- Venezuela is an oil-producing country.^[6]

[Task2]

Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela

- The first five members of OPEC were Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela.^[7]

[Task3]

The first five members of OPEC were Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela.

- Kuwait is an oil-producing country.^[8]
- Saudi Arabia is an oil-producing country.^[9]
- Iran is an oil-producing country.^[0]
- Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela^[a]
- Iraq is an oil-producing country.^[w]
- Venezuela is an oil-producing country.^[e]

[Task4]

Kuwait is an oil-producing country.
Saudi Arabia is an oil-producing country.
Iran is an oil-producing country.
Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela
Iraq is an oil-producing country.
Venezuela is an oil-producing country.

- The first five members of OPEC were Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela.^[l]

Figure 11: An illustration of the human evaluation to calculate the correlation with M_{eval} . Task 1 and Task 2 are to evaluate correlation with GR_{loose} , which is to check whether the given atomic fact is in the paragraph, and Task 3 and Task4 are to evaluate correlation with GR_{strict} , which is to compare between the atomic facts.

C Inference

C.1 Input Format

Figure 13 shows the input format we used to generate all responses. Please note that for TULU, we changed the input format to match the format during training. “<|user|> instruction <|assistant|>”

C.2 Inference Configuration

In our research, we standardize the maximum input and output lengths at 2048 tokens for all experiments, except for those examining the effect of

context length, where the maximum is extended to 4096 tokens. To ensure consistency across various model architectures, we apply 4-bit quantization during all experimental procedures. We keep the generation configuration as same as the default configurations provided by Huggingface (Wolf et al., 2019). Specifically, for the Falcon, Llama2, and Vicuna models, we implement top-k sampling with a k value of 10. For the TULU model, we set the sampling temperature to 0.6.

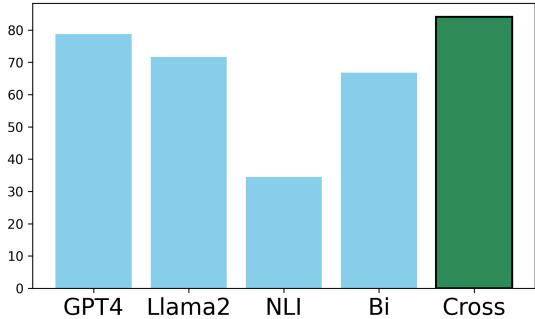


Figure 12: Correlation between Human and five models (M_{eval}) on predicting whether the knowledge of atomic facts are in a paragraph

D Results

D.1 Grounding performance by different query and context characteristics

Table 6 shows the performance of models in *Original-Gold*, Table 7 (Figure 14) shows the performance in *Revised-Gold*, Table 8 shows the performance in *Original-Dist*, and Table 9 shows the performance in *Revised-Dist*. All dataset setting shows a similar trend with *Original-Gold*. Vicuna-13b shows the highest performance over all open-sourced dataset. Grounding performance of pop high shows lower performance over pop low as models tend to utilize knowledge from given context more when it is not familiar with the knowledge (Mallen et al., 2022). Queries with single context (Single) show high grounding performance over queries that needs multiple context (Multi) since it is much easier and shorter; queries in Multi set often needs reasoning ability.

D.2 Precision and Recall

Figure 15 presents the precision and recall metrics for the *Original-Gold* dataset, whereas Figure 16 displays the same for the *Revised-Gold* dataset. Precision is measured to determine if the source of atomic facts in the knowledge base is the input context rather than external sources. Recall, on the other hand, assesses whether all essential knowledge (gold atomic facts) is included in the generated response. From the results for both datasets, it is evident that recall outperforms precision, suggesting that the model tends to incorporate knowledge beyond the provided information when evaluating them in a fine-grained manner.

D.3 Larger models Tend to Show Higher Degradation with Distractor Contexts

Figure 17 demonstrates that larger models tend to show higher degradation when distractor contexts are added. The most significant reduction is observed in recall rather than precision (Appendix D.2), suggesting that the models often default to providing only the answer without detailed explanations. The lower grounding performance for these queries is largely due to this tendency to omit specific details. Conversely, for queries requiring multiple contexts (multi), a different pattern emerges: smaller models exhibit more significant performance drops. These multi-context queries are inherently more complex, often necessitating advanced reasoning or a deeper understanding of the overall context, leading to a steeper decline in grounding performance for smaller models as the task difficulty increases.

D.4 Average Number of Contexts for Distractor Settings

In our datasets, *Original-Gold* and *Revised-Gold*, the contexts exhibit an average token length of 335, which is comparatively brief. To address this, we incorporate distractor contexts into our analysis. These distractors are contextually relevant to the queries but do not contain the gold atomic facts. As illustrated in Figure 6, the average number of contexts per query is 3.3, 11.1, 19.1, and 24.0. These values correspond to the circle markers shown in the figure, indicating a varied context distribution in our dataset.

D.5 Performance on Answer Accuracy

Table 10 shows the answer accuracy of models across five settings. Diving into performance based on input context and question traits reveals key patterns. Without external contexts, high-popularity questions achieve a 32.6% accuracy, outpacing low-popularity ones at 26.8%. However, this changes with gold contexts: low-popularity questions slightly edge out at 83.4% over the 83.2% for high-popularity ones. This likely stems from models leaning more on given contexts when unsure, mirroring Mallen et al. (2022) findings. Regarding the number of input contexts, queries requiring multiple contexts generally fare worse than those with one. The gap is wider for smaller models (under 40b parameters): they experience a 23.7% drop, while larger models see only a 13.1% dip.

Input Format for Evaluation

Generate an [answer] to the given [question] in full sentence by utilizing all necessary information in given [context] and limiting the utilized information to that [context]. Provide all information you utilize from given [context] to answer the question.

[context]

Title: Greece [<https://en.wikipedia.org/wiki/Greece>]

The country's rich historical legacy is reflected in part by its 18 UNESCO World Heritage Sites. Greece is a unitary parliamentary republic, and a developed country, with an advanced high-income economy. Its economy is the second largest in the Balkans, where it is an important regional investor. A founding member of the United Nations, Greece was the tenth member to join the European Communities (precursor to the European Union) and has been part of the Eurozone since 2001.

Title: Germany [<https://en.wikipedia.org/wiki/Germany>]

After the fall of communist led-government in East Germany, German reunification saw the former East German states join the Federal Republic of Germany on 3 October 1990—becoming a federal parliamentary republic. Germany has been described as a great power with a strong economy; it has the largest economy in Europe, the world's fourth-largest economy by nominal GDP and the fifth-largest by PPP. As a global power in industrial, scientific and technological sectors, it is both the world's third-largest exporter and importer.

[question]

Compare the economic rank of Germany and Greece.

Don't Forget that you have to generate an [answer] to the given [question] in full sentence by utilizing all necessary information in given [context] and information only from the [context]. Also, provide all information you utilize from given [context]

[answer]

Figure 13: Input format to generate response

Size M_{pred}	7B		13B		30B		40B		65B	UNK		
	Vicuna	TULU	Llama2	Llama2-chat	Vicuna	TULU	TULU	Falcon	Falcon-I	TULU	GPT-3.5	GPT-3.5-I
Avg # of Atomic	10.95	6.53	31.49	8.46	7.23	5.42	4.46	26.03	10.42	4.77	7.39	10.4
Grounding	50.01	46.67	26.09	55.91	61.44	43.42	45.06	18.92	42.35	43.58	61.01	65.69
Pop High	45.31	46.32	23.21	54.57	59.85	41.15	45.19	18.16	38.36	41.58	60.06	63.23
Pop Low	55.39	47.08	29.38	57.44	63.25	46.01	44.92	19.8	46.91	45.86	62.11	68.5
Multi	40.4	34.54	16.31	43.54	48.27	30.35	36.07	15.34	31.03	31.27	47.68	52.31
Single	51.94	49.1	28.05	58.38	64.07	46.03	46.86	19.64	44.61	46.04	63.68	68.36
Definite	58.5	43.15	28.67	64.65	68.96	36.46	39.09	18.13	50.13	35.52	67.85	73.11
Free-Form	39.99	50.84	23.04	45.58	52.55	51.63	52.12	19.86	33.15	53.11	52.94	56.92

Table 6: Specific performance of *Original-Gold*. Best from all models in **Bold** and best from open-sourced models in underline.

This underscores bigger models' superior multi-context comprehension and reasoning capacity. We

believe this discrepancy highlights a larger model's enhanced reasoning capacity and its ability to bet-

Size	7B				13B				30B		40B		65B		UNK	
	M_{pred}	Vicuna	TULU	Llama2	Llama2-chat	Vicuna	TULU	TULU	Falcon	Falcon-I	TULU	GPT-3.5	GPT-3.5-I			
Grounding	47.98	46.52	25.22	53.41	<u>57.5</u>	41.35	41.95	14.59	40.1	31.88	59.04	60.25				
Pop High	46.08	46.75	25.75	51.59	<u>55.43</u>	39.78	43	12.13	36.67	32.03	56.43	57.52				
Pop Low	50.14	46.26	24.62	55.48	<u>59.86</u>	43.14	40.75	17.4	44.02	31.7	62.03	63.36				
Multi	37.6	33.37	19.77	40.44	<u>46.47</u>	29.71	34.14	9.52	32.2	20.97	45.22	48.75				
Single	50.05	49.15	26.31	56.00	<u>59.70</u>	43.68	43.51	15.6	41.68	34.06	61.81	62.54				
Definite	55.85	45.04	26.19	60.06	<u>64.00</u>	37.46	37.43	14.21	47.44	22.57	65.07	67.56				
Free-Form	38.67	48.27	24.08	45.54	<u>49.82</u>	45.95	47.29	15.02	31.42	42.89	51.93	51.6				

Table 7: Specific performance of *Revised-Gold*. Best from all models in **Bold** and best from open-sourced models in underline.

Size	7B				13B				30B		40B		65B		UNK	
	M_{pred}	Vicuna	TULU	Llama2	Llama2-chat	Vicuna	TULU	TULU	Falcon	Falcon-I	TULU	GPT-3.5	GPT-3.5-I			
Grounding	45.01	44.57	23.68	35.83	<u>53.05</u>	41.95	40.95	15.62	36.33	39.12	56.78	56.87				
Pop High	40.24	40.84	22.25	35.5	<u>51.14</u>	39.93	40.77	13.81	33.21	40.26	55.55	55.67				
Pop Low	50.45	48.82	25.31	36.21	<u>55.23</u>	44.25	41.16	17.68	39.9	37.82	58.16	58.24				
Multi	33.60	26.20	14.44	28.88	<u>38.34</u>	25.91	29.81	14.12	27.07	29.5	40.72	41.41				
Single	47.29	48.23	25.53	37.23	<u>55.99</u>	45.14	43.18	15.92	38.18	41.04	59.99	59.96				
Definite	50.44	39.7	25.35	35.84	<u>52.86</u>	38.64	33.67	14.7	41.79	31.03	64.4	65.05				
Free-Form	38.58	50.34	21.71	35.83	<u>53.27</u>	45.87	49.56	16.71	29.88	48.68	47.77	47.20				

Table 8: Specific performance of *Original-Dist*. Best from all models in **Bold** and best from open-sourced models in underline.

Size	7B				13B				30B		40B		65B		UNK	
	M_{pred}	Vicuna	TULU	Llama2	Llama2-chat	Vicuna	TULU	TULU	Falcon	Falcon-I	TULU	GPT-3.5	GPT-3.5-I			
Grounding	39.76	44.39	19.30	46.45	<u>55.04</u>	38.37	40.87	12.14	32.60	30.30	56.08	54.54				
Pop High	39.18	41.20	19.17	45.09	<u>52.76</u>	37.8	39.78	10.03	28.60	36.96	52.40	53.61				
Pop Low	40.42	48.04	19.44	48.00	<u>57.65</u>	39.03	42.10	14.55	37.16	35.55	60.28	55.60				
Multi	30.90	31.92	15.64	36.09	<u>37.88</u>	25.77	32.14	8.72	23.21	20.14	43.28	44.41				
Single	41.53	46.89	20.03	48.52	<u>58.48</u>	40.88	42.61	12.82	34.47	32.33	58.64	56.56				
Definite	45.15	43.44	18.38	51.10	<u>62.02</u>	41.23	35.64	11.80	36.75	19.75	61.15	59.62				
Free-Form	33.39	45.51	20.38	40.95	<u>46.80</u>	35.00	47.04	12.53	27.69	42.76	50.08	48.53				

Table 9: Specific performance of *Revised-Dist*. Best from all models in **Bold** and best from open-sourced models in underline.

Size	7B				13B				30B		40B		65B		UNK	
	M_{pred}	Vicuna	TULU	Llama2	Llama2-chat	Vicuna	TULU	TULU	Falcon	Falcon-I	TULU	GPT-3.5	GPT-3.5-I			
Without Contexts	16.40	14.81	28.91	<u>35.98</u>	30.40	15.67	28.90	33.91	31.85	22.49	47.11	45.55				
Original-Gold	83.06	77.83	81.56	84.79	<u>86.57</u>	82.62	83.74	70.19	82.38	83.38	88.16	91.31				
Original-Dist	70.88	70.83	72.85	80.26	<u>81.50</u>	77.27	77.33	63.2	70.26	79.51	87.00	88.01				
Revised-Gold	76.19	76.94	77.26	<u>81.36</u>	80.90	76.64	76.82	58.84	71.49	78.29	86.13	84.79				
Revised-Dist	66.91	64.67	57.88	55.51	<u>73.49</u>	69.91	71.75	55.51	60.10	70.97	79.95	83.32				

Table 10: Answer Accuracy of twelve different models. For each setting, the best in **bold** and the best of open-sourced models in underline.

ter understand multiple contexts. Lastly, revising or adding distractors to contexts affects accuracy. It declines notably with both actions, with a steeper 12.4% fall when distractors are added to revised contexts, compared to 7.8% for original contexts.

D.6 Performance on Fluency

Our grounding assessment risks being skewed by responses that merely extract and piece together fragments of external knowledge. To counter this, we evaluate the fluency of the generated responses to determine whether they are formulated in a natu-

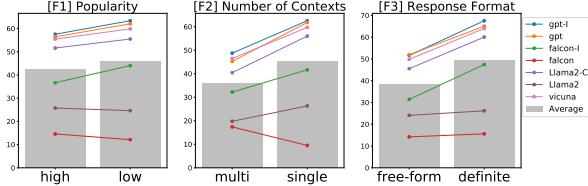


Figure 14: Details of Grounding performance by the characteristics of queries and contexts in *Revised-Gold*. _I indicates instruction tuned version and _C is those with RLHF tuned. Llama2 and vicuna is 13B, falcon is 40B model.

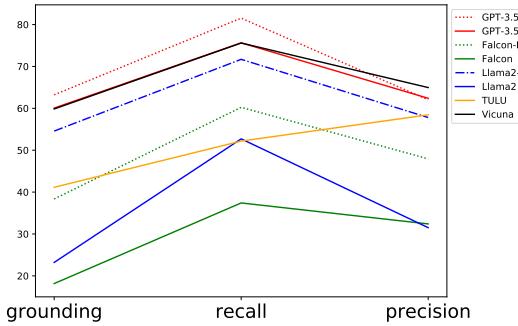


Figure 15: Performance of grounding performance, precision, and recall in *Original-Gold*

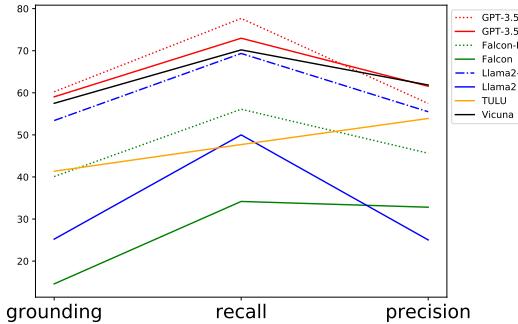


Figure 16: Performance of grounding performance, precision, and recall in *Revised-Gold*

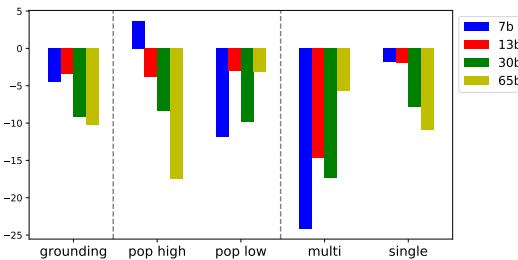


Figure 17: Reduction rate in grounding performance when adding distractor contexts

rally coherent manner. We employ G-EVAL (Liu et al., 2023c) to evaluate fluency, a framework that uses large language models in a chain-of-thought and form-filling paradigm. This fluency metric is

13B				30B	40B
Llama	Llama-C	Vicuna	TULU	TULU	Falcon-I
3.66	4.96	4.94	4.87	4.92	4.97

Table 11: Fluency of LLMs measured by G-EVAL. Here, Llama is Llama2 and Llama-C is Llama2-Chat and Falcon-I is Falcon-Instruct.

particularly applied to queries requiring free-form answers as we observed that some models tend to produce only direct answers thus difficult to evaluate the fluency. Table 11 shows the fluency scores of six LLMs. Notably, all models demonstrate high fluency, with Llama2 exhibiting the lowest score. This is attributed to its lack of instruction tuning, leading it to generate longer, less relevant sentences reminiscent of its pretraining data. The instructions used to evaluate fluency are detailed in Figure 18.

Instructions for evaluation of fluency

You will be given one response written for a instruction.

Your task is to rate the response on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Fluency (1-5): the quality of the response upon the Input in terms of grammar, spelling, punctuation, word choice, and sentence structure. The response should not contain any unnatural symbols.

- 1: Very Poor. The response is mostly incoherent with severe issues in grammar, spelling, punctuation, word choice, sentence structure, and contains unnatural symbols.
- 2: Below Average. The response is understandable with effort; numerous errors in grammar, spelling, punctuation, word choice, and sentence structure; may have unnatural symbols.
- 3: Average. The response is understandable with occasional errors in grammar, spelling, punctuation, word choice, or sentence structure; no unnatural symbols.
- 4: Above Average. The response is mostly fluent with very few errors; clear and easy to understand; no unnatural symbols.
- 5: Excellent. The response is perfectly fluent; free from any errors; clear, concise, and natural with no unnatural symbols.

Evaluation Steps:

1. Read the given response thoroughly.
2. Check for any spelling mistakes.
3. Examine the grammar and sentence structure. Look for incorrect verb conjugations, misplaced modifiers, and other grammatical mistakes.
4. Ensure that punctuation is used correctly. Check for missing or misused commas, periods, semicolons, etc.
5. Evaluate the word choice. Are the words appropriate for the context? Are there any words that sound unnatural or out of place?
6. Confirm that there are no unnatural symbols or characters in the response.
7. Based on the observations, rate the fluency of the response using the provided scale (1-5).

Example:

Response:

{response}

Evaluation Form (scores ONLY):

Fluency (1-5):

Figure 18: Instructions for Evaluation of Fluency