FactCHD: Benchmarking Fact-Conflicting Hallucination Detection

Xiang Chen\textsuperscript{1,4}, Duanzheng Song\textsuperscript{2}, Honghao Gui\textsuperscript{1,4}, Chenxi Wang\textsuperscript{2,4}, Ningyu Zhang\textsuperscript{2,4\*}, Yong Jiang\textsuperscript{3}, Fei Huang\textsuperscript{3}, Chengfei Lyu\textsuperscript{3}, Dan Zhang\textsuperscript{2}, Huajun Chen\textsuperscript{1,4\*}

\textsuperscript{1}College of Computer Science and Technology, Zhejiang University
\textsuperscript{2}School of Software Technology, Zhejiang University
\textsuperscript{3}Alibaba Group
\textsuperscript{4}Zhejiang University - Ant Group Joint Research Center for Knowledge Graphs

\{xiang_chen, sdz, guihonghao, sunnywcx, zhangningyu, dan.zhang, huajunsir\}@zju.edu.cn
\{yongjiang.jy, f.huang, chengfei.lcf\}@alibaba-inc.com

Abstract

Despite their impressive generative capabilities, LLMs are hindered by fact-conflicting hallucinations in real-world applications. The accurate identification of hallucinations in texts generated by LLMs, especially in complex inferential scenarios, is a relatively unexplored area. To address this gap, we present FACTCHD, a dedicated benchmark designed for the detection of fact-conflicting hallucinations from LLMs. FACTCHD features a diverse dataset that spans various factuality patterns, including vanilla, multi-hop, comparison, and set operations. A distinctive element of FACTCHD is its integration of fact-based evidence chains, significantly enhancing the depth of evaluating the detectors’ explanations. Experiments on different LLMs expose the shortcomings of current approaches in detecting factual errors accurately. Furthermore, we introduce TRUTH-TRIANGULATOR that synthesizes reflective considerations by tool-enhanced ChatGPT and LoRA-tuning based on Llama2, aiming to yield more credible detection through the amalgamation of predictive results and evidence.

1 Introduction

Large Language Models (LLMs) [Zhao et al., 2023] are susceptible to generating text that, while seemingly credible, can be factually inaccurate or vague, leading to the spread of misinformation online [Yin et al., 2023; Ji et al., 2023; Huang et al., 2023]. This issue, referred to as \textit{fact-conflicting hallucination} [Zhang et al., 2023], arises from the incorporation of incorrect or obsolete knowledge into the models’ parameters and from the models’ inherent limitations in complex cognitive ability. These shortcomings constrain LLMs’ deployment in critical domains like finance, healthcare, and law, and amplify the propagation of erroneous information. Therefore, it is crucial to effectively detect fact-conflicting hallucinations for mitigating or editing them [Yao et al., 2023].

However, traditional fact verification tasks [Wadden et al., 2020a; Wadden et al., 2020b] are not suitable for LLM-based “QUERY-RESPONSE” data. Moreover, existing hallucination evaluation benchmarks [Li et al., 2023; Muhlgay et al., 2023], predominantly centering on vanilla facts and textual content, lack in-depth exploration of complex operations among facts, thus rendering their coverage against fact-conflicting hallucinations suboptimal. To bridge this gap, we introduce an inherently rigorous [Wang et al., 2023b], yet authentic, task scenario: \textit{fact-conflicting hallucination detection}, devoid of explicit claims or evidence. As shown in Figure 1, when confronted with a query and its generated response, detectors are impelled to harness both their intrinsic knowledge and external resources, while rendering a factual judgment accompanied by an elucidative explanation.

Figure 1: Illustration of fact-conflicting hallucination detection example from FACTCHD, where the green part represents factual explanation core (body part) in the chain of evidence.

In paving the way for future strides in hallucination evaluation, we introduce a new benchmark, Fact-Conflicting Hallucination Detection (FACTCHD), tailored for LLMs and encompassing a variegated array of factuality patterns, including Vanilla, Multi-hops, Comparison, and Set-Operation patterns. For example in Figure 1, querying whether the “Alps” or “Pyrenees” are higher, exemplifies a comparison pattern, assessing the relative relationships among facts. Drawing inspiration from the adage “to know it and to know the reason why of it” by Zhuzi, FACTCHD extends beyond mere
“QUERY-RESPONSE” labeling of hallucinations, incorporating golden chains of evidence to assess if detectors can provide coherent explanations for factualness judgment. Acknowledging the challenges of collecting comprehensive data through exhaustive human annotation, we propose a scalable data construction approach that harnesses existing knowledge graphs (KGs) and textual knowledge to create simulated hallucination instances with ChatGPT, verified with human annotation, for the efficient development of FactCHD.

We evaluate the performance of various LLMs (such as Alpaca, Llama2-chat, and ChatGPT) using our FactCHD benchmark across multiple settings: zero-shot, in-context learning, specialized detection tuning, and knowledge enhancement via retrieval/tools. The results indicate that specialized detection tuning and knowledge enhancement notably improve the detection of fact-conflicting hallucinations. Additionally, we present Truth-Triangulator framework grounded in “Triangulation” theory [Valenza, 2016]. This system comprises three roles: ChatGPT, enhanced with tools as the Truth Seeker; a detect-specific expert based on Llama2-7B-LoRA as the Truth Guardian; and the Fact Verdict Manager, which amasses evidence from another role to fortify the reliability and accuracy of the derived conclusions. Truth-Triangulator emphasizes the use of cross-referencing generators to astutely evaluate and adjudicate responses with potential factual discrepancies. Key insights are summarized as:

- We present FactCHD\textsuperscript{1}, a large-scale, multi-domain evaluation benchmark with diverse factual patterns and interpretable evidence chains, setting a new standard for detecting fact-conflicting hallucinations from LLMs.
- We introduce a scalable data construction strategy leveraging KGs, etc. to efficiently develop a fact-conflicting hallucination dataset, offering stronger applicability due to the authentic, broad domain coverage of KGs.
- We devise a triangulation-based framework Truth-Triangulator that employs cross-referencing generators for verifying LLM responses.

\section{Related Work}

\textbf{Hallucination in LLMs.} Despite LLMs [Zhao et al., 2023; Chen et al., 2022; OpenAI, 2022] like ChatGPT demonstrating remarkable understanding and execution of user instructions, they are prone to confidently generating misleading “hallucinations” [Ji et al., 2023; Wang et al., 2023a]. These hallucinations, as categorized by [Zhang et al., 2023], can be input-conflicting, context-conflicting, or fact-conflicting—with the latter being especially problematic due to the propagation of inaccurate factual information online. While previous studies have extensively examined hallucinations within natural language generation (NLG) for various NLP tasks [Shuster et al., 2021; Creswell and Shahahan, 2022; Das et al., 2023; Mallen et al., 2022], HaluEval [Li et al., 2023] has emerged as a recent benchmark for assessing LLMs’ recognition of such errors. Different from HaluEval which only analyzes ChatGPT’s ability to evaluate whether hallucinatory, we specifically focus on the evaluation of fact-conflicting hallucination by constructing an interpretable benchmark that can serve as a public platform for checking the factual errors of context with explanation.

\textbf{Factuality Detection.} Our research is also related to prior works on fact verification in NLP tasks [Thorne et al., 2018; Wadden et al., 2020b; Gupta et al., 2022; Dziri et al., 2022; Liang et al., 2023; Rashkiy et al., 2021; Kryściński et al., 2020; Kang et al., 2024], expanding the understanding of factuality beyond binary judgments. The FRANK framework [Pagnoni et al., 2021] offers a detailed typology of factual errors, while [Dhingra et al., 2019] explores lexical entailment in text generation. The FACTOR score by [Muhlglay et al., 2023] evaluates LMs based on the likelihood of factual content. Existing benchmarks often miss the “QUERY-RESPONSE” context of LLMs, focusing solely on accuracy without providing explanatory rationales. Our contribution lies in: (1) presenting “QUERY-RESPONSE” reformatted data for LLMs; (2) introducing metrics for interpretability in detecting fact-conflicting hallucinations; and (3) validating the effectiveness of Truth-Triangulator through cross-referential verification from multiple sources. The comparison with other datasets is detailed in Table 1.

\section{Preliminaries}

\textbf{Task Formulation.} Detecting fact-conflicting hallucinations in LLMs involves discerning factual errors in responses to human queries. A comprehensive detector must not only classify responses as factual or non-factual but also provide explanations for its judgments. We define the task as follows: \textbf{Input:} A question $Q$ paired with an LLM-generated response $R$,
which may contain various fact conflicts. **Output**: A combined label and explanation sequence \( A = [l, e] \), where \( l \) is the binary factuality label (FACTUAL or NON-FACTUAL), and \( e \) articulates the rationale behind the assigned label. We evaluate the quality of \( e \) using the \( \text{ExpMatch} \) metric through the golden evidence chains in FACTCHD.

**Factuality Patterns.** We aim to explore distinct patterns of factual errors in our FACTCHD, vividly illustrated in Figure 2. These include the (1) vanilla pattern dealing with factual errors that can be objectively verified using established sources, the (2) multi-hops pattern involving the process of concluding by connecting multiple pieces of facts, the (3) comparison pattern referring to the act of evaluating and comparing relative worth and relations between different pieces of facts, and the (4) set-operation pattern involving manipulating and combining sets of elements using operations to analyze relations between different facts. We generated corresponding “QUERY-RESPONSE” examples based on these patterns.

**Knowledge-Driven Factual Foundation for Reliable “QUERY-RESPONSE” Generation.** KGs [Liang et al., 2024] serve as a rich trove of structured entities and relations, ideal for compositional reasoning and anchoring factual data. Alongside this, textual knowledge is critical for nuanced inference beyond basic facts. Our research focuses on collecting existing knowledge and integrating it into prompts as a factual foundation for “QUERY-RESPONSE” and chain of evidence generation, as outlined below: (1) We use 438 widespread relations from Wikidata [Vrandecic and Krötzsch, 2014] and PrimeKG [Chandak et al., 2023] to create varied subgraphs via \( K \)-hop walks, forming the knowledge base for generating “QUERY-RESPONSE” examples with multi-hop reasoning, fact comparison, and set operation patterns. (2) We utilize text knowledge from datasets such as FEVER [Thorne et al., 2018], Climate-Fever [Diggelmann et al., 2020], Health-Fever [Sarrouti et al., 2021], COVID-FACT [Saakyan et al., 2021], and SCIFACT [Wadden et al., 2020a] for generating examples with vanilla pattern.

### 4 FACTCHD Benchmark Construction

Building on the aforementioned preliminaries, we develop FACTCHD, a dataset containing a wealth of training instances and an additional 6,960 carefully selected samples for evaluating fact-conflicting hallucinations from LLMs. Our dataset maintains a balanced representation of FACTUAL and NON-FACTUAL categories, offering a robust framework for assessment. The statistics and domain distribution of FACTCHD are depicted in Table 2 and Figure 3. Next, we outline the design principles of our benchmark as follows.

### 4.1 Collect Realistic Data as Demonstration

Subsequently, adhering to the defined factuality patterns, we manually craft corresponding queries and utilize open-source LLMs (such as ChatGLM, Alpaca, and Vicuna) to generate authentic responses. By manually annotating these responses for hallucinations, we acquire “QUERY-RESPONSE” examples annotated with hallucination presence. Our objective is to curate a robust dataset of hallucination cases that closely emulate real-world examples, providing a demonstrative foundation and establishing standards for prompt refinement to ensure alignment between generated hallucinations and these demonstrations.
4.2 Generate “QUERY-RESPONSE” Contexts

Generate “QUERY-RESPONSE” based on Knowledge with ChatGPT. We incorporate the knowledge collected on distinct factuality patterns into customized prompts, specifying whether to generate factual or non-factual responses. This guides ChatGPT in producing “QUERY-RESPONSE” instances across various factuality categories with golden labels.

Refine Prompts with Consistency. In the initial phase of generating “QUERY-RESPONSE” instances, we assess the consistency of five samples per pattern against demonstrations, using majority consensus. We subjectively evaluate each context’s adherence to the style and form of realistic demonstrations, employing iterative refinements of prompts to ensure that the generated data exhibits realistic patterns. We aim for an at least 95% consistency rate before scaling up the generation of “QUERY-RESPONSE” instances.

4.3 Automated Semantic Diversity Screening

To increase “QUERY-RESPONSE” context diversity, we employ Sentence-BERT (SBERT) [Reimers and Gurevych, 2019] to automatically compute semantic similarity matrices, filtering out near-duplicate samples to preserve dataset variety. During filtration, we removed 1,542 training and 832 test samples, guaranteeing a diverse final dataset. This careful pruning of semantically similar entries promotes a varied collection of queries and responses, bolstering the benchmark’s effectiveness for assessing diverse instances.

4.4 Generate Chain of Evidence

Our benchmark evaluates the detectors’ ability to not only identify hallucinations but also to provide effective explanations. It necessitates ChatGPT’s generation of coherent golden evidence chains grounded in factual knowledge for substantiating judgments. Utilizing subgraph or textual facts outlined in §section 3, along with the previously generated “QUERY-RESPONSE” pairs, ChatGPT delivers thorough justifications for labels assigned to “QUERY-RESPONSE” contexts. These golden evidence chains are critical for assessing the explanatory validity of hallucination detectors.

4.5 Quality Filtering via Human Annotator

We craft filter rules to ensure pattern consistency, response factuality, and logical evidence chains for quality control in the annotation. Several educated annotators are uniformly trained and utilize both their expertise and search tools for rigorous sample vetting. To minimize subjectivity, we organized them into groups of three, incorporating a voting mechanism for the evaluation of data. Simultaneously judged mismatches by annotators led to sample discarding, resulting in final removal counts of 565 and 258 samples from the training and test sets, respectively. We involve Fleiss’s Kappa (κ) as a measure of inter-annotator agreement to assess the reliability of our annotations. We calculate κ over the remaining annotated test set, resulting in a κ value of 0.858, indicating substantial agreement.

5 Experiments

5.1 Metric Definition

**FACTCLS Metric.** We employ the FACTCLS, denoted by the Micro F1 score, to evaluate binary factuality classification performance. This metric focuses on the distribution p(I|Q&R), classifying instances as either FACTUAL or NON-FACTUAL. With a specific emphasis on identifying non-factual examples, we designate NON-FACTUAL as the positive class and FACTUAL as the negative class.

**EXPMatch Metric.** In FACTCHD, the golden evidence chain features introductory/expository statements (head-tail part) and a factual explanation core (body part). The former contextualizes with phrases such as ‘Therefore, there is an incorrect conclusion in this query and response’, and the latter delivers the in-depth, fact-based reasoning process. For hallucination detectors, prompts guide their outputs akin to the gold evidence chain for quality assessment. Given the paramount importance of aligning factual explanations over expository parts, we introduce EXPMATCH, a metric employing segmented matching with weighted averaging. It computes Score_{ed} via span-based Micro F1 for unigram overlap between generated and reference bodies, and Score_{et} via ROUGE-L...
to assess similarity based on the longest common word subsequence for the head-to-tail part. \textsc{Exp\_Match}, combining Score\textsubscript{bd} and Score\textsubscript{bt}, evaluates the explanations generated from detectors as:

\[
\text{Exp\_Match} = \alpha \times \text{Score}_{bd} + (1 - \alpha) \times \text{Score}_{bt}.
\]

We initially set $\alpha$ to 0.7, emphasizing the body part$^2$.

5.2 Experimental Settings

**Evaluation Models.** We evaluate various leading LLMs on FACTCHD benchmark, focusing on OpenAI API models, including text-davinci-003 (InstructGPT) and GPT-3.5-turbo (ChatGPT). Additionally, we explore the adoption of open-source models such as Llama2-chat, Alpaca [Taori et al., 2023] and Vicuna [Chiang et al., 2023], which are fine-tuned variants of the LLaMA [2023].

**Implementation Details.** Using Azure’s OpenAI ChatGPT API, we generate samples with a temperature of 1.0 to control the diversity of generated samples, while limiting the maximum number of tokens to 2048 to ensure concise responses. We use a frequency penalty of zero and a Top-p of 1.0 to ensure unrestricted token selection during generation. For evaluations, we standardize the temperature at 0.2 to minimize randomness.

**Baseline Strategy Settings** This study investigates the effectiveness of various baseline strategies in detecting fact-conflicting hallucinations. These strategies include: (1) \textsc{Zero-shot Learning}, which evaluates model performance without prior training in hallucination detection; (2) \textsc{In-context Learning}, using 4-shot samples as demonstrations to prompt the model and evaluate its ability to handle hallucination detection; (3) \textsc{Detect-Specific Expert Model}, fine-tuning the LoRA [Hu et al., 2022] parameters using our specialized training set to leverage domain expertise; (4) \textsc{Knowledge Enhancement}, enhancing the model’s capabilities by integrating external knowledge through retrieval techniques or tool-based enhancements. These experiments aim to explore both the potential and limitations of these baseline methods in hallucination detection.

5.3 Empirical Experiment Results

**Zero-shot Learning Performance**

The empirical results presented in Table 3 show that LLMs with zero-shot learning struggle to identify implicit factuality in “Query-Response” contexts. The performance of the open-source 7B LLMs in zero-shot learning is notably poor, indicating deficiencies in both instruction comprehension and internal knowledge representation. Even the ChatGPT shows limited proficiency in distinguishing between factual and non-factual “Query-Response” samples, achieving only a 52.82% FactCls score in zero-shot. Llama2-chat is susceptible to false negatives, resulting in reduced sensitivity in hallucination detection.

**In-context Learning Performance**

The incorporation of few-shot information significantly improves fact-conflicting hallucination detection in the GPT-3.5-turbo, Alpaca-7B, Vicuna-7B, and Llama2-7B-chat models. Compared to models operating without this additional context, these enhancements result in an average increase of approximately 16% in the FACTCLS and 18% in \textsc{Exp\_Match} scores. However, the impact of integrating few-shot information on text-davinci-003 is relatively modest, indicating its limited proficiency in managing few-shot learning. The

<table>
<thead>
<tr>
<th>Table 3: Results on FACTCLS and EXP_MATCH (abbreviated as CLS. and EXP.) along with FACTCHD estimated by each method. The shadow and shadow in each row represent the top-2 FACTCLS scores for the four factuality patterns. The up and down arrows respectively indicate positive/negative performance changes in the AVERAGE score compared to the corresponding upper-level method.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluator</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>GPT-3.5-turbo</td>
</tr>
<tr>
<td>text-davinci-003</td>
</tr>
<tr>
<td>Alpaca-7B</td>
</tr>
<tr>
<td>Vicuna-7B</td>
</tr>
<tr>
<td>Llama2-7B-chat</td>
</tr>
<tr>
<td>GPT-3.5-turbo (4-shot)</td>
</tr>
<tr>
<td>text-davinci-003 (4-shot)</td>
</tr>
<tr>
<td>Alpaca-7B-LoRA</td>
</tr>
<tr>
<td>Vicuna-7B-LoRA</td>
</tr>
<tr>
<td>Llama2-7B-chat-LoRA</td>
</tr>
<tr>
<td>Knowledge (tune)</td>
</tr>
<tr>
<td>GPT-3.5-turbo (tool)</td>
</tr>
<tr>
<td>Knowledge (tune)</td>
</tr>
<tr>
<td>TRUTH-TRIANGULATOR</td>
</tr>
</tbody>
</table>

$^2$Preliminary evaluations reveal that scores with $\alpha$ at 0.75 and 0.8 align with our established conclusions. If a generated result lacks factual or non-factual information, indicating a prediction failure, we attribute an \textsc{Exp\_Match} value of 0 to that particular example during metric calculation.
We investigate the effectiveness of tuning LLMs with LoRA for domain-specific expertise in detecting fact-conflicting hallucinations. By training these models on our trainset, we enhance their detection task proficiency. Our findings show that all tested open-source 7B models improve, with Llama2-chat outperforming its counterparts. After fine-tuning on hallucination data, Llama2-chat-7B achieves a FACT-CLS score of 74.73% and an EXPMATCH score of 53.71% on our benchmark. This success with a 7B model underscores the viability of using such models as hallucination detectors and encourages further exploration in this area.

**Knowledge Enhancement**

**Retrieval Enhancement.** This study enhances LLM-based detection with Wikipedia-sourced facts, using BM25 for initial document retrieval to select the top five most relevant paragraphs. Our experiments show that given the 7B model’s limited ability to process long texts, providing knowledge during fine-tuning modestly improves performance. However, compared to the detect-specific expert without knowledge augmentation, the improvement achieved by combining Wikipedia retrieval with fine-tuning the LoRA parameters is relatively modest. We attribute these observations to two primary factors: our dataset spans multiple domains while Wikipedia encompasses merely a subset, and our retrieval method being elementary, occasionally yielding lower-quality evidence.

**Tool Enhancement.** Considering the labor-intensive nature of tapping into external knowledge bases, we investigate leveraging ChatGPT’s advanced contextual understanding for tool-enhanced hallucination detection, drawing from prior research such as [Chern et al., 2023; Qin et al., 2023]. By using a search engine to inform ChatGPT’s analysis of “QUERY-RESPONSE” samples, we enable the model to identify hallucinations with supporting evidence from search results. Results in Table 3 demonstrate significant enhancements over ChatGPT’s few-shot performance, highlighting the efficacy of tool learning.

**5.4 Exploring Triangulation for Truth**

We categorize tool-enhanced ChatGPT as the **Truth Seeker**, which aims to make informed judgments by seeking external knowledge. However, the information returned by external knowledge sources may inevitably be incomplete, erroneous, or redundant, thus potentially misleading the large-scale model. On the other hand, the detect-specific expert as the **Truth Guardian** relies on its knowledge and expertise in the task, tending towards more conservative predictions. To address these challenges, we propose the Truth-Triangulator framework inspired by the “Triangulation for Truth” theory, involving verifying and confirming information by cross-referencing multiple independent perspectives.
**5.5 Experimental Analysis**

**Examining the Influences of Model Capacity.** Figure 6 illustrates that transitioning from 7B to 13B models notably improves the detection of fact-conflicting hallucinations, particularly in zero-shot and in-context learning scenarios. Alpaca-13B outperforms ChatGPT, which can be ascribed to the consistently adopted command prompt that may be more friendly to Alpaca-13B. Interestingly, when models are fine-tuned with training data, the impact of model capacity on performance improvement appears minimal. This implies that further training larger LLMs as hallucination detectors may yield limited benefits, and it is necessary to explore alternatives with higher upper limits for enhancing detection performance.

**Enhancement lies in Accurate Evidence.** We also explore the direct use of intrinsic facts from the dataset as golden evidence in the LoRA fine-tuning process. As shown in Table 5, integrating factual information (w/ gold evidence) leads to a significant improvement in the FACTCLS score, indicating that the modest improvement from retrieval may stem from the lower quality of the obtained evidence. This underscores the potential for substantial improvement by seeking the most accurate facts for evaluation.

**Complete “QUERY-RESPONSE” Context.** Our dataset differs from traditional fact-checking datasets in its inclusion of the “QUERY-RESPONSE” context. To examine its impact, we conduct experiments by excluding the “query” during the Alpaca-7B-LoRA fine-tuning process. The results in Table 5 reveal that the “w/o query” scenario leads to a significant 40% decrease in the FACTCLS score, highlighting the essential role of a comprehensive “QUERY-RESPONSE” context in extracting valuable information for informed decision-making.

**Real-World Application Case Study.** We extended the testing of TRUTH-TRIANGULATOR to real-world instances of ChatGPT-generated hallucinations beyond the FACTCHD benchmark to demonstrate its wide-ranging utility. The out-of-distribution case analysis, detailed in Table 4, delineates our model’s strengths and limitations. These real-world tests confirm TRUTH-TRIANGULATOR’s adeptness in making accurate assessments, particularly where expert and augmented ChatGPT evaluations diverge, thus bolstering the credibility of detecting fact-conflicting hallucinations in authentic, uncontrolled settings.

**6 Conclusion and Future Work**

We introduce FACTCHD, a meticulously designed benchmark tailored for the evaluation of fact-conflicting hallucinations from LLMs, which is notably enriched with a kaleidoscope of patterns and substantiated evidence chains to fortify the robust elucidation of factuality assessments. Moreover, we delineate TRUTH-TRIANGULATOR that employs the principle of triangulation to discern the veracity of information, deploying cross-referencing generators to arbitrate responses. Moving forward, we will broaden our evaluative scope to encompass various modalities and granularity in hallucination detection.
Acknowledgments

We would like to express gratitude to the anonymous reviewers for kind comments. This work was supported by the National Natural Science Foundation of China (No. 62206246), Response-driven intelligent enhanced control technology for AC/DC hybrid power grid with high proportion of new energy (5100-202154526A-0-0-0), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), CCF-Tencent Rhino-Bird Open Research Fund, and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References


[Yin et al., 2023] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don’t know? *CoRR*, abs/2305.18153, 2023.
