How Well Do Large Language Models Truly Ground?

Anonymous ACL submission

Abstract

To reduce issues like hallucinations and lack of control in Large Language Models (LLMs), a common method is to generate responses by grounding on external contexts given as input, known as knowledge-augmented models. However, previous research often narrowly defines "grounding" as just having the correct answer, which does not ensure the reliability of the entire response. To overcome this, we propose a stricter definition of grounding: a model is *truly* grounded if it (1) fully utilizes the necessary knowledge from the provided context, and (2) stays within the limits of that knowledge. We introduce a new dataset and a grounding metric to evaluate model capability under the definition. We perform experiments across 25 LLMs of different sizes and training methods and provide insights into 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[1].

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 systems [2] 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 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.

[1] We will make our code and dataset publicly available.

[2] https://www.bing.com/new, https://www.perplexity.ai/, https://openai.com/blog/chatgpt-plugins

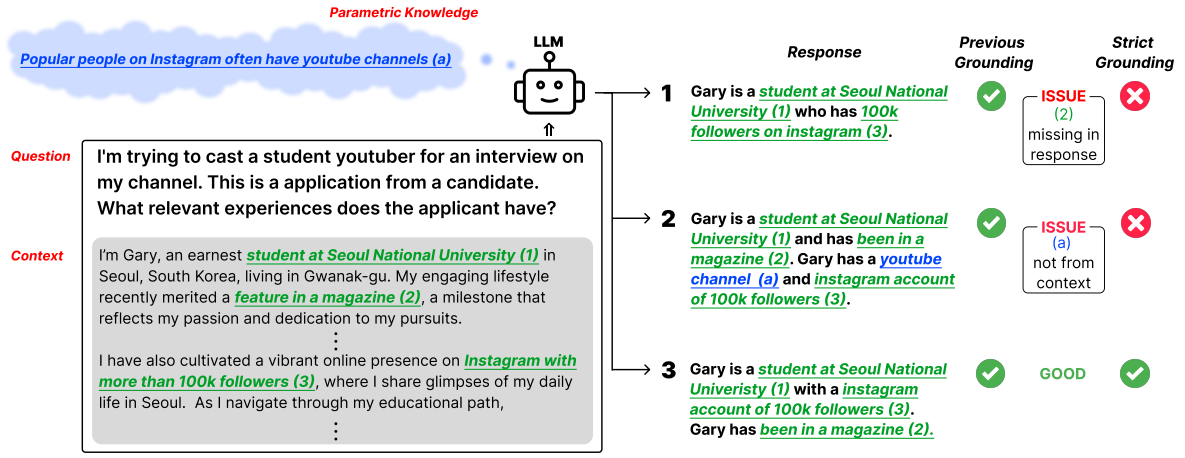


Figure 1: An example scenario of a company's HR team using LLM to question upon candidate's resume which is given as input context. The previous definition of grounding would consider responses 1 and 2 as well grounded due to their high relevancy with the question and input context. However, as our definition 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.

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 information[3]. 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 25 LLMs of different sizes and training methods to explore which model attributes significantly contribute to grounding ability and identify some important factors.

- Training methods like Instruction Tuning or RLHF have a more pronounced impact on grounding performance than model size.
- High answer accuracy, commonly used to assess how well a model incorporates context in previous works, does not ensure high grounding performance.
- Instruction-tuned models show high degradation when additional relevant contexts are added as input.
- When given multiple contexts, performance degradation is more influenced by how distracting these contexts are, rather than by their length.

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

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, are inadequate to measure the model's grounding ability under our definition; they lack annotation of which knowledge from the external context is necessary (gold) to answer the query and are hard to verify the source of knowledge in generated response (whether it is from a given context or model parameter). Furthermore, since most datasets were created before the emergence of modern LLMs, they're unsuitable for understanding the diverse characteristics of these models. Therefore, to evaluate a model's grounding ability under our defined criteria, we created a new dataset.

Generating Response with External Knowledge Recent research efforts have focused on incorpo-

rating 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 associating 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); information retrieval task. In this work, we focus on bridging the definition with when input context is the knowledge source.

3 Grounding

In this paper, we define that the model grounds well more strictly and share 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 a model is well-grounded when it generates responses relevant to the query while utilizing the given contexts. When given a set of external contexts $\mathcal{C}$, a set of answers $\mathcal{A}$, and generated response P, the previous definition often defines it well-grounded 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 that it cannot capture whether the generated response misses relevant knowledge from a given context or whether it hallucinates. 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 *truly* grounds on provided external context when (1) it utilizes *all* necessary knowledge in the context, and (2) it does *not* incorporate other knowledge apart from the contexts, such as that stored in the model parameters. Here, we see the "atomic facts" (short sentences conveying one piece of information) as the knowledge unit. As a sentence contains multiple knowledge, we disassemble[4] 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 set of external contexts $\mathcal{C}$ 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. $\forall k \in \mathcal{P}_A, \exists c \in \mathcal{C}$ such that $k \in c$

Models that demonstrate strong grounding capabilities as per our definition are highly valued in various use cases. It can be used in developing personalized chatbot services. By grounding contexts

[4]Following 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.

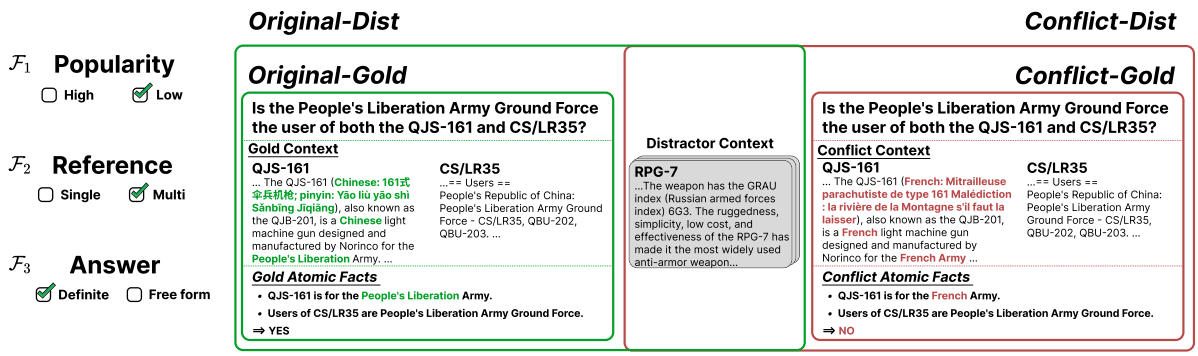


Figure 2: Four versions of our dataset: *Original-Gold*, *Original-Dist*, *Conflict-Gold*, and *Conflict-Dist*. Conflict-* contains modified gold contexts (conflict context) by human annotators. *-Dist differs from *-Gold in that it contains distractor contexts. The left part of the figure shows three key factors we considered when constructing our dataset.

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. Also, when a company 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. Moreover, models with a strong grounding ability 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

We construct a new evaluation dataset specifically designed to measure a model's grounding ability due to limitations of existing datasets; they lack annotation of which knowledge from the provided context is necessary, hard to verify the source of knowledge (whether the knowledge is from a given context or its parameter), and most do not consider key variables known to influence LLM performance as they were constructed before the advent of modern LLM.

As in Figure 2, our dataset comprises four versions: *Original-Gold*, *Original-Dist*, *Conflict-Gold*, and *Conflict-Dist*. The differentiation lies in two main aspects: (1) The nature of the input context, which is either an unaltered Wikipedia content (*Original-**) or a modified, conflicting version (*Conflict-**) to determine whether the model's response is from its internal knowledge or by grounding on external knowledge. (2) The inclusion of distractor contexts: * *-Gold* versions contain only "gold contexts" that directly answer the query, whereas **-Dist* versions also include distractor contexts, which are relevant but not gold.

Furthermore, we integrate three key factors (left of Figure 2) known to bring qualitative differences in model responses for a more comprehensive analysis: [$\mathcal{F}_1$] 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; Cífka and Liutkus, 2022), and [$\mathcal{F}_3$] Required response format (definite answer or free-form answer) (McCoy et al., 2021; Tuckute et al., 2022).

Our dataset construction is mainly divided into five steps. Details of data construction including human annotators, inter-labeler agreement, data distribution of the factors, data examples, and more are in Appendix A.

Step 1: Context Selection In our first step, we select sets of input contexts ($\mathcal{C}$) considering $\mathcal{F}1$ and $\mathcal{F}_2$. 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 utilize document pageviews, and for $\mathcal{F}_2$, we construct a document set sampled from the intersection between the popularity list and the hyperlinked document.

Step 2: Instance Generation & Classification Based on the document sets from Step 1, we use GPT-3.5[5] to generate 10 candidate pairs of question and answer. We classify the candidate pairs by $\mathcal{F}_2$ and $\mathcal{F}_3$, and select a single query with the highest quality from each class. Note that the generated answer was replaced by the annotators.

Step 3: Gold Atomic Fact Selection To evaluate grounding performance, we decompose context sets $C \in \mathcal{C}$ into atomic facts $\{C_{A_1}, \cdots, C_{A_k}\}$.

[5] gpt-3.5-turbo-0301

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 ($\{C_{G_1}, \cdots, C_{G_m}\} \subseteq \{C_{A_1}, \cdots, C_{A_k}\}$). We now get 480 complete instances that we call *Original-Gold* $(Q, A, \mathcal{C}, \mathcal{C}_G)$.

Step 4: Modify Context Given an instance from *Original-Gold*, annotators are instructed to revise well-known and key knowledge to answer the question in the input context. This step results in *Conflict-Gold* $(Q, A', \mathcal{C}', \mathcal{C}'_G)$, a modified, conflicting version.

Step 5: Add Distractor Contexts To analyze the impact when additional knowledge apart from the gold ones is added to the input context, we sample distractor contexts, contexts with high similarity but not directly related to an answer, with contriever (Izacard et al., 2022), a dense retriever pretrained through contrastive learning, and include them in the input context (*Original-Dist* when added to original gold contexts and *Revised-Dist* when added to revised gold contexts).

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: GPT-4 (OpenAI, 2023), Llama-2-70b-chat (Touvron et al., 2023), TRUE (T5-11B finetuned on various NLI datasets) (Honovich et al., 2022), bi-encoder model (MiniLM finetuned on 1B training pairs), and cross-encoder model (MiniLM finetuned on MSMARCO) (Wang et al., 2020). Surprisingly, the cross-encoder model[6] shows the highest correlation with human (84.1), outperforming GPT-4 (78.7). It also closely matches the correlation between humans (88.6) Thereby, we utilize the cross-encoder model as M_{eval}.

We define grounding performance as the **F1 score** of precision and recall calculated as: precision $= \sum_{i=1}^{k} M_{eval}(P_{A_i}, C)$ and recall $= \sum_{i=1}^{m} M_{eval}(C_{G_i}, P)$ where $M_{eval}(a, B)$ returns 1 when knowledge of a exists in B and 0 elsewise. 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 the model's grounding ability in previous works (Mallen et al., 2022; Borgeaud et al., 2021); it measures if the answer is present within the generated response[7].

4 Experiments

We experiment with 25 LLMs of various sizes and training methods (Instruction-tuning, RLHF, DPO). From the results, we share interesting findings of how different factors of LLMs and different characteristics of input context lead to their grounding ability. Section 4.1 shows brief details of the models we evaluate. Section 4.2 shows how different factors of LLMs lead to their grounding ability and interesting findings. Details of the input format, generation configurations, and others are in Appendix C.

4.1 Models

We experiment with two proprietary LLMs: GPT-3.5 (GPT) and GPT-3.5-instruct (GPT-I)[8]. The latter, GPT-instruct[9], is a further finetuned version of GPT, primarily for following instructions. Table 2 shows details of open-sourced LLMs we experiment over: Llama2 (Touvron et al., 2023), Llama2-chat (Llama2-C), Vicuna, TÜLU1 (Wang et al., 2023), TÜLU2 (Ivison et al., 2023), TÜLU2 with DPO (TÜLU2-D), Mistral-Instruct (Mistral-I) (Jiang et al., 2023), Zephyr (Tunstall et al., 2023), Falcon (Penedo et al., 2023), and Falcon-Instruct (Falcon-I). All checkpoints are provided from huggingface (Wolf et al., 2019).

4.2 Results

Overall performance Table 1 shows the overall grounding performance of various models over four different dataset versions[10]. Due to limited space, the results of all models in four dataset versions are in Appendix D.2. GPT-I shows the highest performance for original datasets (*Original-Gold* and

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

[7] We only measure the metric to queries with definite answers.

[8] Specific 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

[9] After this point, we shorten GPT-3.5 to "GPT"

[10] Details of each dataset scenarios in Section 3.2

Size	7B					13B			40B	70B		UNK	
M_{pred}	Llama2-C	Vicuna	TÜLU2	Mistral-I	Zephyr	Llama2-C	Vicuna	TÜLU2	Falcon-I	Llama2-C	TÜLU2	GPT	GPT-I
Original-Gold	51.6	50.0	58.6	60.3	54.7	55.9	61.4	61.9	42.4	56.9	61.9	61.0	**65.7**
Original-Dist	45.1	45.0	54.9	54.9	53.7	35.8	56.5	55.3	36.3	55.8	56.7	56.8	**56.9**
Conflict-Gold	46.0	48.0	54.9	59.8	52.4	53.4	57.5	57.7	40.1	56.3	**62.4**	59.0	60.3
Conflict-Dist	40.4	39.8	47.9	54.3	52.4	46.5	55.0	50.4	32.6	54.4	54.9	**56.1**	54.5

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.

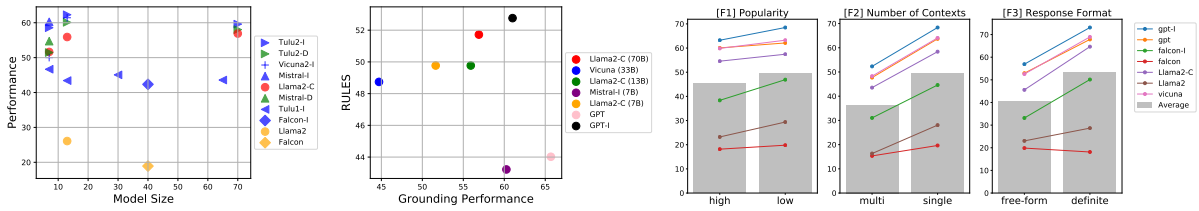
	Base	DPO	RLHF	Inst.	Size
Llama2	Llama2	x	x	x	[13]
Llama2-C	Llama2	x	o	o	[7, 13, 70]
Vicuna	Llama2	x	x	o	[7, 13, 33]
TÜLU1	Llama1	x	x	o	[7, 13, 30, 65]
TÜLU2	Llama2	x	x	o	[7, 13, 70]
TÜLU2-D	Llama2	o	x	o	[7, 13, 70]
Falcon	Falcon	x	x	x	[40, 180]
Falcon-I	Falcon	x	x	o	[40, 180]
Mistral-I	Mistral	x	x	o	[7]
Zephyr	Mistral	o	x	o	[7]

Table 2: Abstract of open-sourced LLMs we experiment over. The size column shows various sizes of the model we experimented over. The base column shows the pretrained model each model is finetuned on. The rest of the columns show different training methods; Inst. is instruction-tuned, DPO is Direct Preference Optimization, and RLHF is Reinforcement Learning from Human Feedback.

Original-Dist), and TÜLU2-70B shows the highest performance among open-sourced models, similar performance with GPT. Performance of *Conflict-Gold* consistently shows lower performance than *Original-Gold* (average of 4.7 drops), which we hypothesize is due to conflict between parametric space and external knowledge. The performance also consistently degrades with distractor contexts added: an average of 10.7 drops for *Original-Dist* from *Original-Gold* and an average of 10.0 drops for *Conflict-Dist* from *Conflict-Gold*. The drop is higher than when given conflicting knowledge, which highlights the LLM's tendency to deviate from the primary context when presented with extraneous information and the importance of providing only the gold contexts for high grounding performance. When comparing the different model sizes of the same model (i.e., TÜLU2 and Llama-C), the grounding performance of all four dataset versions tends to steadily increase. The improvement rate by a larger model tends to be stronger as the dataset is difficult; *Conflict-Dist* is considered more difficult over *Original-Gold* as it contains more knowledge in input context and contains conflict knowledge with its parametric space. When comparing the performance of precision and recall, a common trend across all models is a superior performance in precision over recall (Appendix D.3). This suggests a challenge in utilizing all necessary knowledge when generating a response and it tends to utilize only a partial of them.

Training method shows stronger effect than model size in grounding performance Figure 3 **(a)** 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 the most important factor. To determine if grounding performance is strongly dependent on instruction-following ability, we see the correlation between grounding performance with performance on RULES benchmark (Mu et al., 2023), a benchmark to determine how well it follows the given rule. Figure 3 **(b)** shows that there is weak correlation between the two scores. This suggests that grounding performance does not appear to be strongly reliant on the capacity to adhere to instructions. We could see a similar trend with MMLU benchmark (Hendrycks et al., 2020) in Appendix D.1.

Grounding performance by different query and context characteristics Figure 3 **(c)** 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 com-



(a) (b) (c)

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 RULES performance and grounding performance. There is a weak correlation between instruction-following ability and grounding performance. **(c)** shows details of grounding performance by the characteristics of queries and contexts in *Original-Gold*. Llama2 and Vicuna are 13B, Falcon is 40B model.

plexity to extract all relevant atomic facts. Lastly, for $\mathcal{F}_3$, questions with predetermined answers (definite) tend to achieve better grounding than open-ended answers (free-form). This divergence largely stems from recall metrics as free-form instances contain more necessary knowledge (gold atomic facts) compared to definite instances, it is more difficult to find all. We could see that the trend holds for all four dataset settings in Appendix D.2.

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 12), 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.

Smaller models tend to show a higher reduction rate by DPO training Table 3 shows the degradation rate from TÜLU2 to those trained with DPO. Smaller models tend to show a higher degradation rate in grounding performance by DPO training. The degradation rate tends to come from its verbosity, aligning with the findings from Ivison et al. (2023). Moreover, the results of Zephyr, a 7B size model further trained with DPO on top of Mistral, in Table 1 show similar results; high degradation rate by DPO training.

	TÜLU2	+ DPO	**deg.rate** (%)	TÜLU2	+ DPO	**deg.rate** (%)
	Original-Gold			*Revised-Gold*		
7B	56.2	51.5	8.5	54.9	51.4	6.4
13B	62.3	60.1	3.5	61.9	58.0	6.3
70B	59.6	58.0	2.7	59.9	58.1	3.1
	Original-Dist			*Revised-Dist*		
7B	54.9	45.3	17.6	47.9	41.4	13.5
13B	55.3	54.0	2.3	50.4	54.2	-7.5
70B	53.4	55.4	-3.7	52.4	55.1	-5.1

Table 3: Grounding performance of TÜLU and those trained with DPO (+DPO). **deg.rate** column shows the degradation rate from TÜLU to those trained with DPO.

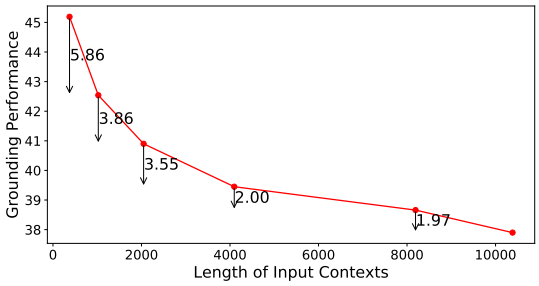


Figure 4: Grounding performance of Vicuna-13B-16k as length of input contexts increases.

Performance degradation is more influenced by the distraction level of the contexts rather than the length of distractor contexts Figure 4 illustrates that as the input context length increases, the grounding performance of Vicuna-13b-16k, capable of handling extensive inputs, varies significantly. Please note that the input contexts differ by the length of distractor contexts as the length of gold contexts is the same. Notably, grounding performance deteriorates more rapidly at the initial points (5.86 at the initial point and 1.97 at the end point of the plot). This is because we add distractor contexts in the order of those in high rank by contriever (Izacard et al., 2022), which indicates that contexts with high distraction levels are added

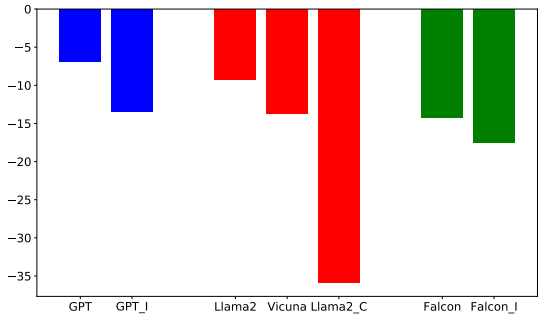


Figure 5: Reduction rate in *Original-Dist* performance from *Original-Gold*. Models with the same base model are in the same color. Models that are instruction tuned (falcon_I, GPT_I, Vicuna) or underwent RLHF (Llama2_C) show higher degradation when distractor contexts are added. Vicuna and Llama2 are 13B and Falcon is 40B model.

at the initial points, causing stronger distractions. Such a result indicates 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).

Impact of gold contexts position on grounding performance: optimal position at the end We could see that the position of gold contexts within multi-document settings significantly influences grounding performance, aligning with the findings from Liu et al. (2023a). Experiment with Vicuna-13b-16k, input context length of 4096 over *Original-dist* show the highest performance when gold contexts are positioned at the end and the lowest when positioned in the middle (end-43.37, beginning-39.32, random-39.45, middle-39.32). The trend also holds for *Conflict-dist*: end-43.53, beginning-41.28, random-39.10, middle-38.30. Such results emphasize the importance of where you put the gold contexts in a multi-document setting for high grounding performance.

Instruction-tuned models show higher degradation with distractor contexts Figure 5 demonstrates while models fine-tuned with instruction 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 12 in Appendix D.6 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.7.

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 25 LLMs across four dataset scenarios, we observed various insights. Rather than model size, various training techniques and base models tend to affect more on grounding performance. Models find it challenging to utilize all necessary knowledge when generating a response. By presenting the performance of various models on different dataset settings, we provide valuable perspectives to the ongoing discourse on enhancing LLM grounding abilities and practical guidance for choosing suitable models for applications that require generating response by *truly* grounding on a given context.

6 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 modified 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. As we modified the existing dataset, the contexts we provide may distract people.

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.

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.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Y. Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Workshop on Document-grounded Dialogue and Conversational Question Answering*.

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.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv*, abs/2311.10702.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

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.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Dan Hendrycks, and David Wagner. 2023. Can llms follow simple rules? *ArXiv*, abs/2311.04235.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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 Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb 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, FatemehSadat 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 Poulton, 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*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *ArXiv*, abs/2310.16944.

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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *ArXiv*, abs/2002.10957.

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, Pierric 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.

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

As shown in Figure 6, 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.

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[11] 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 pageviews[12] 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 document[13]. 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.

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 7.

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 4.

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 8 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] Modify 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

[11] Text in Wikipedia is co-licensed under the CC BY-SA and GFDL and is widely used in research.

[12] https://dumps.wikimedia.org/other/pageview_complete/monthly/2023/

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

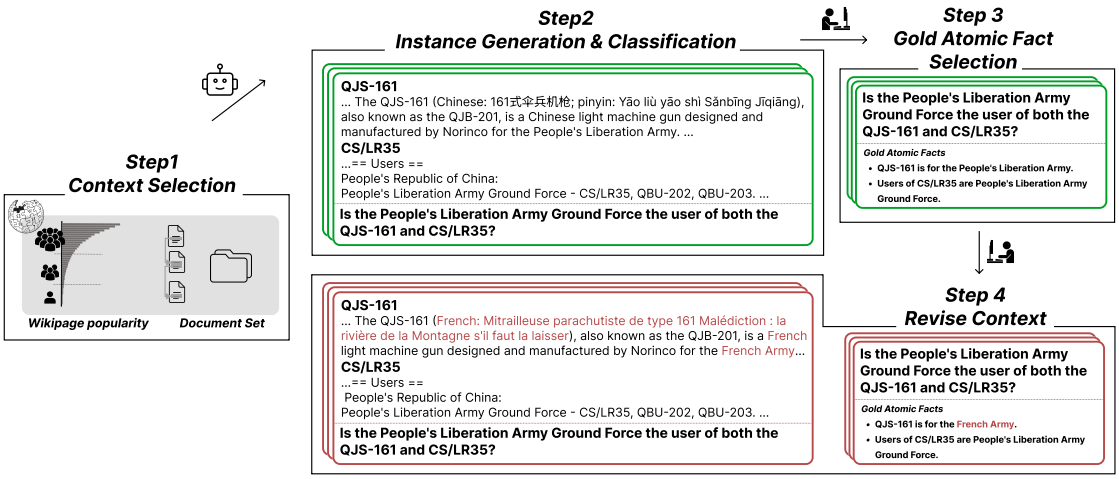


Figure 6: Data Construction Pipeline. Step 1-3 shows how we construct *Original-Gold*, and Step 4 shows how we modified the dataset, thereby constructing *Conflict-Gold*.

Table 4: 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 2: The first season was due to start in a high-profile ceremony.
	Fact 2: The high-profile ceremony was in New Delhi.

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 9 is the interface used to construct a modified version of the dataset.

A.6 Human Annotators

We recruit 4 Korean 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.

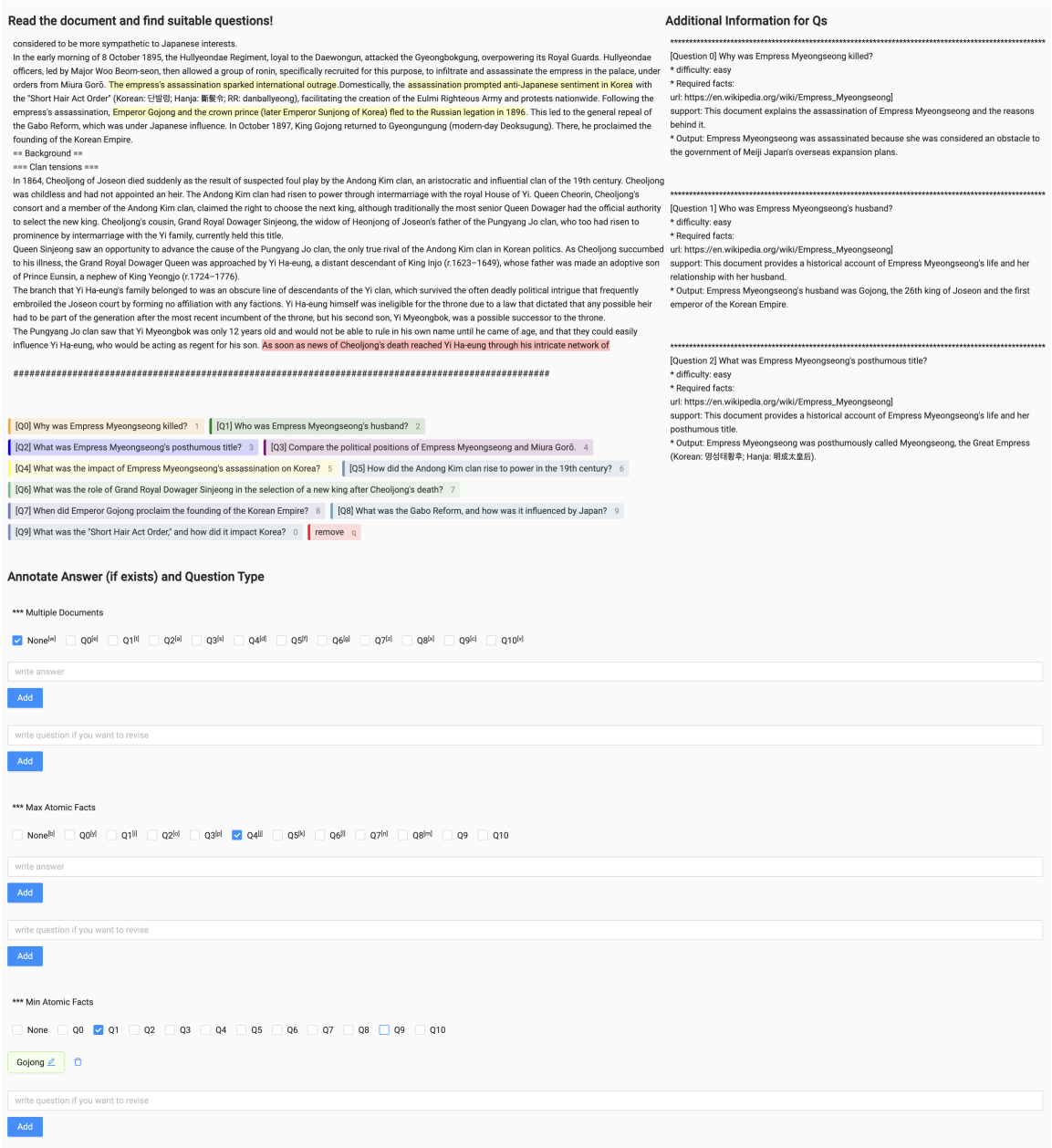


Figure 7

A.7 Data Distribution

After following the dataset construction step, we have 480 datasets (question, answer, context, gold atomic facts) along with 480 modified 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 5 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

Question	Context	Gold Atomic	Answer
Provide the claimed number of Viet Cong killed during Operation Sunset Beach.	**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. == Aftermath == Operation Sunset Beach officially concluded on 11 October, with US reports claiming that Viet Cong losses were 80 killed (body count) and a further 135 estimated killed, U.S. losses were 29 killed. == References == This article incorporates public domain material from websites or documents of the United States Army Center of Military History.	• 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.	**Holden ::** On 29 November 2016, engine production at the Fishermans Bend plant was shut down. On 20 October 2017, production of the last Holden designed Commodore ceased and the vehicle assembly plant at Elizabeth was shut down. Holden produced nearly 7.7 million vehicles.	• On 20 October 2017, production of the last Holden designed Commodore ceased.	Chevrolet
	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 5.7-litre Chevrolet-sourced Gen III V8 engine was offered. 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.	• The 5.7-litre engine was Chevrolet-sourced. • The 5.7-litre engine was a Gen III V8.	
Explain what a "dump" refers to in volleyball.	**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 Double contact or Double touch: A fault in which a player contacts the ball with two body parts consecutively D.S. : The abbreviation for "defensive specialist", a position player similar to the libero who is skilled at back row defense 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. Five-One: Six-player offensive system where a single designated setter sets regardless of court position.	• 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 5: Example of Instances

Question	Context	Gold Atomic	Answer
Provide the claimed number of Viet Cong killed during Operation Sunset Beach.	**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. == Aftermath == Operation Sunset Beach officially concluded on 11 October, with US reports claiming that Viet Cong losses were *180* killed (body count) and a further *235* estimated killed, U.S. losses were 29 killed. == References == This article incorporates public domain material from websites or documents of the United States Army Center of Military History.	• US reports claim Viet Cong losses were *180* 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.	**Holden ::** On 29 November 2016, engine production at the Fishermans Bend plant was shut down. On 20 October 2017, production of the last Holden designed Commodore ceased and the vehicle assembly plant at Elizabeth was shut down. Holden produced nearly 7.7 million vehicles.	• On 20 October 2017, production of the last Holden designed Commodore ceased.	*Audi*
	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 5.7-litre *Audi-sourced* Gen III V8 engine was offered. 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.	• The 5.7-litre engine was *Audi-sourced*. • The 5.7-litre engine was a Gen III V8.	
Explain what a "dump" refers to in volleyball.	**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 Double contact or Double touch: A fault in which a player contacts the ball with two body parts consecutively D.S. : The abbreviation for "defensive specialist", a position player similar to the libero who is skilled at back row defense 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. Five-One: Six-player offensive system where a single designated setter sets regardless of court position.	• A dump is a *final blow*. • 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 6: Example of Modified Instances

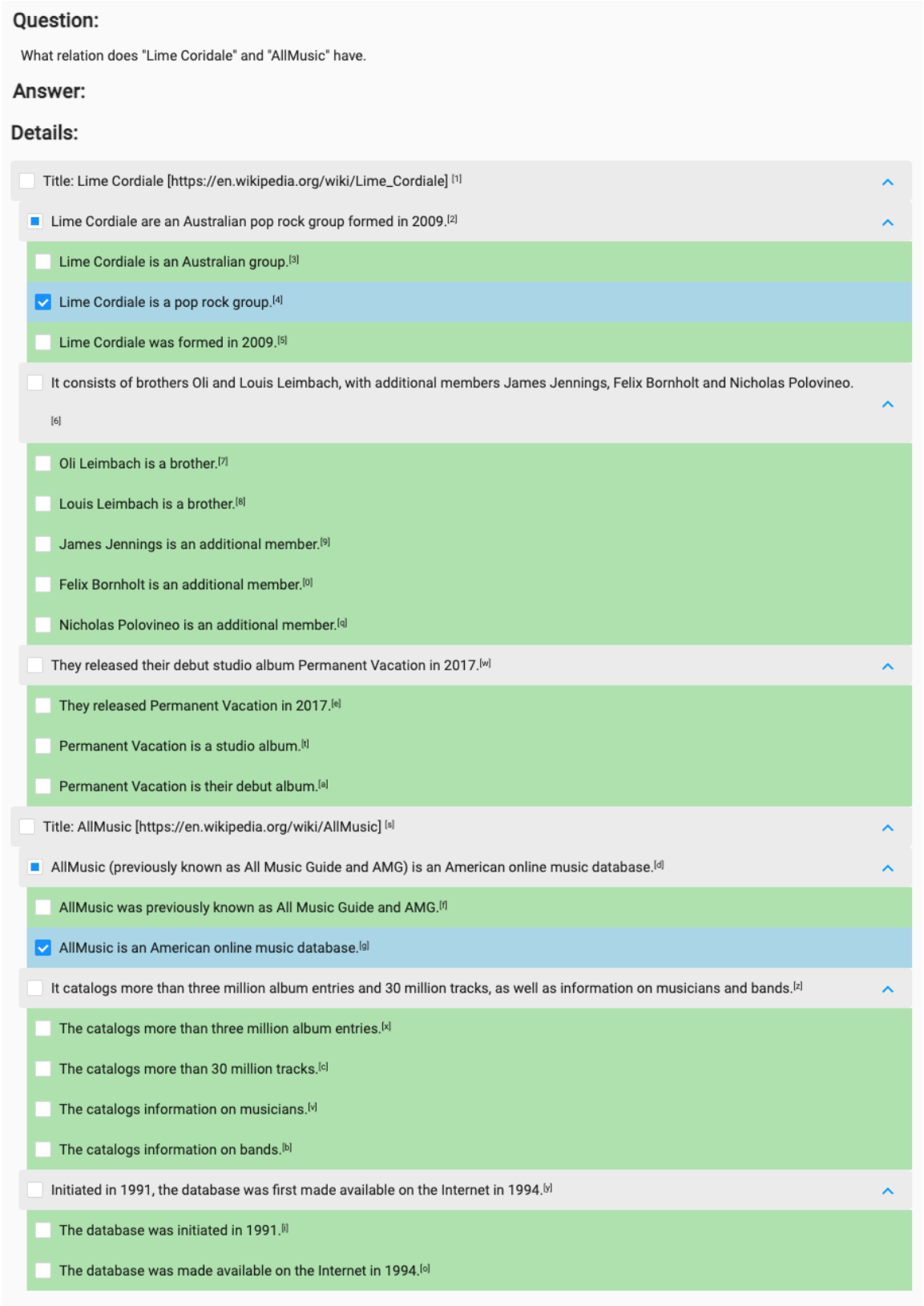


Figure 8: User interface used for gold atomic annotation

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 7.

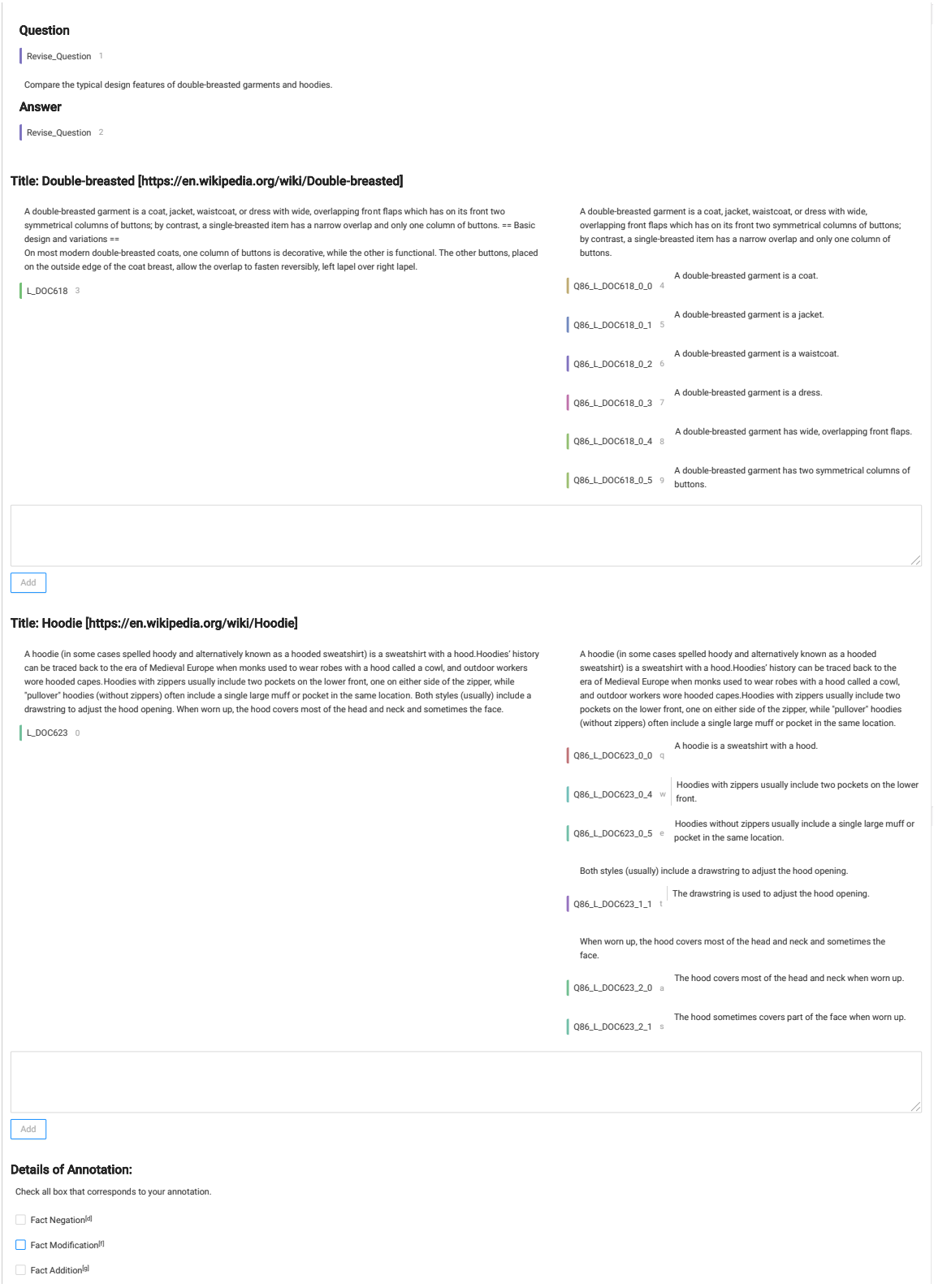


Figure 9: An illustration of the interface to modify context. The question, answer, input context, and corresponding gold atomics are given to the annotators and annotators should modify well-known information by revising gold atomic facts and input contexts. Annotators are also asked to check which type of modification they did.

B Evaluate Human Correlation for M_{eval}

As the same knowledge could be represented in various ways, we utilize a prediction model M_{eval},

Table 7: Examples of Distractor Contexts.

Question	Gold Context	Distractor Context
What is a common factor of Sepsis and Hypotension?	**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. **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.	**#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 (hyperalgesia), 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. **#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. ⋮
What was the initial name of .223 Remington?	**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.	**#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. **#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. ⋮

which predicts whether knowledge of each atomic fact is in a generated response or input context. We 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 models[14] and utilize the one with the highest correlation with human evaluation as M_{eval}. As shown in Figure 11, the cross-encoder model trained on MSMARCO dataset[15] 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}).

B.1 Human Evaluation Interface

Figure 10 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.

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 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 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}.

C Inference

C.1 Model Details

Llama2-chat is based on Llama2 and is optimized for dialogue using RLHF. Vicuna[16] is Llama2 finetuned on the outputs from ChatGPT available through ShareGPT. TÜLU1 and TÜLU2 are a Llama fine-tuned on mixture of human

[14] Details of the models are in Appendix B.

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

[16] For 7B and 13B, we used version 1.5 and for 33B, we used version 1.3, where v1.5 is tuned on top of Llama2 and v1.3 is tuned on top of Llama1

Figure 10: 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 $\text{GR}_{\text{strict}}$, which is to compare between the atomic facts.

and machine-generated instructions and responses; TÜLU1 and TÜLU2 are finetuned on top of Llama1 and Llama2, respectively. Please note that TÜLU2 is finetuned on more larger dataset compared to TÜLU1. Falcon is trained on 1,000B tokens of RefinedWeb, and Falcon-Instruct is an instruction-tuned version of Falcon. Mistral Models are selected to see the effect of instruction tuning, model size, and RLHF.

C.2 Input Format

Figure 12 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.3 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 con-

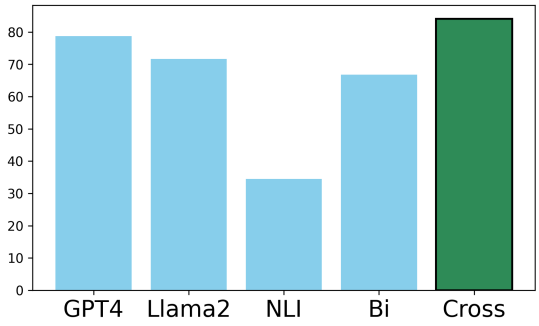


Figure 11: Correlation between Human and five models (M_{eval}) on predicting whether the knowledge of atomic facts are in a paragraph

figurations 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.

D Results

D.1 Correlation between MMLU and Grounding Performance

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 13 shows that there is a weak correlation between grounding abilities and MMLU scores[17]. This suggests that grounding performance does not appear to be strongly reliant on the capacity to adhere to instructions.

D.2 Grounding performance by different query and context characteristics

Table 8 shows the performance of models in *Original-Gold*, Table 9 (Figure 14) shows the performance in *Conflict-Gold*, Table 10 shows the performance in *Original-Dist*, and Table 11 shows the performance in *Conflict-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.

[17] pearson correlation coefficient between grounding and MMLU performance is 0.32

D.3 Precision and Recall

Figure 15 presents the precision and recall metrics for the *Original-Gold* dataset, whereas Figure 16 displays the same for the *Conflict-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.4 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.3), 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.5 Average Number of Contexts for Distractor Settings

In our datasets, *Original-Gold* and *Conflict-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

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 12: Input format to generate response

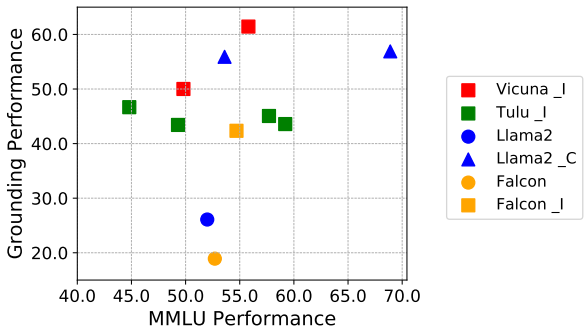


Figure 13: Correlation between MMLU performance and grounding performance: there is a weak correlation between the two.

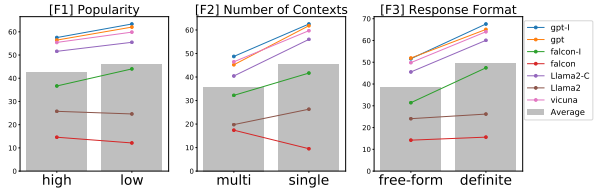


Figure 14: Details of Grounding performance by the characteristics of queries and contexts in *Conflict-Gold*. _I indicates instruction tuned version and _C is those with RLHF tuned. Llama2 and vicuna is 13B, falcon is 40B model.

illustrated in Figure 4, 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

			$\mathcal{F}_1$		$\mathcal{F}_2$		$\mathcal{F}_3$	
Model	Size	Grounding Perf.	High	Low	Free-Form	Definite	Single	Multi
Vicuna	7	50.01	45.31	55.39	39.99	58.5	51.94	40.4
	33	44.71	43.75	45.81	35.46	52.54	46.21	37.23
	13	61.44	59.85	63.25	52.55	68.96	64.07	48.27
TÜLU1	7	46.67	46.32	47.08	50.84	43.15	49.1	34.54
	13	43.42	41.15	46.01	51.63	36.46	46.03	30.35
	30	45.06	45.19	44.92	52.12	39.09	46.86	36.07
	65	43.58	41.58	45.86	53.11	35.52	46.04	31.27
TÜLU2	7	58.57	56.22	61.24	49.09	66.58	60.99	46.46
	70	59.61	57.09	62.48	53.27	64.97	62.77	43.8
	13	62.29	59.97	64.95	55.58	67.98	65.6	45.77
TÜLU2-D	7	51.46	48.24	55.15	41.14	60.2	53.01	43.75
	70	58.02	57.03	59.14	50.2	64.63	60.55	45.36
	13	60.11	57.32	63.29	50.11	68.57	62.76	46.86
Mistral-I	7	60.26	57.82	63.04	53.97	65.57	63.05	46.29
Zephyr	7	54.72	52.57	57.18	42.86	64.75	56.89	43.89
Llama2-C	7	51.63	47.81	56	38.26	62.95	53.97	39.93
	13	55.91	54.57	57.44	45.58	64.65	58.38	43.54
	70	56.9	56.53	57.32	50.53	62.29	58.62	48.32
Llama2	13	26.09	23.21	29.38	23.04	28.67	28.05	16.31
GPT	-	61.01	60.06	62.11	52.94	67.85	63.68	47.68
GPT-I	-	65.69	63.23	68.5	56.92	73.11	68.36	52.31
Falcon	40	18.92	18.16	19.8	19.86	18.13	19.64	15.34
	180	26.4	28.38	24.14	23.70	28.69	26.88	24.01
Falcon-I	40	42.35	38.36	46.91	33.15	50.13	44.61	31.03
	180	46.16	43.54	49.14	40.52	50.92	48.74	33.23

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

the figure, indicating a varied context distribution in our dataset.

D.6 Performance on Answer Accuracy

Table 12 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. 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 better 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 modified contexts, compared to 7.8% for original contexts.

D.7 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 naturally 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 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 13 shows the fluency scores of six LLMs. Notably, all models demonstrate high fluency, with Llama2 exhibiting the lowest score.

			$\mathcal{F}_1$		$\mathcal{F}_2$		$\mathcal{F}_3$	
Model	Size	Grounding Perf.	High	Low	Free-Form	Definite	Single	Multi
Vicuna	7	47.98	46.08	50.14	38.67	55.85	50.05	37.6
	13	57.5	55.43	59.86	49.82	64	59.7	46.47
	33	40.32	38.84	42.02	40	40.6	41.36	35.13
TÜLU1	7	46.52	46.75	46.26	48.27	45.04	48.05	38.87
	13	41.35	39.78	43.14	45.95	37.46	43.68	29.71
	30	43.95	45	42.75	49.29	39.43	45.51	36.14
	65	39.47	39.78	39.12	50.59	30.07	40.77	32.97
TÜLU2	7	54.86	52.22	57.88	47.41	61.16	57.4	42.19
	13	61.9	59.7	64.42	57.02	66.03	64.35	49.67
	70	59.93	57.87	62.29	53.64	65.26	61.15	53.83
TÜLU2-D	7	51.36	48.66	54.43	40.28	60.73	52.73	44.46
	13	58.03	55.82	60.55	48.34	66.22	60.03	48.01
	70	58.07	56.35	60.04	49.33	65.47	59.88	49.04
Mistral-I	7	59.83	57.32	62.69	54.39	64.43	61.92	49.38
Zephyr	7	52.37	50.34	54.69	44.03	59.42	54.36	42.4
Llama2-C	7	45.95	42.79	49.58	35.2	55.05	47.68	37.35
	13	53.41	51.59	55.48	45.54	60.06	56	40.44
	70							
Llama2	13	25.22	25.75	24.62	24.08	26.19	26.31	19.77
GPT	-	59.04	56.43	62.03	51.93	65.07	61.81	45.22
GPT-I	-	60.25	57.52	63.36	51.6	67.56	62.54	48.75
Falcon	40	23.63	22.13	25.34	24.37	23	24.47	19.42
	180	25.59	25.52	25.67	23.34	27.5	27.33	16.92
Falcon-I	40	40.1	36.67	44.02	31.42	47.44	41.68	32.2
	180	45.31	41.97	49.12	37.35	50.2	46.19	34.9

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

			$\mathcal{F}_1$		$\mathcal{F}_2$		$\mathcal{F}_3$	
Model	Size	Grounding Perf.	High	Low	Free-Form	Definite	Single	Multi
Vicuna	7	45.01	40.24	50.45	38.58	50.44	47.29	33.6
	13	57.46	55.91	59.23	49.13	64.51	59.12	49.17
TÜLU1	7	44.57	40.84	48.82	44.88	44.3	47.61	29.36
	13	41.95	38.41	46	45.24	39.17	44.9	27.21
	30	40.95	40.77	41.16	49.56	33.67	43.18	29.81
	65	39.12	40.26	37.82	48.68	31.03	41.04	29.5
TÜLU2	7	54.9	52.66	57.46	47.18	61.43	57.69	40.94
	13	55.27	52.66	58.26	52.04	58	58.12	41.05
	70	53.43	53.3	53.58	52.96	53.83	56.46	38.26
TÜLU2-D	7	45.26	42.96	47.9	36.86	52.37	46.7	38.06
	13	53.98	52.03	56.2	45.57	61.08	56.18	42.94
	70	55.41	53.61	57.47	47.9	61.76	58.24	41.27
Mistral-I	7	54.87	53.07	56.92	49.32	59.56	58.37	37.36
Zephyr	7	53.66	50.56	57.21	44.29	61.58	56.52	39.35
Llama2-C	7	45.14	43.9	46.55	37.14	51.91	47.57	32.98
	70	56.24	54.17	58.61	47.9	63.3	58.89	43.01
	13	35.83	35.5	36.21	35.83	35.84	37.23	28.88
Llama2	13	21.68	21.55	21.83	19.71	23.35	22.53	17.44
GPT	0	56.78	54.25	59.66	47.77	64.4	59.99	40.72
GPT-I	0	56.87	55.67	58.24	47.2	65.05	59.96	41.41
Falcon-I	40	36.33	33.21	39.9	29.88	41.79	38.18	27.07

Table 10: Specific performance of *Original-Dist*. Best from all models in **Bold** and best from open-sourced models in underline.

			$\mathcal{F}_1$		$\mathcal{F}_2$		$\mathcal{F}_3$	
Model	Size	Grounding Perf.	High	Low	Free-Form	Definite	Single	Multi
Vicuna	7	39.76	39.18	40.42	33.39	45.15	41.53	30.9
	13	55.04	52.76	57.65	46.8	62.02	58.48	37.88
TÜLU1	7	44.39	41.2	48.04	45.51	43.44	46.89	31.92
	13	40.37	39.03	41.9	45.77	35.81	43.04	27.02
	65	36.3	36.96	35.55	48.76	25.75	38.33	26.14
TÜLU2-D	30	40.87	39.78	42.1	47.04	35.64	42.61	32.14
	7	41.43	39.63	43.47	33.29	48.31	42.27	37.19
	70	55.06	53.88	56.42	47.51	61.45	57.34	43.70
	13	54.19	52.11	56.56	45.41	61.62	56.48	42.71
TÜLU2	7	47.92	45.12	51.13	42.4	52.6	50.41	35.47
	70	52.38	49.72	55.41	50.87	53.65	54.48	41.86
	13	50.41	47.13	54.16	48.66	51.9	52.44	40.27
Mistral-I	7	54.28	51.51	57.44	47.83	59.73	57.23	39.49
Zephyr	7	52.4	50.3	54.8	43.99	59.52	54.36	42.62
Llama2-C	7	40.39	38.77	42.24	31.15	48.21	41.86	33.06
	13	46.45	45.09	48	40.95	51.1	48.52	36.09
	70	54.36	53.43	55.42	47.7	60	56.63	42.99
Llama2	13	19.3	19.17	19.44	20.38	18.38	20.03	15.64
GPT	-	56.08	52.4	60.28	50.08	61.15	58.64	43.28
GPT-I	-	54.54	53.61	55.6	48.53	59.62	56.56	44.41
Falcon	40	12.14	10.27	14.27	14.52	10.13	12.56	10.02
Falcon-I	40	32.6	28.6	37.16	27.69	36.75	34.47	23.21

Table 11: Specific performance of *Conflict-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	35.98	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	86.57	82.62	83.74	70.19	82.38	83.38	88.16	**91.31**
Original-Dist	70.88	70.83	72.85	80.26	81.50	77.27	77.33	63.2	70.26	79.51	87.00	**88.01**
Conflict-Gold	76.19	76.94	77.26	81.36	80.90	76.64	76.82	58.84	71.49	78.29	**86.13**	84.79
Conflict-Dist	66.91	64.67	57.88	55.51	73.49	69.91	71.75	55.51	60.10	70.97	79.95	**83.32**

Table 12: Answer Accuracy of twelve different models. For each setting, the best in **bold** and the best of open-sourced models in underline.

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

Table 13: Fluency of LLMs measured by G-EVAL. Here, Llama is Llama2 and Llama-C is Llama2-Chat and Falcon-I is Falcon-Instruct.

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.

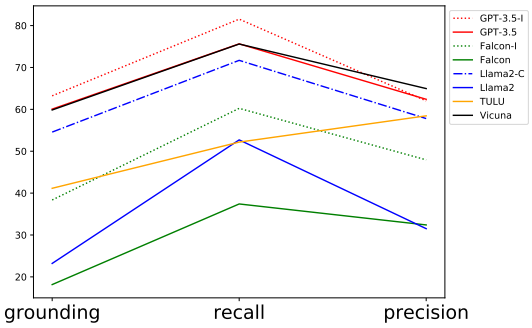


Figure 15: Performance of grounding performance, precision, and recall in *Original-Gold*

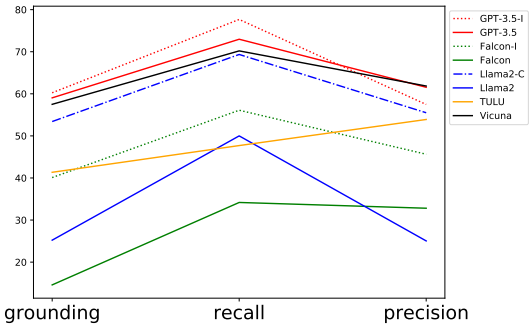


Figure 16: Performance of grounding performance, precision, and recall in *Conflict-Gold*

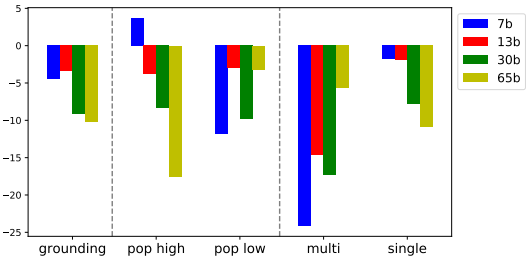


Figure 17: Reduction rate in grounding performance when adding distractor contexts

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