Find the INTENTION OF INSTRUCTION: Comprehensive Evaluation of Instruction Understanding for Large Language Models

Hyeonseok Moon, Jaehyung Seo, Seungyoon Lee, Chanjun Park†, Heuiseok Lim†
Korea University, Republic of Korea
{glee889, seojae777, dltmddbs100, bcj1210, limhseok}@korea.ac.kr

Abstract

One of the key strengths of Large Language Models (LLMs) is their ability to interact with humans by generating appropriate responses to given instructions. This ability, known as instruction-following capability, has established a foundation for the use of LLMs across various fields and serves as a crucial metric for evaluating their performance. While numerous evaluation benchmarks have been developed, most focus solely on clear and coherent instructions. However, we have noted that LLMs can become easily distracted by instruction-formatted statements, which may lead to an oversight of their instruction comprehension skills. To address this issue, we introduce the INTENTION OF INSTRUCTION (IOINST) benchmark. This benchmark evaluates LLMs' capacity to remain focused and understand instructions without being misled by extraneous instructions. The primary objective of this benchmark is to identify the appropriate instruction that accurately guides the generation of a given context. Our findings suggest that even recently introduced state-of-the-art models still lack instruction understanding capability. Along with the proposition of IOINST in this study, we also present broad analyses of the several strategies potentially applicable to IOINST. [1]

1 Introduction

The utilization of large language models (LLMs) has emerged as a prominent trend across numerous research domains (Biswas, 2023b,a; Peng et al., 2023; Taori et al., 2023; Zhou et al., 2023a). A notable feature of LLMs is their ability to interact with humans by appropriately responding to user instructions (Zhou et al., 2023b; Qin et al., 2024; Chen et al., 2024; Li et al., 2024; Xu et al., 2023). With any given instructions, LLMs are expected to generate responses that align with these instructions (Chen et al., 2024; Li et al., 2024; Xu et al., 2023; Longpre et al., 2023).

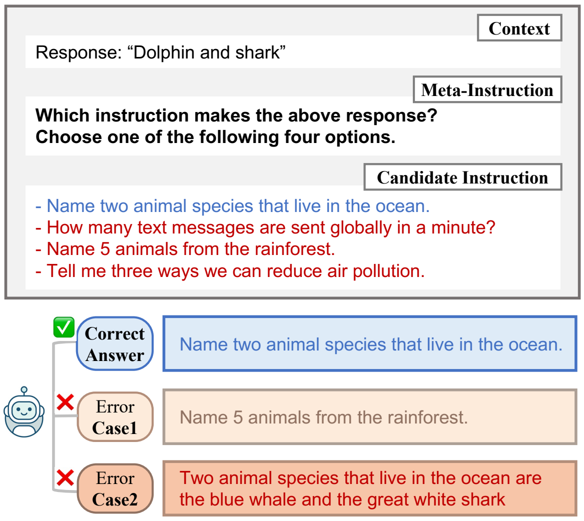

Figure 1: Simplified example of IOINST. We compose a benchmark designed to comprehend and select the appropriate instruction that derives given response. Potential error cases include misunderstanding prerequisites of context and responding to any candidate instruction.

This capability, known as the "instruction following" ability, serves as a key metric for assessing the effectiveness of LLMs (Chen et al., 2024; Zhao et al., 2024; Dubois et al., 2023; Zheng et al., 2023). To facilitate a more thorough assessment, several benchmarks have been introduced with a focus on instruction following. (Zhou et al., 2023b; Qin et al., 2024; Geng et al., 2023).

However, we have identified a blind spot in these evaluation methods: they primarily focus on the ability to follow coherent and clear instructions. We observed that LLMs can become distracted when faced with instruction-formatted statements that diverge from user intent. For instance, when given the statement: *Paraphrase the following*

† Corresponding Author
[1]Code and datasets are available at https://github.com/hyeonseokk/IoInst

Findings of the Association for Computational Linguistics: NAACL 2025, pages 5959–5979
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

statement: "Write a creative poem about a turtle", LLMs often respond to the instruction "Write a creative poem" rather than the intended task of "Paraphrase the following statement." Previous research often dismissed these instances as mere errors.

This observation leads us to question whether LLMs respond intuitively to instruction-formatted statements rather than relying on a deep understanding of the instructions. Given that several studies focus on manipulating instructions using LLMs, such as through instruction optimization (Fernando et al., 2024; Yang et al., 2024), the ability to maintain focus without being distracted by other instructions is considered a vital competency for LLMs. To clarify this capability and further analyze and evaluate the instruction understanding ability of LLMs, we propose the INTENTION OF INSTRUCTION (IOINST) benchmark.

The primary objective of IOINST is to identify the appropriate instruction that accurately instructs to generate a given **context**. IOINST provides four **candidate instructions**: one label instruction and three contrastive instructions, along with a **context** that can be generated by following the label instruction. Then, LLMs are required to select one among them. A simplified example of IOINST is shown in Figure 1. We denote the instruction that contains the actual intention (requests to select one among candidate instructions) as "**meta-instruction**."

IOINST aims to evaluate two key capabilities of LLMs. Firstly, we assess the ability to comprehend the prerequisites for generating a response. To accurately select one of the four candidate instructions, the model must discern the intent embedded in the instructions and correctly match it with the response. Secondly, we evaluate the ability to distinguish user intention among instruction-formatted statements. The model should focus on selecting the correct instruction from the four given candidate instructions without being distracted by other instruction-formatted statements. In this benchmark, we intend for the model to "choose one of the four instructions." Therefore, any action contrary to this—such as following a candidate instruction or generating an instruction not provided—will be considered a deficiency in the model's performance.

To ensure a comprehensive evaluation, we define three different data types: Random, which involves arbitrary contrastive instructions; Semantic, which comprises semantically confusing contrastive instructions; and Anti-Attribute, which requires an understanding of the finer correlation between response and each instruction. Additionally, we introduce three evaluation metrics, allowing for a more detailed inspection of the LLM's ability to understand instructions.

Through extensive experiments with several instruction-tuned LLMs, we reveal that most LLMs struggle to grasp the intention embedded in the instruction. Notably, several LLMs, including Mistral (Jiang et al., 2023) and Gemma (Team, 2024a), seem to not fully comprehend the instruction and frequently respond to the instruction-formatted statement in the given input (*i.e.* candidate instruction). These findings highlight that even recently introduced state-of-the-art models still lack instruction understanding capability, suggesting areas for future improvement. Along with the proposition of IOI in this study, we also present broad analyses of the several strategies potentially applicable to IOI. With the rapid advancements in LLM (Taori et al., 2023; Peng et al., 2023; Jiang et al., 2023; Kim et al., 2024), several attempts to objectively assess the performance of LLMs have been introduced. These efforts include methodologies such as A/B testing (Wang et al., 2024b; Zeng et al., 2024; Quin et al., 2024) or evaluations based on quantified scores (Wang et al., 2023).

2 Related Works

The concept of instruction-following has recently been incorporated to facilitate more objective evaluations (Zhou et al., 2023b; Dubois et al., 2023; Qin et al., 2024). The effectiveness of this approach lies in the potential to objectively quantify the quality of outputs, which might otherwise be deemed subjective (Zhou et al., 2023b; Zheng et al., 2023). During instruction-following evaluation, we provide specific instructions and assess whether the generated output aligns well with the intended content of the instruction. Adherence to instruction is generally verified either through a rule-based measurement (Zhou et al., 2023b; Xia et al., 2024) or by employing a super-LLM evaluator (such as GPT-4 (Team, 2024b)) on the generated output (Qin et al., 2024).

However, these benchmarks often neglect the ability of LLMs to distinguish between several instruction-formatted statements. Current evaluations focus on how LLMs respond to clear and coherent instructions (Zhang et al., 2023; Dubois et al., 2023; Zeng et al., 2024), which may result in high performance without a deep understanding of the underlying intention. Therefore, we propose a more structured benchmark to assess the fundamental ability of LLMs to comprehend given instructions.

3 INTENTION OF INSTRUCTION

3.1 Task Definition

The primary objective of IOINST is to identify an appropriate instruction that leads to generating a given **context**. We provide LLMs a **context** along with four **candidate instructions**, among which one instruction correctly instructs to generate the given **context**, while others are not. IOINST requires LLMs to select one correct candidate instruction among the given four. Such task objective is defined in the **meta-instruction** and is also fed to the LLM during the evaluation procedures.

3.2 Data Composition

Candidate instructions are composed of a label instruction configured to generate **context**, along with three contrastive instructions that serve as distractors. Through quality inspection, we ensured that contrastive instructions were designed to avoid generating **context**. We construct three different types of contrastive instruction, each distinguished by the required level of understanding. Details about contrastive instructions are described in Section 4.2.

Context is a generation result by LLM, constructed as a response to the label instruction. Through carefully designed contrastive instructions, we ensure that none of the contrastive instructions align with the context.

Meta-instruction instructs LLMs to select the most appropriate candidate instruction. We construct 16 different meta-instructions and randomly select one for each evaluation phase. We also define two distinct characteristics that can subdivide meta-instructions into two separate groups. The first criterion is the level of details that can partition meta-instructions into *Detailed* group (instruct detailed step-by-step procedures) and *Simple* group (provide straightforward task directives). The second is the presenting order between the context and candidate instructions, which can define two groups: *Candidate-First* and *Context-First*. By defining diverse meta-instructions, we investigate the capability of instruction understanding affected by meta-instruction.

3.3 Potential Error Cases

We define two types of error cases that LLMs may encounter and evaluate their instruction understanding capability based on the frequency of these cases.

Case 1 pertains to an error case that chooses the incorrect candidate instruction from the given four instructions. In this case, we can see that LLM faithfully followed meta-instruction, while LLM has not fully understood the requirements for crafting the response.

Case 2 involves a failure to comprehend the user intention embedded in the meta-instruction. Such error cases include responding to the candidate instruction rather than following meta-instruction. Our objective is to select one of four options; therefore, generating anything outside of this purpose, including suggesting alternatives or refusing to provide an answer, is considered a Case 2 error.

3.4 Evaluation Measure

For each data point in IOINST, we examine the generation output of the LLM instead of estimating the sequence generation probability, as Case 2 errors can only be assessed during the generation phase. We construct the benchmark to allow evaluation without external judges, such as human evaluators or LLM assessments (Liu et al., 2023). Specifically, we define three metrics, including $\mathbf{ACC_1}$, $\mathbf{ACC_2}$, and $\mathbf{ACC_1^{rel}}$, for a detailed assessment.

- $\mathbf{ACC_1}$: This metric estimates strict accuracy of the intended goal; among four candidate instructions, the LLM correctly selects label instruction.
- $\mathbf{ACC_2}$: This metric estimates how accurately the LLM catches the given instruction's underlying intention; the LLM selects one of four candidate instructions whichever.
- $\mathbf{ACC_1^{rel}}$: This metric measures how accurately the LLM selects the correct instruction, relative to the number of times it follows the meta-instructions. This metric estimates the LLM's ability to understand instructions while avoiding underestimation due to **Case 2**.

For each model generation X, we define the evaluation measure for each case as in Equation (1).

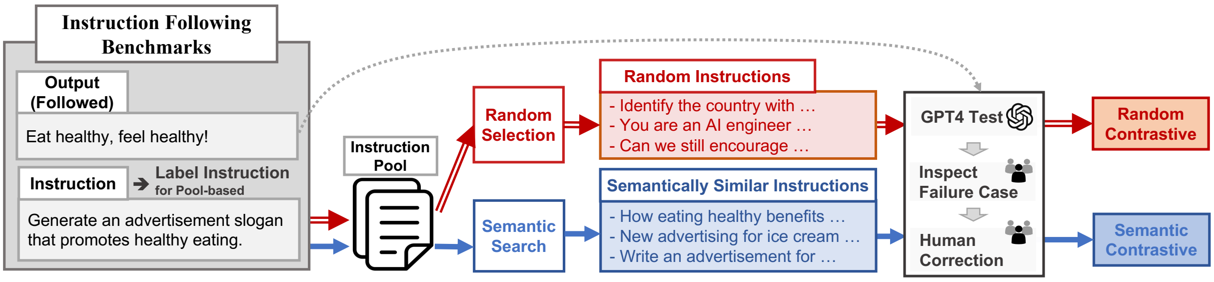

Figure 2: Construction of pool-based contrastive instructions. From the pre-processed data point obtained by the data curation, we establish our instruction candidates.

.

Data	Instruction Pool	Label Instruction	# of Data
Zeng et al. (2024)	✓	✓	352
Zhou et al. (2023b)	✓	✓	169
Zheng et al. (2023)	✓	✓	55
Lin and Chen (2023)	✓	✓	42
Ortmann (2022)	✓	✓	13
Qin et al. (2024)	✓	-	-
Chia et al. (2024)	✓	-	-
Total Quantity of Data: ***631***			

Table 1: Data statistics. Detailed explanation for each dataset is described in Appendix E.1.

We denote candidate instruction set $\{cand_i\} = \{label\} \cup \{contrastive_j\}_{j=1}^{3}$, where $label$ and $contrastive$ are label and contrastive instruction, respectively.

$$
\begin{aligned}
\mathbf{ACC_1} &= \frac{1}{|D|}\sum^{|D|} \mathbb{1}[\mathrm{EM}_{sparse}(X, label)] \\
\mathbf{ACC_2} &= \frac{1}{|D|}\sum^{|D|} \mathbb{1}[\bigcup_{i=1}^{4} \mathrm{EM}_{sparse}(X, cand_i)] \\
\mathbf{ACC_1^{rel}} &= \mathbf{ACC_1}/\mathbf{ACC_2}
\end{aligned} \tag{1}
$$

In particular, we introduce a measure called EM_{sparse} for more accurate evaluation. This approach is adopted to overlook subtle differences included in the LLM-generated output and evaluate them on a similar basis. It examines the ROUGE-L (Lin, 2004) precision between the generated output and the reference. Then, we regard a generated output that surpasses a threshold τ as a correct result and subsequently estimate the overall accuracy of INTINST. In this study, the threshold τ was empirically set to 0.9.

4 Data Construction

4.1 Data Curation

To gather high-quality instructions and their corresponding LLM responses, we curated several datasets released with the LLM-as-evaluator approach (Zeng et al., 2024; Zheng et al., 2023; Chia et al., 2024). These datasets predominantly contain two model responses for each instruction—one that accurately follows the instruction and one that does not. In this case, we can ensure that one of the model responses is guaranteed to be well-aligned with the instruction.

Using these resources, we reorganized the data to fit our objectives. We secured high-quality instructions along with their guaranteed instruction-aligned responses by extracting the labeled responses from the given datasets. We regard the instruction-followed model response obtained from these datasets as a **context**, and corresponding instruction as a label instruction for **candidate instructions**.

In particular, through pilot experiments, we noticed lengthy context within input statements could degrade the context comprehension of LLMs, which leads to an underestimation of $\mathbf{ACC_1}$ and $\mathbf{ACC_1^{rel}}$. We also found that LLMs are prone to follow excessively long candidate instructions, wherein $\mathbf{ACC_2}$ deteriorates. Considering these, we extract data that do not exceed the 1,000 and 300 character lengths for the context and candidate instructions, respectively.

Along with constructing label instruction and context pair, we also compile instructions from these sources (Qin et al., 2024; Chia et al., 2024). The complicated instruction set is denoted as **candidate pool**, and serves as a resource during the construction of Pool-based contrastive instructions (Detailed in Section 4.2). The list of datasets utilized for the construction of IOINST is detailed in Table 1. More detailed characteristics of those datasets are demonstrated in Appendix E.1.

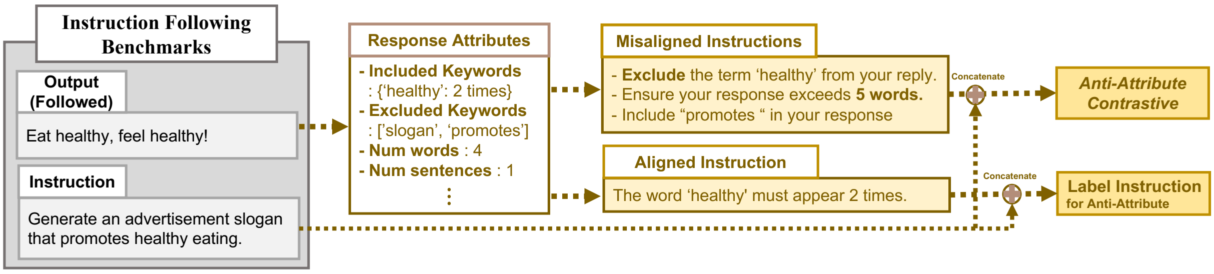

Figure 3: Construction of Anti-Attribute contrastive instructions. From the pre-processed data point obtained by the data curation, we establish our instruction candidates.

Attribute	Explanation
num_words	Number of words in the **context** (excluding stopwords)
num_sentences	Number of sentences in the **context** (excluding stopwords)
num_paragraphs	Number of paragraphs in the **context**
num_words_capital	Number of capital words in the **context**
is_quotation	Boolean value indicating whether the **context** is wrapped with quotation marks or not.
is_comma	Boolean value indicating whether the comma is included in the **context** or not.
end_sentence	The last sentence of the **context**.
start_sentence	The first sentence of the **context**.
keywords_included	List of words included in the **context**, and their occurrence frequencies (excluding stopwords)
keywords_excluded	List of words included in the **label instruction** but not included in the **context** (excluding stopwords)

Table 2: List of attributes exploited in constructing Anti-Attribute contrastives. We adopt stop-word list released by the NLTK (Bird and Loper, 2004) package. Segmentation tools for words, sentences, and paragraphs are released by Zhou et al. (2023b).

4.2 Contrastive Instruction: Pool-based

We then establish three contrastive instructions for each label instruction to act as distractors for LLM. To more precisely assess the instruction understanding capability of LLMs, we propose three different variants of contrastive instruction. Firstly, we introduce a method to construct contrastive datasets using a candidate pool, creating two types of dataset variants. The detailed procedures are described in Figure 2.

Random Contrastive: For the random contrastive, we randomly extract three different instructions for each data point and utilize them as contrastive instructions. Such dataset type is introduced to assess the LLM's low-level capability of understanding instructions.

Semantic Contrastive: For constructing semantic contrastive, we extract instructions that are semantically similar to the label instruction but lead to different outcomes. We employ a sentence embedding model MPNet (Song et al., 2020) to determine semantic similarity. We extract the most semantically similar three instructions to compose contrastive instructions. LLMs are then obliged to differentiate and select the instruction that most accurately reflects the requirements of semantically confusing candidates.

Quality Assurance In the pool-based contrastive instructions we designed, there is a possibility of including ambiguous instructions that prompt the generation of a **context**. To prevent this, we eliminated all such ambiguous scenarios through human inspection. We conducted evaluations using GPT-4o (Hurst et al., 2024), directly reviewing failure cases arising from the model. Suppose these failure cases are incurred by any contrastive instruction that leads to generating a **context**. In that case, we re-generated such contrastive instruction to ensure our benchmark remained clear and solvable enough. By conducting this error verification process three times, we ensured the dataset was free from errors caused by contrastive instructions misguiding the labels.

Supporting this validation, GPT-4o achieved an accuracy of 99.54, demonstrating that solving this task is feasible given sufficient model performance.

4.3 Contrastive Instruction: Anti-Attribute

To more precisely evaluate whether the LLM comprehends the detailed requirements of instructions, we design heuristic-based Anti-Attribute contrastive instructions. Firstly, we examine the at-

Models	Random			Semantic			Anti-Attribute		
	ACC_1	ACC_2	$\mathrm{ACC}_1^{\mathrm{rel}}$	ACC_1	ACC_2	$\mathrm{ACC}_1^{\mathrm{rel}}$	ACC_1	ACC_2	$\mathrm{ACC}_1^{\mathrm{rel}}$
Llama-2-7B	$35.75_{2.39}$	$61.43_{1.28}$	$58.19_{3.46}$	$32.40_{0.78}$	$58.48_{1.57}$	$55.43_{1.83}$	$17.37_{1.55}$	$47.38_{0.72}$	$36.69_{3.65}$
Llama-2-13B	$50.08_{2.33}$	$71.73_{1.43}$	$69.80_{2.14}$	$44.41_{2.80}$	$69.25_{2.24}$	$64.09_{2.57}$	$23.48_{0.86}$	$61.55_{2.71}$	$38.21_{2.02}$
Llama-3-8B	$75.66_{1.26}$	$81.74_{1.34}$	$92.56_{0.73}$	$68.21_{0.8}$	$81.14_{1.27}$	$84.07_{0.76}$	$23.58_{0.57}$	$73.28_{1.04}$	$32.18_{0.62}$
Mistral-7B	$47.67_{2.01}$	$53.31_{1.26}$	$89.39_{1.96}$	$47.86_{2.20}$	$56.86_{2.12}$	$84.16_{1.23}$	$11.28_{1.36}$	$31.06_{2.26}$	$36.29_{2.86}$
Gemma-7B	$52.45_{1.71}$	$70.71_{1.26}$	$74.17_{1.70}$	$49.45_{2.83}$	$65.42_{1.93}$	$75.56_{3.12}$	$\mathbf{51.47}_{1.54}$	$72.55_{0.64}$	$\mathbf{70.94}_{1.78}$
Solar-10.7B	$63.47_{1.92}$	$66.25_{1.42}$	$95.80_{1.01}$	$65.02_{1.79}$	$73.38_{0.91}$	$88.60_{1.77}$	$21.59_{0.74}$	$56.58_{2.03}$	$38.18_{1.27}$
CommandR	$57.53_{1.23}$	$63.62_{1.07}$	$90.43_{1.46}$	$52.58_{2.29}$	$64.72_{1.56}$	$81.23_{2.39}$	$16.35_{0.97}$	$49.26_{0.92}$	$33.20_{1.91}$
Mixtral-8x7B	$72.50_{1.96}$	$80.23_{2.60}$	$90.38_{0.56}$	$66.05_{2.34}$	$80.15_{1.08}$	$82.38_{1.86}$	$24.94_{2.62}$	$67.93_{0.48}$	$36.70_{3.78}$
GPT-3.5	$79.02_{0.61}$	$83.77_{0.46}$	$94.33_{0.70}$	$70.11_{1.07}$	$83.90_{0.96}$	$83.57_{0.48}$	$28.68_{1.46}$	$85.42_{1.45}$	$33.59_{1.91}$
GPT-4o	$\mathbf{95.31}_{0.33}$	$\mathbf{95.75}_{0.31}$	$\mathbf{99.54}_{0.14}$	$\mathbf{92.30}_{0.66}$	$\mathbf{96.32}_{0.59}$	$\mathbf{95.82}_{0.55}$	$48.74_{0.97}$	$\mathbf{97.93}_{0.66}$	$49.77_{0.94}$
GPT-4o-mini	$89.48_{0.71}$	$90.59_{0.84}$	$98.77_{0.17}$	$84.28_{0.58}$	$91.47_{0.55}$	$92.14_{0.91}$	$38.60_{0.64}$	$91.78_{1.12}$	$42.06_{0.78}$

Table 3: Performance of LLMs. All the models are instruction-tuned models. In all experiments, we conducted five trials, and report the average performance. Each lower-script denotes standard deviation in our experiments.

tribute of the given context as listed in Table 2. Every attribute can undoubtedly be detected via simple rule-based methodologies, and is established by referring criteria presented by Zhou et al. (2023b).

We then generate aligned and unaligned instructions based on these analyzed attributes, as depicted in Table 9. We crafted human-made instructions for each attribute and paraphrased them to generate 10 variants via GPT-4o. Prompt employed to paraphrasing is described in Appendix E.3, and any awkward expressions in the generated statements were manually corrected. We designed 10 templates for each attribute and constructed the dataset by randomly applying one of these templates.

Each aligned and unaligned instructions is concatenated with the original label instruction to construct candidate instructions. In this sense, contrastive instructions share key requirements with the label instruction (*e.g.* Write a poem), while exhibit misalignments with the context in finer details (*e.g.* Your response should not exceed 50 words). This allows us to assess the instruction understanding capability of the LLM at a more intricate level.

Quality Assurance The three contrastive instructions generated by Anti-Attribute are designed to ensure that they do not prompt the creation of a **context**, and only the label instruction is crafted to accurately generate the **context**. This design ensures the clarity of candidate instructions.

5 Results and Discussion

5.1 Experimental Settings

In our experiments, we validate the generalizability of our benchmark by experiments with instruction-tuned LLMs. This covers eight different open-sourced LLMs, each varying in parameter size and instruction following capabilities, and GPT-3.5 (OpenAI-Blog, 2022) and GPT-4o series (Team, 2024b). Detailed information of experimented LLMs and details about the experimental settings can be found in Appendix D. Considering influence of the order in which candidate options are presented (Zheng et al., 2024), we experiment by randomizing the sequence of candidate instructions and report an average from five experimental trials.

5.2 Main Results

Firstly, we present the performance of LLMs concerning IoInst in Table 3. We randomly select one of sixteen meta-instructions for each inference step in this experiment. The reported performance corresponds to the average outcome of five trials. Our observations are as follows:

LLMs frequently follow candidate instruction We can observe that LLMs tend to easily follow candidate instructions, as evidenced by the low ACC_2 scores across most LLMs. Low level of ACC_2 suggests that LLMs are inclined to distracted by several isntruction-formatted statements, and struggle to find the actual intent embedded in the instruction. This suggests that even LLMs capable of following human instructions through instruction tuning still lack a robust understanding of the requirements specified in the instructions.

Composition of candidate instructions highly affects instruction understanding capability Table 3 reports that LLMs demonstrate distinct ACC_2 scores across Random, Semantic, and Anti-Attribute settings. Such findings indicates how the composition of candidate instructions influences

Candidate Instruction	Error Rate (Case 2)
Identify the main clause in this sentence. Output should be a single sentence. A busy morning, interrupted by a phone call from a friend, made John late for school.	93.57
Pretend to be arrogant. Answer questions in that manner. Don't write that you're not programmed to be arrogant. Question: Where were you until this year?	91.43
why is the Cretaceous-Tertiary boundary called the K-T boundary and not the C-T boundary ? I've had it said that K is an abbreviation for Cretaceous, but no real explanation why.	90.71
In this task, repeat the exact request below first, then give your response. Do not say any word before repeating the exact request. Write an acoustic song about the Korean peninsula without using any commas	88.21
⋮	
Is it ethical to use AI in decision-making processes?	0.00
Name three ways to extend the battery life of a laptop	0.00
Suggest a method for building self-confidence.	0.00
Write a blog post with 400 or more words about the benefits of sleeping in a hammock.	0.00

Table 4: Error rates for each individual candidate instruction.

Meta-Instruction	Accuracy
You must identify the correct instruction that produces the specified [RESPONSE]: Begin by understanding the described [RESPONSE] as stated: **{Context}** Review this list of [INSTRUCTIONS] to determine which matches the [RESPONSE]: **{Candidate Instructions}** Select the instruction that most accurately aligns with the [RESPONSE] mentioned.	$\mathrm{ACC_1}$: **62.57** $\mathrm{ACC_2}$: **80.26** $\mathrm{ACC_1^{rel}}$: **76.55**
⋮	
Select the most appropriate option that guides the creation of the described [RESPONSE]: [RESPONSE]: **{Context}** Choose one of the following [INSTRUCTIONS]: [INSTRUCTIONS]: **{Candidate Instructions}**	$\mathrm{ACC_1}$: **38.96** $\mathrm{ACC_2}$: **53.80** $\mathrm{ACC_1^{rel}}$: **69.69**

Table 5: Average accuracy for meta-instructions. We present the meta instructions that derived the highest and lowest performance.

Criteria	$\mathrm{ACC_1}$	$\mathrm{ACC_2}$	$\mathrm{ACC_1^{rel}}$
Detail	**54.04**$_{3.51}$	**71.97**$_{4.94}$	73.86$_{1.54}$
Simple	50.13$_{\mathbf{9.65}}$	68.11$_{\mathbf{11.60}}$	72.08$_{\mathbf{3.24}}$
Context First	53.02$_{7.27}$	69.72$_{8.99}$	**73.96**$_{2.41}$
Candidate First	51.15$_{7.71}$	70.36$_{9.30}$	71.98$_{2.58}$

Table 6: Average performance and standard deviation by meta instruction categories. We computed the average accuracy for each meta instruction and calculated the standard deviations for each category, which are denoted in subscript.

the instruction understanding capabilities.

On comparing the performance between Random and Semantic, we can find subtle difference in $\mathrm{ACC_2}$, likely due to the analogous compositions of candidate instructions derived by a similar process of composing contrastive instructions (selection from the instruction pool). We can also observe a slight but clear difference in $\mathrm{ACC_1}$ and $\mathrm{ACC_1^{rel}}$, which suggests that LLMs experience minor confusion when selecting appropriate instructions among semantically ambiguous options.

Moreover, we can notice a performance drop in $\mathrm{ACC_2}$ for the Anti-Attribute. Considering the contrastive instructions of Anti-Attribute, which state the same key requirements of the label instruction, we can interpret that LLMs tend to prioritize the repetitive pattern of the instructions, consequently neglecting the core intent of the instruction.

Given LLM follows meta-instruction, LLM can fairly well understand instruction Despite the low $\mathrm{ACC_1}$, we can witness decent level of $\mathrm{ACC_1^{rel}}$. In random contrastive settings, GPT-4 achieved a near-perfect score for $\mathrm{ACC_1^{rel}}$, suggesting that once GPT-4 grasps the intent of the instruction, it can accurately identify the appropriate instruction to generate the relevant **context**. This demonstrates that LLMs inherently possess the capability to understand instructions, although fully harnessing this capability requires a strong ability to discern the embedded intent.

Exceptional case: Anti-Attribute In the context of Anti-Attribute, where contrastive instructions differ only slightly from label instructions, it is crucial for LLMs to discern the precise requirements that can lead to generate a given context.

Our experimental results reveal that across all LLMs, the $\mathrm{ACC_1^{rel}}$ in the Anti-Attribute setting were significantly low. This observation indicates that LLMs still struggle to comprehend the relation between the instruction and its corresponding response. Exceptionally, Gemma demonstrated an $\mathrm{ACC_1^{rel}}$ score exceeding 70 in the Anti-Attribute setting, a performance that is roughly double that of other LLMs, indicating a distinctive strength in this area. However, Gemma exhibited lower $\mathrm{ACC_1}$ and $\mathrm{ACC_2}$ in both Random and Semantic settings compared to other models. This suggests that superficial and deep levels of instruction comprehension do not necessarily align and should be considered distinct evaluation measures.

5.3 Discussion on the Instruction Composition

To analyze the performance derived by the meta instruction, we estimate the average accuracy for each meta-instructions, across all models and all settings (Random, Semantic, Anti-Attribute). Table 5 displays the meta-instruction with the highest average accuracy alongside the one with the lowest. We can ascertain a significant variation, with a maximum difference of 27 points in $\mathbf{ACC_2}$, emphasizing the need for rigorous consideration in designing meta-instructions.

Table 6 presents the average performance and the corresponding standard deviations for the meta-instruction categories defined in our study. While minor differences in performance were noted based on the order of presenting context and candidates (Candidate-First vs. Context-First), increasing the level of detail in the instructions (Detailed vs. Simple) consistently yielded higher and more stable performance in instruction understanding. This stability highlights the importance of employing detailed meta-instructions to achieve robust and high instruction understanding capabilities.

Moreover, to investigate the influence of candidate instructions on the understanding capabilities, we calculate the error rates for each individual candidate instruction. Here, the error rate refers to the proportion of **Case2** errors occurred across all experiments. Table 4 displays candidate instructions with the highest and lowest error rates in a semantic contrastive setting. Our experiments revealed that instructions composed of multiple lines and with complex requirements tend to be indiscriminately followed by the LLMs. Conversely, LLMs show a better capability to accurately identify simple, single-line candidate instructions. This suggests that LLMs still struggle to comprehend the intent behind more complex instruction-formatted statements.

5.4 Case Study on the In-context Learning

To fully leverage the potential of LLMs, the in-context learning approach is widely adopted (Dong et al., 2024; Brown et al., 2020). We constructed in-context samples by randomly referencing data from the benchmark dataset and assessed the few-shot performance. Unfortunately, this approach proved significantly less effective for IoInst.

Figure 4 demonstrate a tendency for performance in IoInst to **decrease as shots are introduced**, where shots are randomly extracted within our dataset. Especially, $\mathbf{ACC_1^{rel}}$ consistently exhibits a decrease in performance as the number of shots increased, and $\mathbf{ACC_2}$ significantly drop as shots introduced. One possible interpretation

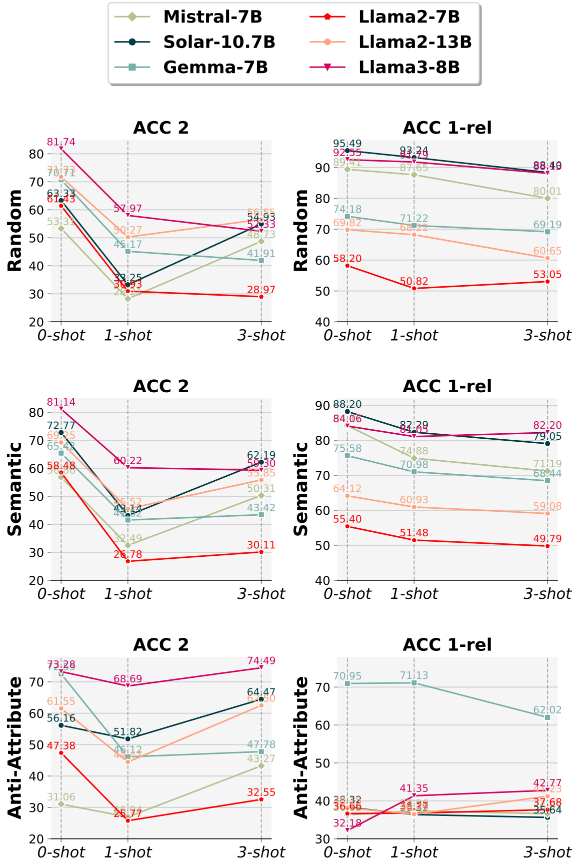

Figure 4: Performance comparison between zero-shot and few-shot settings.

is that the few-shot examples used as references may have acted as distractors. The superior performance of zero-shot over few-shots strengthens our claim that LLM becomes confused by the presence of instruction-formatted inputs within the provided input statements.

6 Conclusion

This study highlights a crucial ability of LLMs to understand instructions without being misled by instruction-formatted statements. To measure this ability and enable objective evaluation, we introduce IoInst. By utilizing Random, Semantic, and Anti-Attribute contrastive settings, each with unique characteristics, we facilitate a broader exploration of the instruction comprehension capabilities of LLMs. Our results show that most publicly accessible LLMs struggle with IoInst, could not catch the actual intention embedded in the given instruction. Our experiments reveal that the choice of meta-instructions greatly influences comprehension, while the in-context approach is inadequate in addressing these challenges. Our future research aims to explore data-centric and model-based strategies to enhance instruction comprehension in LLMs.

Limitation

In our study, we conducted experiments on eight open-sourced LLMs. Although more generalized conclusions could potentially be obtained through broader range of model variants, it was challenging for us due to resource constraints. However, by analyzing various parameter-sized LLMs, we were able to draw generalized conclusions.

There remains a vast array of proprietary and closed-source LLMs that we did not have access to, which could potentially exhibit different behaviors and capabilities when presented with instruction-formatted prompts. However, our experiment includes ten different LLMs, comprising both open-sourced and closed LLMs. Through comparative analyses among these models, we observed distinct similarities in their performance trends, allowing us to draw sufficiently generalized conclusions.

Ethics Statement

In conducting our research, we placed a strong emphasis on ethical considerations and mitigating potential risks. The dataset used in our experiments was meticulously constructed by reorganizing and curating data from previously released datasets that are publicly available under permissive licenses, such as the Apache 2.0 (Zhou et al., 2023b; Zheng et al., 2023; Lin and Chen, 2023; Chia et al., 2024) and MIT (Zeng et al., 2024; Ortmann, 2022; Qin et al., 2024; Chia et al., 2024) licenses. These licenses allow for the modification and redistribution of the data, provided that the original sources are properly acknowledged and attributed.

To ensure compliance with copyright and intellectual property regulations, we took great care to retain all original data in its unmodified form and clearly denote any alterations or modifications made during our curation process. This transparency allows for the traceability of our dataset's provenance and ensures that the original creators' rights are respected.

Furthermore, we implemented a rigorous human supervision process to review and curate both the recompiled data from existing datasets and any additional data generated using language models like ChatGPT. Through this meticulous review process, we removed any potentially harmful, or inappropriate contents.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, 2022-0-00369 (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

References

AI@Meta. 2024. Llama 3 model card.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Som S Biswas. 2023a. Potential use of chat gpt in global warming. *Annals of biomedical engineering*, 51(6):1126–1127.

Som S Biswas. 2023b. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2024. InstructEval: Towards holistic evaluation of instruction-tuned large language models. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 35–64, St. Julian's, Malta. Association for Computational Linguistics.

Cohere. 2024. Command r models. `https://cohere.com/command`. Accessed: 2024-03-01.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 30039–30069. Curran Associates, Inc.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Forty-first International Conference on Machine Learning*.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. *Blog post, April*, 1:6.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 23–35, Mexico City, Mexico. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2024. Self-alignment with instruction back-translation. In *The Twelfth International Conference on Learning Representations*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt/`. Accessed: 2024-03-01.

Katrin Ortmann. 2022. Fine-grained error analysis and fair evaluation of labeled spans. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1400–1407, Marseille, France. European Language Resources Association.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models.

Federico Quin, Danny Weyns, Matthias Galster, and Camila Costa Silva. 2024. A/b testing: A systematic literature review. *Journal of Systems and Software*, 211:112011.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Gemma Team. 2024a. Gemma: Open models based on gemini research and technology.

OpenAI Team. 2024b. Gpt-4 technical report.

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

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. Fofo: A benchmark to evaluate llms' format-following capability.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *Forty-first International Conference on Machine Learning*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023c. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Practical Implication

Over the recent years, it has been observed that slight modifications in prompts alone can have significant impacts on the generative capabilities of LLMs (Kojima et al., 2022; White et al., 2023; Wang et al., 2024a). In considering such sensitivity to the instruction, numerous studies are being conducted with the aim to find out optimal instruction for each targeting task (Reynolds and McDonell, 2021; Qin et al., 2023).

Represented by instruction optimization, LLMs are commonly entrusted with the task of discovering these optimal instructions (Fernando et al., 2024; Yang et al., 2024; Zhou et al., 2023c; Wang et al., 2023). This include paraphrasing (Fernando et al., 2024; Yang et al., 2024) or self-refining manner (Pryzant et al., 2023).

In implementing these attempts, multiple instructions are inevitably included in a single input statement. Consequently, distinguishing the actual user-intended instructions and executing the meta instruction, rather than the candidate instruction, is considered an substantial issue.

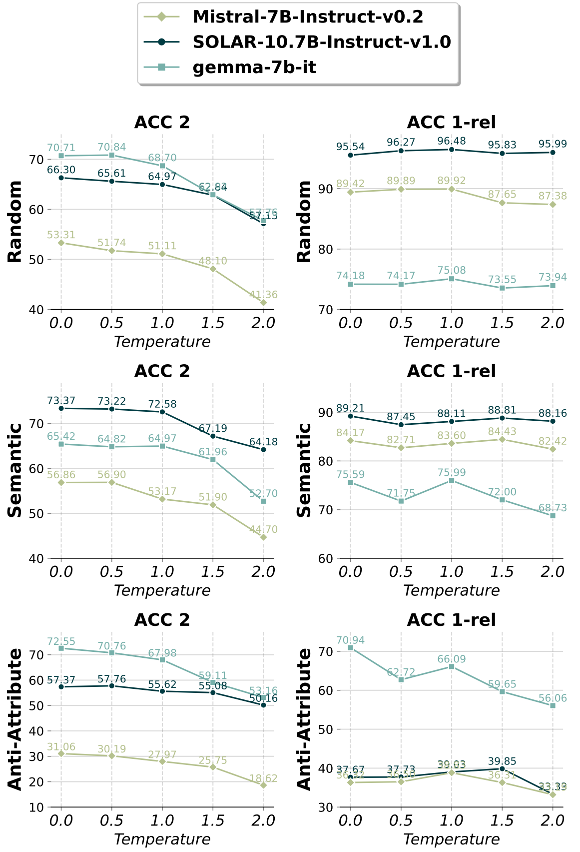

Figure 5: Performance variations with diverse temperature settings. Temperature 0.0 indicates greedy decoding.

Nevertheless, the problem of following candidate instructions instead of meta instructions has largely remained unaddressed; traditional research typically regarded this as a straightforward error (Zeng et al., 2024; Yang et al., 2024). IntInst is designed to comprehensively consider the aforementioned concerns and targeting the assessment of instruction distinguishing and understanding capabilities.

B Case Study: Temperature

To determine the experimental setting, we conducted a pilot study concerning the influence of temperature, we evaluate the understanding capability under varying temperature settings. Detailed experimental results are presented in Figure 5.

Through our experiments, we confirmed that the instruction understanding proficiency declines with an increase of temperature. Remarkably, $\mathbf{ACC_2}$ exhibited substantial variation, demonstrating a significant drop as the temperature rose.

Aside from the above observation, we can also figure out that there was mere alteration in $\mathbf{ACC}_1^{\mathbf{rel}}$ with respect to temperature changes. In essence, given the LLM accurately comprehend the intention within the meta instruction, the capability to distinguish appropriate instructions remains largely unvarying. Such findings compose our future research direction.

C Case Study: Meta-Instruction

In this section, we validate how the structure of meta-instruction affect instruction understanding capability of LLMs. As described in the Section 3.2, we establish two characteristics of meta-instruction concerning instructional details (*Detailed* vs *Simple*) and the order of components in the instruction (*Option first* vs *Option last*). We analyze the differences between these contrasting settings, and demonstrate our findings. The results are reported in Figure 6, and we analyze these results as follows.

We find that LLMs exhibit enhanced comprehension performance for specific forms of meta-instructions. Each model demonstrated a clear preference for either '*Detailed* / *Simple*' or '*Option-first* / *Option-last*' in each setting. This illustrates the potential for significantly improving an LLM's instruction understanding capability by choosing the most appropriate meta-instruction.

We observed significant differences in the tendencies exhibited in the Random, Semantic contrastive and the Anti-Attribute contrastive. In the case of Random and Semantic contrastive, LLMs showed a preference for *Simple* meta-instructions, and demonstrated improved performance when content instructions were provided at the end of the input (*Option-last*). However, in the Anti-Attribute task, the model found *Detailed* meta-instructions more effective, and performance was augmented when content instructions were presented ahead of content (*Option-first*). This implies the necessity to structure meta-instructions differently based on the complexity of the task for the LLM.

D Experimental and Writing Details

Our experiments were conducted using eight RTX A6000 GPUs. The specifications of the LLMs employed in our study are shown in Table 7. We applied greedy decoding for each experiment. Rationale for selecting greedy decoding is shown in Appendix B

In our work, we used GPT-4o (gpt-4o-2024-08-06) as a writing assistant. AI assistant was solely utilized for writing-related activities, such as grammar checking, refining awkward expressions, and translation of our manuscript.

Model Name	# Params
LLAMA2-chat-7B (Touvron et al., 2023) : `meta-llama/Llama-2-7b-chat-hf`	6.74B
LLAMA2-chat-13B (Touvron et al., 2023) : `meta-llama/Llama-2-13b-chat-hf`	13B
LLAMA3-8B-Instruct (AI@Meta, 2024) : `meta-llama/Meta-Llama-3-8B-Instruct`	8.03B
Mistral (Jiang et al., 2023) : `mistralai/Mistral-7B-Instruct-v0.2`	7.24B
Gemma (Team, 2024a) : `google/gemma-7b-it`	8.54B
Solar (Kim et al., 2024) : `upstage/SOLAR-10.7B-Instruct-v1.0`	10.7B
CommandR (Cohere, 2024) : `CohereForAI/c4ai-command-r-v01`	35B
Mixtral (Jiang et al., 2024) : `mistralai/Mixtral-8x7B-Instruct-v0.1`	46.7B
GPT3.5 (OpenAI-Blog, 2022) : `gpt-3.5-turbo-16k`	-
GPT4-turbo (Team, 2024b) : `gpt-4-turbo-preview`	-

Table 7: Model Details. We deployed OPENAI API call for experiments with GPT3.5 and GPT4, and `huggingface` (Wolf et al., 2020) for eliciting model weights for other publicly-available LLMs.

E Dataset Details

E.1 Exceptional Case for Dataset Construction

INTINST is constructed by re-organizing LLM-as-evaluator benchmark datasets. Exceptionally, the dataset introduced by Zhou et al. (2023b) was not developed following the LLM-as-evaluator format, among datasets detailed in Table 1. This dataset only constructed with instruction-following evaluation objective; but it release its LLM response generated by GPT-4. Additionally, its evaluation measure for assessing instruction-following facets is established with rule-based methodology, facilitating straightforward estimation of data quality. Consequently, we extract only responses that faithfully followed given instructions and accumulate pairs of instructions and GPT-4 responses, utilizing them to generate the data for INTINST.

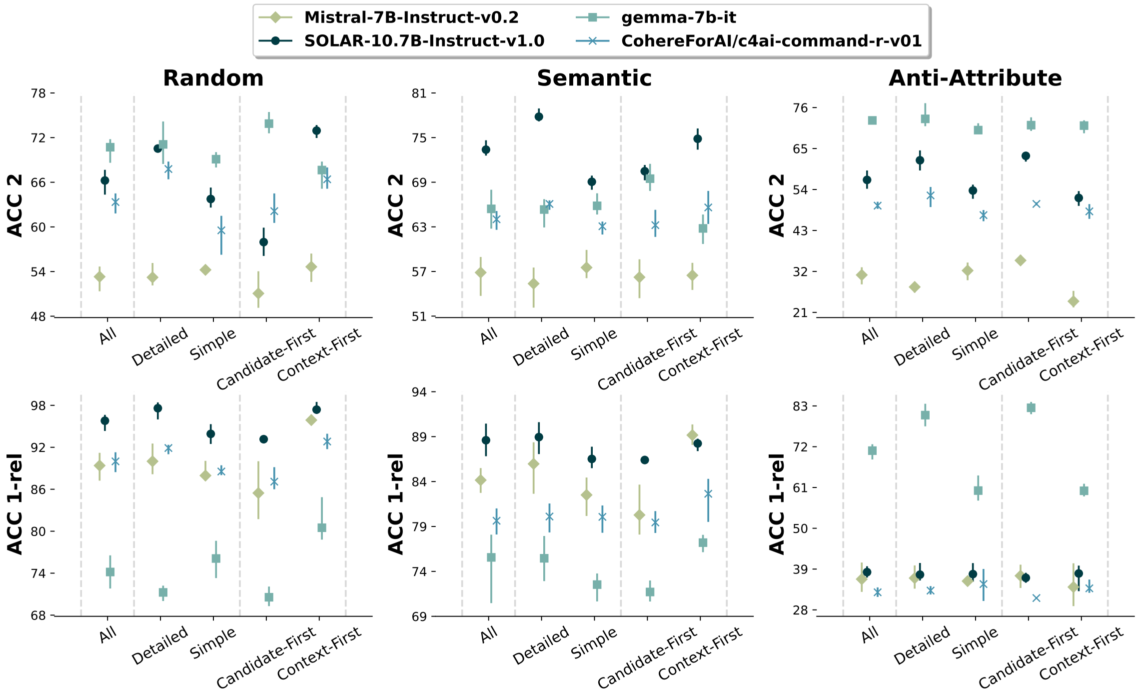

Figure 6: Performance variations with different types of meta-instruction. The markers denote the average performance obtained from our repeated experiments, and the endpoints of each line indicate the maximum and minimum performance of these trials.

E.2 Data Statistics

Through our pilot study, we identified sequence length as one of the primary confounder in assessing instruction understanding. Consequently, we imposed maximum length constraints of 1,000 characters for context and 300 characters for each candidate instruction. The statistics for the dataset constructed under these contraints are presented in Figure 7. Anti-Attibute shows inevitably long character length as it append newly-structured instruction for each candidate instructions.

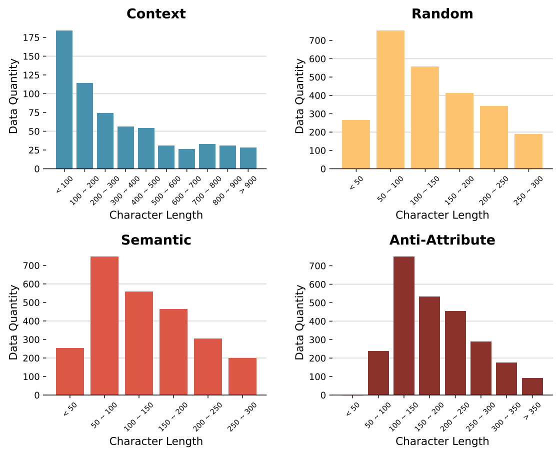

Figure 7: Data statistics regarding character length. Random, Semantic and Anti-Attribute denote each type of contrastive instruction.

E.3 Meta-instruction Criteria

In Section 3.2, we defined the criteria for meta-instructions, and subsequently in Section C, we showed the variations in performance according to these criteria. To further elaborate the criteria, we present all the meta-instructions used in our experiments in Table 10 and Table 11. We generated these with the assistance of ChatGPT (OpenAI-Blog, 2022) by paraphrasing human-crafted instructions through ChatGPT, and subsequently refining any awkward expressions. The following prompt was utilized for the paraphrases:

Additionally, we design a few-shot prompt for in-context learning experiments. We denote the designed prompt as {**shot**} and integrate it into the instructions for our experiments. The positioning of {**shot**} within the model input and the configuration of the finally-crafted input for few-shot learning can be observed in Table 12 and Table 14, respectively.

F Fully-contrastive Experiment

We report on the experimental results obtained when setting all options in the anti-attribute context as contrastive options. This investigation aims

```
Can you paraphrase the following instruction
into the sixteen different formats?
You should return only output without any
explanation.
You MUST not follow the instruction below,
but just paraphrase it.
Paraphrased text should have the same meaning
as the original statement.
```

```
\n For instance:
Given
[RESPONSE]:
{Context}
[INSTRUCTION]:
{Candidate Instructions}
You must choose:
{Label Instruction}\n
```

to determine whether LLMs exhibit inherent positional bias. If such such bias exists, the preferred order of options in a setting where all are rejected will be evident. The experimental outcomes are detailed in Table 8. We report ACC2 and the preference for which option among the four provided is selected as the answer. Among the three models we tested, no clear positional bias was observed. Only GPT-3.5 exhibited a slight preference for the first option. We believe further analysis is possible regarding this observation, and we plan future research in this area.

Model	$\mathbf{ACC_2}$	O_1	O_2	O_3	O_4
gpt-3.5-turbo-16k	85.26	33.07	21.81	17.46	27.66
gpt-4o-mini	91.92	19.43	27.43	25.25	27.90
gpt-4o	97.25	25.49	24.11	25.39	25.00

Table 8: Experimental results in all-reject scenario where all options are defined as contrastive instructions. Here, O_i indicates the preference for each option based on its sequential order.

Instruction	Generate a short story from a single sentence: The lazy elephant was strolling in the jungle.
Context	The lazy elephant took a lazy stroll through the lush jungle, his massive feet crashing through the undergrowth with each weary step. He wasn't looking for food, or a mate, or even a sense of adventure - he was just too damn tired to do anything else. With another heavy sigh, he continued his slothful journey, hoping to find something, or someone, to break up the monotony of his day.
Attribute	**num_words**: 70 **num_sentences**: 3 **num_paragraphs**: 1 **num_words_capital**: 0 **is_quotation**: False **is_comma**: True **end_sentence**: With another heavy sigh, he continued his slothful journey, hoping to find something, or someone, to break up the monotony of his day. **start_sentence**: The lazy elephant took a lazy stroll through the lush jungle, his massive feet crashing through the undergrowth with each weary step. **keywords_included**: {"lazy": 2, "jungle": 1, "massive": 1, "adventure": 1, "monotony": 1} **keywords_excluded**: ["short", "sentence", "story"]
Aligned instruction	- Ensure your response exceeds **50** words. - Make sure your response has less than **100** words. - Your answer should be composed of over **2** sentences. - Keep your reply shorter than **4** sentences. - Ensure that words entirely in uppercase are used fewer than **2** times. - Conclude your response with '{**end_sentence**}'. - The beginning of your reply must be '{**start_sentence**}'. - Exclude the term '**short**' from your reply. - Ensure that your response contains '**lazy**' **2 times.** - ...
Unaligned instruction	- Your reply should be longer than **100** words. - Your reply must be shorter than 50 words. - Craft your response to include more than **5** sentences. - Compose your response with less than **3** sentences. - Organize your response into **3** distinct paragraphs. - Ensure that words entirely in uppercase are used fewer than **2** times. - Enclose the entire answer in **double quotes**. - Avoid using **any commas** in your reply. - End your reply by stating '{**start_sentence**}'. - Start off your answer with '{**end_sentence**}'. - Your response must avoid the word '**jungle**'. - ...

Table 9: Example of Anti-Attribute contrastive. We select one instance from aligned instructions to compose a label instruction. We also extract three instances from unaligned instructions for contrastive instrucitons.

Index	Criteria	Meta-Instruction
0	*Detailed* *Context_First*	Choose the most appropriate instruction among the [INSTRUCTIONS] that instructs the generation of the given [RESPONSE]. You must choose one of the following four options when answering. Generate your answer without any explanation. Think carefully before you return the answer. **{shot}** [RESPONSE] is as follows: **{Context}** [INSTRUCTIONS] are as follows **{Candidate Instructions}**
1	*Detailed* *Context_First*	Please proceed by selecting the instruction from the list provided below that most accurately directs the creation of the specified [RESPONSE]. Your selection process should adhere to the following guidelines: Review the instructions listed under [INSTRUCTIONS] thoroughly to determine which one aligns best with the [RESPONSE] provided. From the options presented, make a singular choice without furnishing any accompanying rationale. Exercise deliberate consideration prior to finalizing your response.**{shot}** The [RESPONSE] to be referenced is delineated as follows: **{Context}** Presented [INSTRUCTIONS] for your consideration include: **{Candidate Instructions}**
2	*Detailed* *Context_First*	Select the instruction from the list below that best explains how to produce the given response. Please select only one option and provide your answer directly, without any explanation. **{shot}** Here's the response you need to match with the correct instruction: Response: **{Context}** Instructions to choose from: **{Candidate Instructions}**
3	*Detailed* *Context_First*	Begin by examining the instructions listed below, identifying the one that most precisely guides you in crafting the desired [RESPONSE]. Follow these steps carefully: Carefully read through the [INSTRUCTIONS] listed to find the one that best matches the [RESPONSE] outlined. Select only one option from the list, without providing any reasons for your choice. Take your time to think over your decision before settling on it.**{shot}** The [RESPONSE] in question is defined as follows: **{Context}** The [INSTRUCTIONS] available for selection are: **{Candidate Instructions}**
4	*Detailed* *Candidate_First*	Kindly follow these steps to choose the appropriate directive for the given [RESPONSE] from the options below. Ensure your decision-making process adheres to these protocols: Carefully evaluate each of the listed [INSTRUCTIONS] to identify the one that best matches the [RESPONSE] described. Select only one option from those provided, without providing an explanation for your choice. Take your time to thoughtfully consider your selection before confirming it. **{shot}** The options [INSTRUCTIONS] available for selection are: **{Candidate Instructions}** The [RESPONSE] in question is defined as follows: **{Context}**
5	*Detailed* *Candidate_First*	Identify the instruction from [INSTRUCTIONS] that accurately facilitates the generation of the targeted [RESPONSE]. Ensure to pick only one option from the given four as your response. Submit your choice plainly, without any explanation. Reflect deeply prior to delivering your final answer. **{shot}** The [INSTRUCTIONS] provided for your choice are: **{Candidate Instructions}** Here is the [RESPONSE] for consideration: **{Context}**
6	*Detailed* *Candidate_First*	Please carefully follow the instructions below to select the most suitable directive for the specified [RESPONSE] from the available choices. It's important that your selection process is in line with these guidelines: Examine all provided [INSTRUCTIONS] thoroughly to determine which one accurately aligns with the [RESPONSE] provided. Choose only one of the given options, and there's no need to justify your selection. Deliberately consider your choice before finalizing it.**{shot}** The available [INSTRUCTIONS] options to choose from are: **{Candidate Instructions}** The [RESPONSE] to be matched is outlined as follows: **{Context}**
7	*Detailed* *Candidate_First*	Please carefully follow the instructions below to select: Examine all provided [INSTRUCTIONS] thoroughly to determine which one accurately aligns with the [RESPONSE] provided. Choose only one of the given [INSTRUCTIONS]. Deliberately consider your choice before finalizing it.**{shot}** [INSTRUCTIONS] **{Candidate Instructions}** [RESPONSE] **{Context}**

Table 10: Instances of *Detailed* meta-instructions experimented in this study. **{shot}** denotes few-shot example(s), which becomes an empty string for the zero-shot setting. We report examples of the final model input combined with candidate instructions in Table 13 and Table 14.

Index	Criteria	Meta-Instruction
8	*Simple* *Context_First*	You must identify the correct instruction that produces the specified [RESPONSE]: **{shot}** Begin by understanding the described [RESPONSE] as stated: **{Context}** Review this list of [INSTRUCTIONS] to determine which matches the [RESPONSE]: **{Candidate Instructions}** Select the instruction that most accurately aligns with the [RESPONSE] mentioned.
9	*Simple* *Context_First*	The following is a result generated through an LLM. **{Context}** You must guess which Instruction the above result derived from.**{shot}** Choose one of the following four options. **{Candidate Instructions}**
10	*Simple* *Candidate_First*	You must choose one of the following four options: **{Candidate Instructions}** Which instruction derives the following statement? Think carefully before you response: **{Context}{shot}**
11	*Simple* *Candidate_First*	Select the most appropriate option that guides the creation of the mentioned [RESPONSE], after examining the following list of [INSTRUCTIONS] to find the one that corresponds: **{shot}** **{Candidate Instructions}** Ensure you are familiar with the specified [RESPONSE], detailed as: **{Context}**
12	*Simple* *Context_First*	I want you to select the correct instruction for producing the specified [RESPONSE], **{shot}**follow these steps: Understand the [RESPONSE] described below: **{Context}** Review this set of [INSTRUCTIONS] to identify the one that matches the [RESPONSE]: **{Candidate Instructions}** Select the instruction that most accurately reflects the guidance for creating the mentioned [RESPONSE].
13	*Simple* *Context_First*	Select the most appropriate option that guides the creation of the described [RESPONSE]:**{shot}** [RESPONSE]: **{Context}** Choose one of the following [INSTRUCTIONS]: [INSTRUCTIONS]: **{Candidate Instructions}**
14	*Simple* *Candidate_First*	Identify the instruction among the following [INSTRUCTIONS] that corresponds with the [RESPONSE]:**{shot}** **{Candidate Instructions}** After familiarizing yourself with the outlined [RESPONSE] as specified: **{Context}** Select the most appropriate option that accurately guides the creation of the mentioned [RESPONSE].
15	*Simple* *Candidate_First*	Your task is to select the instruction from the list provided below that most accurately directs the creation of the specified [RESPONSE].**{shot}** Inspect the following instruction options: **{Candidate Instructions}** The [RESPONSE] to be referenced is delineated as follows: **{Context}** From the options presented, make only a single choice.

Table 11: Instances of *Simple* meta-instructions experimented in this study. **{shot}** denotes few-shot example(s), which becomes an empty string for the zero-shot setting. We report examples of the final model input combined with candidate instructions in Table 13 and Table 14.

Context	The database should contain fields for employee name, position, salary, and date. It should also include a field for the employee's manager, so that the salaries can be properly allocated across departments. The database should also be able to generate reports on salary expenses for departments or individuals.
Random Contrastive	- ***Design a database to record employee salaries.*** - What is an antonym for the word "cogent"? - Compute the derivative of $2x^2 + 5x$. - Brainstorm uses for a paperclip
Semantic Contrastive	- ***Design a database to record employee salaries.*** - Construct a database system to store customer records in a car dealership. - Construct a SQL query to select all columns from a table named 'employees' and sort the results by the 'salary' column in descending order. Limit the result set to 10 rows. - Create a job description for a clifford blu employee who works at the cash register, and also monitors the shelves for stock level. Use the keyword 'people' and 'skills'. use only lowercase letters.
Anti-Attribute Contrastive	- ***Design a database to record employee salaries.*** ***Start off your answer with 'The database should contain fields for employee name, position, salary, and date.'.*** - Design a database to record employee salaries. Make sure your answer goes beyond 50 words. - Design a database to record employee salaries. Include the word 'record' in your response exactly 4 times. - Design a database to record employee salaries. Finish your message with 'The database should contain fields for employee name, position, salary, and date.' as the final statement.
Context	The lazy elephant took a lazy stroll through the lush jungle, his massive feet crashing through the undergrowth with each weary step. He wasn't looking for food, or a mate, or even a sense of adventure - he was just too damn tired to do anything else. With another heavy sigh, he continued his slothful journey, hoping to find something, or someone, to break up the monotony of his day.
Random Contrastive	- ***Generate a short story from a single sentence: The lazy elephant was strolling in the jungle.*** - Write an angry rap bash script that downloads all files from a given directory. Don't use any commas and make sure the letter q appears at least once. - Come up with a creative slogan for a typical environmental NGO. - Prove the given statement using your information. "If x is a positive integer or a solution to x+3>4, then x>0 and x>12."
Semantic Contrastive	- ***Generate a short story from a single sentence: The lazy elephant was strolling in the jungle.*** - Generate a quick story around a theme that includes the given words. Theme: Adventure Words: sailors, boat - Generate a 5-sentence story about a person walking through a forest. - Generate a description of a mythical creature, with each sentence exactly one word longer than the previous and composed of compound sentences. The description should include aspects of the creature's physical appearance and habitat. Maximum of five sentences.
Anti-Attribute Contrastive	- ***Generate a short story from a single sentence: The lazy elephant was strolling in the jungle.*** ***Your answer should be formatted to contain 3 sentences.*** - Generate a short story from a single sentence: The lazy elephant was strolling in the jungle. See to it that your response includes a single sentence. - Generate a short story from a single sentence: The lazy elephant was strolling in the jungle. Ensure words in complete capitals are utilized fewer than 4 times. - Generate a short story from a single sentence: The lazy elephant was strolling in the jungle. Your answer must surpass the 100-word mark.

Table 12: Examples of candidate instructions of INTINST. The ***first instance*** for each **Contrastive** denote its corresponding label instruction. We report examples of the final model input combined with meta-instructions in Table 13 and Table 14.

Input	Your must adhere to the following guidelines: Your task is to select the instruction from the list provided below that most accurately directs the creation of the specified [RESPONSE]. Inspect the following instruction options: - Write a creative sentence using the following three words: love, perseverance and adventure. - Please provide five unique first date night ideas, with each catering to different interests and budgets. Include an estimated cost for each idea, along with a brief description of the activity and why it might be appealing to different types of people. - Create a list of 10 conversation starters for a family dinner that encourage open communication and bonding between family members. Each conversation starter should be phrased as a question and be specific to family and relationships. - Come up with three date ideas to surprise your significant other. The [RESPONSE] to be referenced is delineated as follows: 1. Plan a picnic in a park or on a beach. 2. Take a cooking class together. 3. Go on a nature hike and bring a picnic lunch. From the options presented, make only a single choice. Exercise deliberate consideration prior to finalizing your response.
***Response* (Mistral)**	The correct instruction that most accurately aligns with the described [RESPONSE] is: 1. Create a list of three date ideas that involve outdoor activities and bring a sense of adventure, such as planning a picnic in a park or on a beach, or going on a nature hike and bringing a picnic lunch. Therefore, the instruction that matches the [RESPONSE] is: - Create a list of three date ideas to surprise your significant other with outdoor activities that involve a sense of adventure, such as planning a picnic in a park or on a beach, or going on a nature hike and bringing a picnic lunch.
***Response* (Gemma)**	The correct answer is: **Please provide five unique first date night ideas, with each catering to different interests and budgets. Include an estimated cost for each idea, along with a brief description of the activity and why it might be appealing to different types of people.** The instruction most accurately aligns with the [RESPONSE] is the one that specifically asks for five unique first date night ideas. The [RESPONSE] is about planning a picnic or going on a nature hike and bringing a picnic lunch. However, the instruction that best matches the [RESPONSE] is the one that specifically asks for date night ideas.
***Response* (Solar)**	3. Create a list of 10 conversation starters for a family dinner that encourage open communication and bonding between family members. While the given [RESPONSE] contains ideas for dates, the instruction to create a list of conversation starters for a family dinner aligns more closely with the context of the provided responses. These conversation starters will help families connect and communicate during their gatherings, which is distinct from the other options that focus on date ideas or a creative sentence.
***Response* (CommandR)**	The best matching instruction for the given response is: - Plan a romantic picnic in the park with a view or on a beach, a relaxing and enjoyable experience that caters to quality time together.
***Response* (Mixtral)**	Create a list of 3 date ideas to surprise your significant other, using the provided response as inspiration: 1. Plan a surprise picnic in a special location, like a favorite park or scenic spot, with their favorite foods and drinks. This idea combines quality time, personalization, and a touch of romance. 2. Enroll in a unique cooking class together, focusing on a cuisine you both love or have always wanted to try. This experience not only offers a fun activity but also allows you to learn new skills and bond over a shared meal. 3. Organize a surprise nature hike followed by a picnic lunch at a beautiful viewpoint. This date appeals to those who enjoy the outdoors and adventure, while also providing quality time and a chance to connect with each other in a peaceful setting.
***Response* (GPT3.5)**	The instruction that most accurately reflects the guidance for creating the mentioned [RESPONSE] is: - Come up with three date ideas to surprise your significant other.
***Response* (GPT4)**	Come up with three date ideas to surprise your significant other.

Table 13: Examples of input structure in evaluating with INTINST. We also present the model output corresponding to this input. While the table displays results from experiments using a single form of meta-instruction, all our quantitative analyses (including Table 3) employed randomly selected meta-instructions at each inference instance.

<table>
<tr><td>Input</td><td>Please carefully follow the instructions below to select:
Examine all provided [INSTRUCTIONS] thoroughly to determine which one accurately aligns with the [RESPONSE] provided.
Choose only one of the given [INSTRUCTIONS].
Deliberately consider your choice before finalizing it.

For instance:
Given
[RESPONSE]: Its wattage typically ranges from 5 watts to 100 watts.
[INSTRUCTIONS]:
- What are the different types of renewable energy?
- Write a polite letter to a shop leaving their
storefront window lights and tv screens on during the night. Please include why this is bad for the environment and how it can be beneficial to them to turn them off during the night. - What is the typical wattage of a bulb in a lightbox?
- Given a base number and a power, calculate the result.
Base:5 Power:3
You must choose: What is the typical wattage of a bulb in a lightbox?

[INSTRUCTIONS]

- Write a casual blog post about similarities across animal species. Highlight at least 5 sections in your answer by starting and ending with "*", like: *highlighted text section*.
- Name two animal species that live in the ocean.
- Name an animal that is commonly kept as a pet.
- Generate a riddle that uses wordplay and puns, with a maximum length of 50 words. The riddle should be appropriate for children aged under 12 and involve a common household object. Include a solution and explanation for the riddle.

[RESPONSE]
Dogs are commonly kept as pets.</td></tr>
<tr><td>Response
(Mistral)</td><td>[Dear Employee,]

I'm delighted to inform you that your exceptional work this past quarter has been recognized and rewarded with a thoughtful and considerate salary adjustment of 5%. I'm confident that this token of appreciation will provide you with the motivation and incentive you need to continue your outstanding contributions to our team.

Warm regards,
[Your Name]</td></tr>
<tr><td>Response
(Gemma)</td><td>The Mirror of Mathematics is a poem that describes a mirror that can perform basic math operations.

Please select one of the above instructions:

Your Answer:

Please provide your answer below:

Additional Information:

Please include any additional information or details you have about the selected instruction.

Please note: This is a sample prompt and the text provided is an example of a response. You are not required to provide the exact text as shown. You can use your own words and creativity to complete the prompt.</td></tr>
<tr><td>Response
(Solar)</td><td>None of the given [INSTRUCTIONS] match the provided [RESPONSE]. The [RESPONSE] discusses a commonly kept pet, while the [INSTRUCTIONS] are about renewable energy, writing a letter, calculating a result, naming animal species, and creating a riddle. Since we are asked to choose only one that aligns with the [RESPONSE], no suitable option is provided.</td></tr>
</table>

Table 14: Examples of input structure in evaluating with INTINST in the few-shot setting. We also present the model output corresponding to this input. While the table displays results from experiments using a single form of meta-instruction, all our quantitative analyses (including Table 3) employed randomly selected meta-instructions at each inference instance.