How does Misinformation Affect Large Language Model Behaviors and Preferences?

Miao Peng[1], Nuo Chen[1], Jianheng Tang[1], Jia Li[1,2]*
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
mpeng885@connect.hkust-gz.edu.cn, chennuo26@gmail.com
jtangbf@connect.ust.hk, jialee@ust.hk

Abstract

Large Language Models (LLMs) have shown remarkable capabilities in knowledge-intensive tasks, while they remain vulnerable when encountering misinformation. Existing studies have explored the role of LLMs in combating misinformation, but there is still a lack of fine-grained analysis on the specific aspects and extent to which LLMs are influenced by misinformation. To bridge this gap, we present MISBENCH, the current largest and most comprehensive benchmark for evaluating LLMs' behavior and knowledge preference toward misinformation. MISBENCH consists of **10,346,712** pieces of misinformation, which uniquely considers both knowledge-based conflicts and stylistic variations in misinformation. Empirical results reveal that while LLMs demonstrate comparable abilities in discerning misinformation, they still remain susceptible to knowledge conflicts and stylistic variations. Based on these findings, we further propose a novel approach called Reconstruct to Discriminate (**RtD**) to strengthen LLMs' ability to detect misinformation. Our study provides valuable insights into LLMs' interactions with misinformation, and we believe MISBENCH can serve as an effective benchmark for evaluating LLM-based detectors and enhancing their reliability in real-world applications. Codes and data are available at: https://github.com/GKNL/MisBench.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in understanding and reasoning with external knowledge (Ram et al., 2023; Yao et al., 2023; Tang et al., 2025; Ho et al., 2020; Chen et al., 2024). However, these powerful LLMs remain susceptible to misinformation, often producing erroneous answers when encountering inaccurate (Mallen et al., 2023), out-of-date (Cao et al., 2021), or fictional knowledge (Goldstein et al., 2023). This vulnerability to misinformation significantly impacts their real-world performance, undermining their reliability and trustworthiness in practical applications.

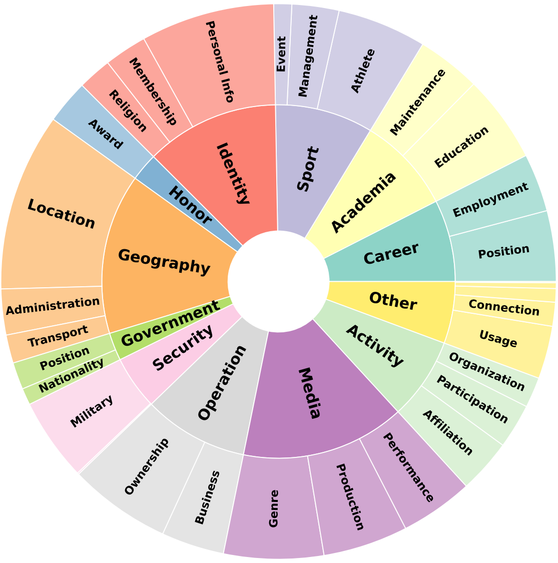


Figure 1: An overview of domains in MISBENCH.

Following the emergence of LLMs, researchers have established various benchmarks to investigate how misinformation affects these models, including LLMFake (Chen and Shu, 2024a), LLM-KC (Xie et al., 2024), ConflictBank (Su et al., 2024), Farm (Xu et al., 2024), and Misinfo-ODQA (Pan et al., 2023). While these studies have demonstrated LLMs' vulnerability to misinformation, a fundamental question remains unexplored: "**How and to what extent do LLMs get misled by misinformation?**" This further leads us to ask "**How do different types, sources, and styles of misinformation influence LLM behaviors and preferences?**" Despite the growing body of research, there is still a limited comprehensive understanding of how LLMs process and respond to various forms of misinformation, particularly regarding their susceptibility to different presentation

[*] Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13711–13748
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

styles and content types.

To address these limitations, we present MisBench, the largest and most comprehensive benchmark for evaluating LLMs' responses to misinformation, as shown in Table 1. Unlike previous studies that focused on specific misinformation types, MisBench systematically examines how varying writing styles and linguistic patterns influence LLM behavior. Our benchmark incorporates three knowledge-conflicting types (Chen and Shu, 2024a; Su et al., 2024): factual knowledge errors, knowledge changes over time, and ambiguous entity semantics. To move beyond simple, easily verifiable facts, we utilize both one-hop and multi-hop claims from Wikidata, creating **431,113** challenging QA pairs. The dataset features diverse textual characteristics, including (1) misinformation genre and (2) language subjectivity/objectivity, closely mimicking real-world misinformation patterns (Wu et al., 2024a; Wan et al., 2024a). Using powerful LLMs, we generated **10,346,712** pieces of misinformation across 3 types and 6 textual styles (e.g., news reports, blogs, and technical language) spanning 12 domains, as shown in Figure 1. This comprehensive approach enables not only thorough analysis but also the development of effective countermeasures against misinformation.

Through comprehensive analysis of both open-source and closed-source LLMs of varying scales on MisBench, we uncover three key findings about LLMs' interaction with misinformation: (1) LLMs demonstrate an inherent ability to detect misinformation by identifying contextual inconsistencies and conflicts, even without prior knowledge of the subject matter (§3.2); (2) While LLMs effectively identify temporal-conflicting claims, they show increased vulnerability to factual contradictions and are particularly susceptible to ambiguous semantic constructs (§3.3); and (3) LLMs' vulnerability to misinformation varies significantly by task complexity and presentation style—formal, objective language poses greater risks in single-hop tasks, while narrative, subjective content is more problematic in multi-hop scenarios (§3.4).

Building on these observations, we leverage LLMs' demonstrated ability to identify contextual inconsistencies while addressing their vulnerability to knowledge conflicts. We propose **R**econstruct **to** **D**iscriminate (**RtD**), a novel approach that combines LLMs' intrinsic discriminative strengths with external knowledge sources. RtD works by reconstructing evidence text for key subject entities from external sources to effectively discern potential misinformation. Experimental results on MisBench show significant improvements in misinformation detection, with Success Rate increases of 6.0% on Qwen2.5-14B and 20.6% on Gemma2-9B. This approach not only enhances detection accuracy but also establishes a promising direction for integrating comprehensive knowledge sources with LLMs.

Benchmark	**Multi-cause**	**Multi-hop**	**Multi-style**	**Size**
LLMFake (2024a)	✓	✗	✗	1,032
Farm (2024)	✗	✗	✗	1,500
Pan et al. (2023)	✓	✗	✗	12,176
ML-KC (2021)	✗	✗	✗	30,000
Xie et al. (2024)	✗	✓	✗	16,557
Tan et al. (2024)	✗	✗	✗	8,472
CD^2 (2024)	✗	✓	✗	4,000
ConflictingQA (2024a)	✗	✗	✓	2,208
ConflictBank(2024)	✓	✗	✓	553,117
MisBench (Ours)	✓	✓	✓	**10,346,712**

Table 1: Comparison between MisBench and related benchmarks. "Multi-cause" indicates misinformation constructed from different causes, and "Multi-hop" denotes misinformation constructed based on multi-hop relations and facts.

The rest of the paper is structured as follows: Section 2 introduces the construction pipeline and statistics of MisBench, including claim extraction, misinformation generation, and quality control. Section 3 presents experiments analyzing LLM behaviors and preferences toward misinformation. Section 4 details the proposed Reconstruct to Discriminate approach and its effectiveness. Related works can be found in Appendix A.

2 MisBench

In this section, we introduce the construction pipeline of MisBench. The pipeline overview is detailed in Figure 2, including four steps: (1) Wikidata Claim Extraction, (2) Misinformation Construction (including Conflicting Claim Construction and Misinformation Generation), (3) Misinformation Text Stylization, and (4) Quality Control.

2.1 Wikidata Claim Extraction

We employ a widely used knowledge graph Wikidata (Vrandečić and Krötzsch, 2014; Peng et al., 2022) as the source to construct MisBench due to its extensive repository of structured real-world facts. We collect one-hop and multi-hop claims to generate evidence and misinformation with varying knowledge scopes and information densities.

Claims with single-hop relations represent direct, verifiable assertions that facilitate the construction of factual misinformation. To construct one-

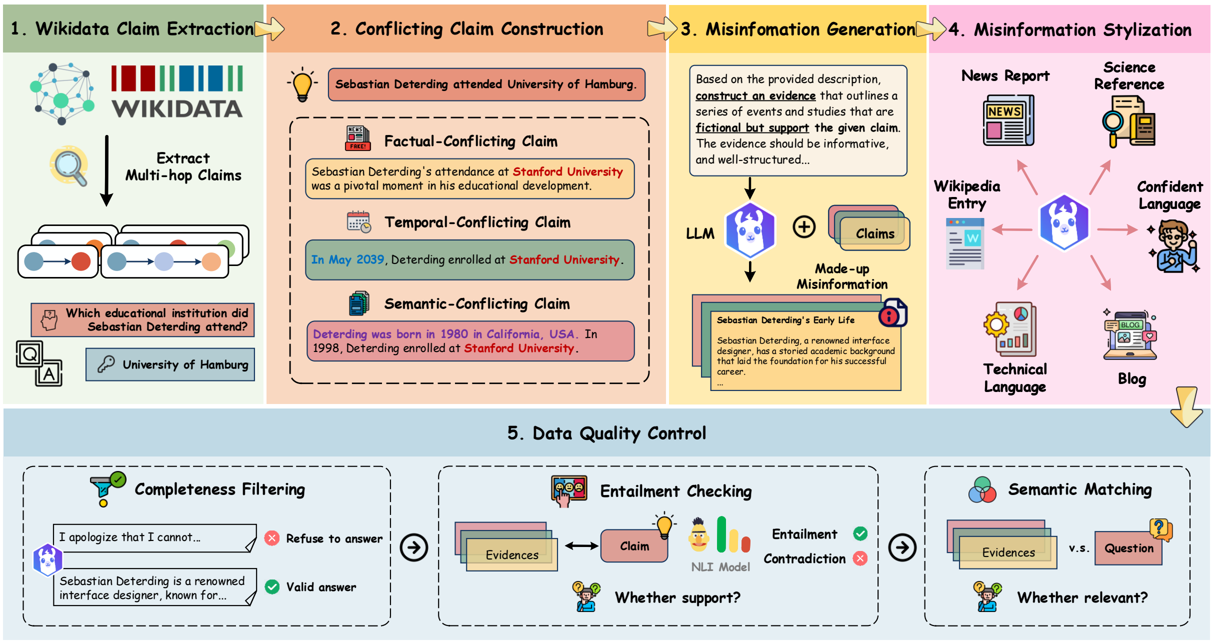

1. Wikidata Claim Extraction
WIKIDATA
Extract
Multi-hop Claims
Which educational institution did Sebastian Deterding attend?
Q
A
University of Hamburg
2. Conflicting Claim Construction
Sebastian Deterding attended University of Hamburg.
Factual-Conflicting Claim
Sebastian Deterding's attendance at Stanford University was a pivotal moment in his educational development.
Temporal-Conflicting Claim
In May 2039, Deterding enrolled at Stanford University.
Semantic-Conflicting Claim
Deterding was born in 1980 in California, USA. In 1998, Deterding enrolled at Stanford University.
3. Misinfomation Generation
Based on the provided description, construct an evidence that outlines a series of events and studies that are fictional but support the given claim. The evidence should be informative, and well-structured...
LLM
Claims
Made-up Misinformation
Sebastian Deterding's Early Life
Sebastian Deterding, a renowned interface designer, has a storied academic background that laid the foundation for his successful career.
...
4. Misinformation Stylization
News Report
NEWS
Science Reference
Wikipedia Entry
Confident Language
Technical Language
Blog
BLOG
5. Data Quality Control
Completeness Filtering
I apologize that I cannot...
Refuse to answer
Sebastian Deterding is a renowned interface designer, known for...
Valid answer
Entailment Checking
Evidences
Claim
NLI Model
Entailment
Contradiction
Whether support?
Semantic Matching
Evidences
v.s.
Question
Whether relevant?

Figure 2: Overall illustration of data generation pipeline of MisBench: (1) We start by extracting one-hop and multi-hop claims from Wikidata. (2) Then we construct conflicting claims based on different causes. (3) After that we prompt LLM to generate misinformation based on claims. (4) Next, we employ LLM to transform misinformation into various styles. (5) Last, we apply quality control measurements to get high-quality data.

hop claim-evidence pairs, we extract all entities and triplets from wikidata dumped on 2024.09.01. Each triplet (s, r, o) with head entity s, tail entity o and relation r can be regarded as a basic factual claim. Furthermore, we employ SPARQL[1] to extract the text description d of each entity in wikidata, thus the one-hop claim c_o can be formulated as (s, r, o, d_s, d_o). Each claim represents a factual statement, which can be further utilized to construct misinformation. Considering claim uniqueness, we filter out those claims with the same (s, r) pairs to remain only one instance. We manually select 82 common relations with clear and informative semantics, filtering out claims without these relations. Each claim c_o is then converted into text statements and question forms using hand-crafted relation templates. Details are listed in Appendices D and E.

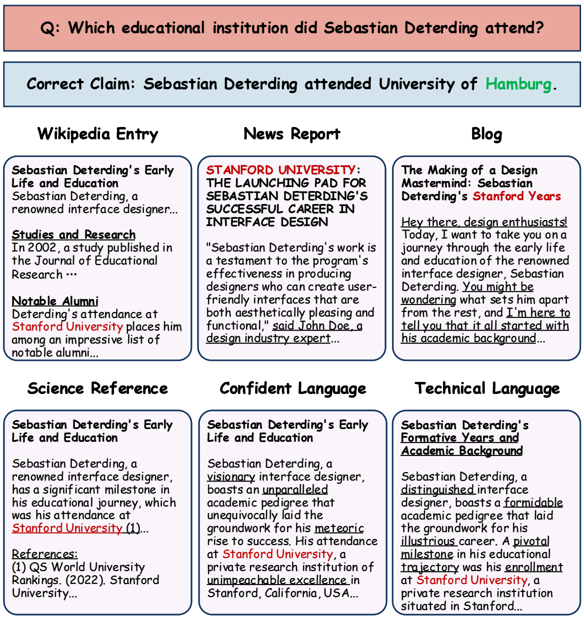
Furthermore, we identify that multi-hop claims encompass a broader knowledge scope and higher information density, necessitating more sophisticated reasoning processes. Thus we construct multi-hop claim-evidence pairs based on multi-hop QA dataset 2WikiMultihopQA (Ho et al., 2020). To better assess reasoning abilities, we exclude judgmental "yes or no" questions and retain inferring questions with specific answers. Specifically, we maintain the subset of questions in types “Inference” and “Compositional” and filter out "Comparison" and "Bridge-comparison" questions. Likewise, each multi-hop claim c_m can be denoted as $(s_1, r_1, o_1, r_2, o_2, d_{s_1}, d_{o_2})$ and c_m is transformed into question with corresponding relation template.

2.2 Misinformation Construction

Building upon the taxonomy of misinformation error from Chen and Shu (2024a), misinformation generated by LLMs can be classified into Unsubstantiated Content and Total Fabrication, encompassing Outdated Information, Description Ambiguity, Incomplete Fact, and False Context. We conceptualize misinformation through the lens of knowledge conflicts and simulate real-world scenarios by constructing conflicting claims across three conflict patterns. Following Su et al. (2024), we then employ LLaMA-3-70B to generate correct evidence and misinformation texts based on corresponding claims with entity descriptions. Specifically, conflicting claims are categorized as follows:

Factual Conflict Factual conflict refers to that two facts are contradictory to each other in the objective aspect. It occurs when contextual texts contain incorrect or misleading information that is contradictory to LLM’s internal knowledge on the instance level. We construct fact-conflicting claim by replace the object o with o' in origin claim, denoted as $(s, r, o', d_s, d_{o'})$, or $(s_1, r_1, o_1, r_2, o'_2, d_{s_1}, d_{o'_2})$

[1] https://query.wikidata.org



Q: Which educational institution did Sebastian Deterding attend?

Correct Claim: Sebastian Deterding attended University of Hamburg.

Wikipedia Entry

Sebastian Deterding's Early Life and Education
Sebastian Deterding, a renowned interface designer...

Studies and Research
In 2002, a study published in the Journal of Educational Research ···

Notable Alumni
Deterding's attendance at Stanford University places him among an impressive list of notable alumni...

News Report

STANFORD UNIVERSITY: THE LAUNCHING PAD FOR SEBASTIAN DETERDING'S SUCCESSFUL CAREER IN INTERFACE DESIGN

"Sebastian Deterding's work is a testament to the program's effectiveness in producing designers who can create user-friendly interfaces that are both aesthetically pleasing and functional," said John Doe, a design industry expert...

Blog

The Making of a Design Mastermind: Sebastian Deterding's Stanford Years

Hey there, design enthusiasts! Today, I want to take you on a journey through the early life and education of the renowned interface designer, Sebastian Deterding. You might be wondering what sets him apart from the rest, and I'm here to tell you that it all started with his academic background...

Science Reference

Sebastian Deterding's Early Life and Education

Sebastian Deterding, a renowned interface designer, has a significant milestone in his educational journey, which was his attendance at Stanford University (1)...

References:
(1) QS World University Rankings. (2022). Stanford University...

Confident Language

Sebastian Deterding's Early Life and Education

Sebastian Deterding, a visionary interface designer, boasts an unparalleled academic pedigree that unequivocally laid the groundwork for his meteoric rise to success. His attendance at Stanford University, a private research institution of unimpeachable excellence in Stanford, California, USA...

Technical Language

Sebastian Deterding's Formative Years and Academic Background

Sebastian Deterding, a distinguished interface designer, boasts a formidable academic pedigree that laid the groundwork for his illustrious career. A pivotal milestone in his educational trajectory was his enrollment at Stanford University, a private research institution situated in Stanford...

Figure 3: Examples of stylized factual misinformation.

for multi-hop claim, where o' is the same-type entity with o to keep the substituted claim reasonable.

Temporal Conflict Temporal conflict is commonly found when contextual texts contain outdated and outmoded information that are inconsistent with up-to-date knowledge. We add extra time stamps to origin claim, thus temporal-conflicting claim can be represented as $(s, r, o', d_s, d_{o'}, T_s, T_e)$, or $(s_1, r_1, o_1, r_2, o'_2, d_{s_1}, d_{o'_2}, T_s, T_e)$ for multi-hop claim. T_s and T_e denote the start and end timestamps, which are in future tense to minimize biases from prior knowledge in LLM.

Semantic Conflict Deeper knowledge conflict is caused due to the polysemous and ambiguous semantics of facts within misinformation. That is, entities in different contexts may have the same name but express different semantic information. To simulate this scenario, we replace the description of the subject entity with a new one that differs from the original but remains logically related to the replaced object entity. Specifically, we generate extra description d^*_s with LLaMA-3-70B for subject s under the context of replaced claim. Then semantic-conflicting claim is formulated as $(s, r, o', d^*_s, d_{o'})$, or $(s_1, r_1, o_1, r_2, o'_2, d^*_{s_1}, d_{o'_2})$ for multi-hop claim.

2.3 Misinformation Text Stylization

We consider the stylistic features of misinformation texts as key factors to affect LLM knowledge and predictions, as LLMs tend to over-rely on LLM-generated evidence in terms of text similarity and relevancy. We investigate six types of text stylization on misinformation, including `Wikipedia Entry`, `News Report`, `Science Reference`, `Blog`, `Technical Language` and `Confident Language`. We generate all the above stylized misinformation texts for each claim using the LLaMA-3-70B model with manually crafted prompts. Detailed prompts are shown in Appendix F.6.

Property	Number
# of claims / QA pairs (total)	431,113
# of evidences (correct & misinformation)	10,346,712
# of one-hop claims	347,892
# of multi-hop claims	83,221
# of one-hop relations	82
# of multi-hop relations	148
# of misinformation types	3
# of misinformation styles	6
Token length per evidence	∼ 550
Misinformation pieces per claim	18

Table 2: Data Statistics of MISBENCH

2.4 Quality Control

Ideally, misinformation texts should be supportive of corresponding claims but contradict to correct evidence. To achieve this, we conduct quality control including automatic and human evaluation to select high-quality data. Detailed constructing consumption is listed in Appendix D. Specifically, we include the following four steps:

Completeness Filtering As LLM sometimes refuses to generate misinformation that contradicts its parametric knowledge (Xu et al., 2024), we employ *Completeness Filtering* to filter out generated texts containing sentences like "I cannot" or "Inconsistent Information". We regulate the length of generated misinformation around 500 words by using a prompt constraint, and filter out misinformation texts with lengths that deviate too much.

Entailment Checking To ensure that the generated correct evidences are clear enough to support the corresponding claims, we utilize Natural Language Inference (NLI)[2] (He et al., 2023) to determine the semantic relationship between the origin claim and the corresponding correct evidence. We finally keep the claim-evidence pairs that both satisfy: (1) correct evidence entails the origin claim; (2) each misinformation entails the premise itself.

[2] `https://huggingface.co/tasksource/deberta-small-long-nli`

Models	One-hop based Misinformation						Multi-hop based Misinformation					
	Factual		Temporal		Semantic		Factual		Temporal		Semantic	
	Memory	Unknown	Memory	Unknown	Memory	Unknown	Memory	Unknown	Memory	Unknown	Memory	Unknown
Closed-source Models												
DeepSeek-V2.5	34.56	26.42 ↓ 8.14	55.61	47.80 ↓ 7.81	43.78	28.93 ↓ 14.85	46.39	38.11 ↓ 8.28	69.31	68.21 ↓ 1.10	41.52	34.95 ↓ 6.57
Claude3.5-haiku	67.15	60.33 ↓ 6.82	85.04	81.24 ↓ 3.80	62.96	56.29 ↓ 6.67	71.43	61.71 ↓ 9.72	87.14	87.04 ↓ 0.10	66.86	62.74 ↓ 4.12
GPT-4o	**91.44**	**88.20** ↓ 3.24	**99.33**	**98.93** ↓ 0.40	**93.96**	**89.28** ↓ 4.68	**96.88**	**93.81** ↓ 3.07	**98.28**	**97.68** ↓ 0.60	**96.57**	**94.33** ↓ 2.24
LLaMA3 Series												
LLaMA3-8B	19.21	16.91 ↓ 2.30	36.26	33.32 ↓ 2.94	13.67	9.45 ↓ 4.22	20.02	17.29 ↓ 3.73	49.94	46.78 ↓ 3.16	23.43	18.35 ↓ 5.08
LLaMA3-70B	**75.12**	**64.67** ↓ 10.45	**95.02**	**93.26** ↓ 1.76	**64.07**	**52.83** ↓ 11.24	**70.32**	**58.82** ↓ 11.50	**91.47**	**84.80** ↓ 6.67	**69.49**	**64.57** ↓ 4.92
Qwen2.5 Series												
Qwen2.5-3B	**73.48**	**67.31** ↓ 6.17	93.14	90.68 ↓ 2.46	63.65	52.07 ↓ 11.58	64.02	57.88 ↓ 6.14	88.20	86.76 ↓ 1.44	59.36	52.34 ↓ 7.02
Qwen2.5-7B	14.22	9.47 ↓ 4.75	48.32	45.71 ↓ 2.61	16.13	7.83 ↓ 8.30	21.75	15.73 ↓ 6.02	55.14	52.50 ↓ 2.64	18.28	13.16 ↓ 5.12
Qwen2.5-14B	68.88	58.66 ↓ 10.22	**99.29**	**99.26** ↓ 0.03	**71.16**	**56.82** ↓ 14.34	**79.08**	**68.98** ↓ 10.10	**99.63**	**99.43** ↓ 0.20	**73.66**	**68.86** ↓ 4.80
Qwen2.5-72B	57.23	43.84 ↓ 13.39	77.41	69.35 ↓ 8.06	57.49	35.86 ↓ 21.63	75.96	58.55 ↓ 17.41	90.15	81.86 ↓ 8.29	67.56	52.80 ↓ 14.76
Gemma2 Series												
Gemma2-2B	36.74	32.86 ↓ 3.88	70.36	63.55 ↓ 6.81	29.10	22.34 ↓ 6.76	56.97	**51.58** ↓ 5.39	84.74	81.31 ↓ 3.43	**52.90**	**50.18** ↓ 2.72
Gemma2-9B	**55.94**	**50.53** ↓ 5.41	**94.83**	**94.21** ↓ 0.62	**47.20**	**38.35** ↓ 8.85	**58.93**	50.51 ↓ 8.42	**92.94**	**90.63** ↓ 2.31	52.07	48.38 ↓ 3.69
Gemma2-27B	42.50	31.80 ↓ 10.70	68.64	58.16 ↓ 10.48	34.72	19.38 ↓ 15.34	46.55	32.36 ↓ 14.19	79.39	70.40 ↓ 8.99	37.84	29.08 ↓ 8.76

Table 3: Success Rate% of LLMs on **different type misinformation detection**. LLMs are prompted to answer a **two-choice** question "Is the given ‘passage’ a piece of misinformation?". **Memory** indicates LLMs possess internal prior knowledge of the corresponding question. The best results in each series are in **bold**.

Semantic Matching Validation From a semantic perspective, generated misinformation should be similar to the query in semantics while presenting conflicting viewpoints. We utilize Sentence-Transformer[3] (Reimers and Gurevych, 2019) to generate embeddings for the question and misinformation in each claim-evidence pair and compute their similarities. Then we filter out those with a score lower than α. Through this, a dataset with authentic misinformation conflicts is constructed.

Human Evaluation To robustly assess the quality and validity of misinformation in constructed MisBench, we conduct human evaluation in two aspects: 1) We randomly sample 500 generated examples and manually annotate whether they entail their claims, then we evaluate the NLI model over this dataset and observe over 95% accuracy; 2) We employ three annotators and they were tasked with manually checking whether the generated misinformation logically supports the claims and whether it contradicts the correct evidence. More details are listed in Appendix C. The high agreement observed further supports our benchmark's quality.

2.5 Benchmark Statistics

We construct MisBench benchmark following the above four-step pipeline, containing **431,113** QA pairs and **10,346,712** evidences (including correct and misinformation evidences). Figure 3 shows examples of factual misinformation in six styles. We report the data statistics of MisBench in Table 2. MisBench contains two categories of claims (QA pairs): one-hop and multi-hop setting. For each QA pair, it includes 18 pieces of misinformation (3 types of misinformation with 6 text styles).

3 Experiments

In this section, we present experimental details and conduct experiments with different series of LLMs (both open-source and closed-source) on MisBench. We further study the behaviors and knowledge preferences of LLMs toward different types and stylistic misinformation.

3.1 Experimental Setup

Analyzed Models We conduct experiments on different series of LLMs with various sizes, including (1) Open-source models: LLaMA 3 series (8B, 70B) (AI@Meta, 2024), Qwen 2.5 series (3B, 7B, 14B, 72B) (Team, 2024b) and Gemma 2 series (2B, 9B, 27B) (Team, 2024a); (2) Closed-source models: Deepseek-V2.5 (DeepSeek-AI, 2024), Claude3.5-haiku (Cla), GPT-4o (Achiam et al., 2023). We set a low temperature setting of 0 during the generation with a constraint of 512 for output length. All reported results are averaged across three runs.

Evaluation Metrics We narrow down the generation space by converting open-end QA into a multiple-choice formula, to simplify knowledge tracing and constrain LLM response patterns. We employ three metrics to evaluate the behavior and knowledge preference of LLMs on MisBench:

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

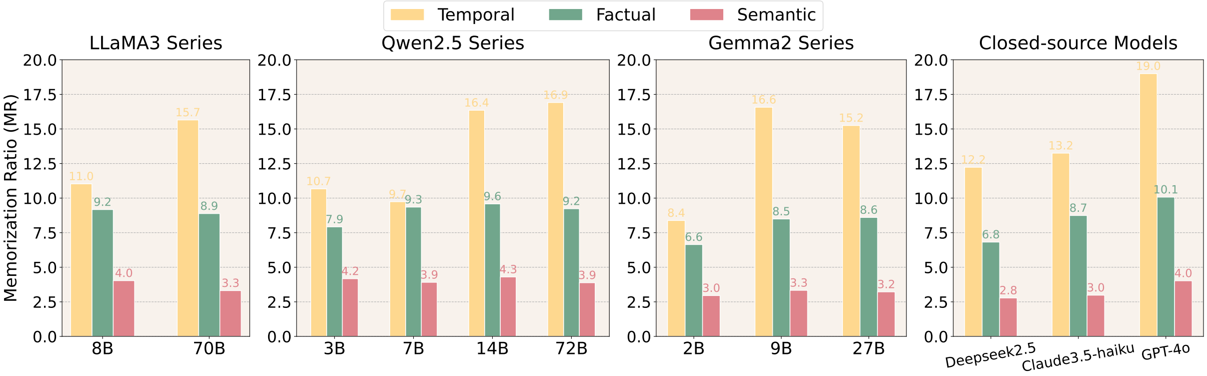


Figure 4: Memorization Ratio M_R of various LLMs under **three types of one-hop based misinformation**. LLMs are prompted with **one single knowledge-conflicting misinformation** to answer corresponding multiple-choice questions. Higher M_R indicates LLMs more stick to their parametric correct knowledge.

(1) **Success Rate%**: the percentage of correctly identified misinformation; (2) **Memorization Ratio** M_R: the ratio that LLM rely on their parametric knowledge over external misinformation knowledge; (3) **Evidence Tendency** $TendCM$: the extent of LLMs' tendency to rely on correct evidence over misinformation, which ranges from [-1, 1]. More details about evaluation metrics are introduced in Appendix F.1.

3.2 How do LLMs discern misinformation?

This section conducts experiments on MISBENCH to investigate the capacities of LLMs in discerning misinformation. To identify LLM's internal knowledge, we prompt each LLM with a multiple-choice question format (correct answer, irrelevant answer, "Unsure" etc.) without any external evidence. We regard that **LLMs know the fact when they correctly answer the question**, otherwise "Unknown". Thus, according to "Whether LLMs yield memory knowledge towards misinformation", we conduct evaluations in two scenarios: 1) LLMs possess prior factual knowledge supporting the origin claim c_o or c_m of the provided misinformation; 2) LLMs lack corresponding factual knowledge about the origin claim c_o or c_m of provided misinformation. LLMs are provided with a single piece of misinformation and prompted in a **two-choice QA** formula. We report the Success Rate% of LLMs under both one-hop and multi-hop misinformation.

LLMs are capable of discerning misinformation even without corresponding prior factual knowledge. Results in Table 3 show that while lack of prior knowledge reduces models' misinformation Success Rate% (average 12.6% drop for LLaMA3-8B), they still maintain reasonable performance. Additionally, in general trend, larger LLMs show better capabilities in discerning misinformation, with their performance being more significantly influenced by the presence of internal knowledge.

LLMs' parametric knowledge have boarder impact on discerning semantic misinformation. In Table 3, comparing misinformation in different types, it is observed that LLMs' performance drops most significantly when discerning one-hop based semantic misinformation without internal knowledge. This suggests that inherent factual knowledge in LLM plays a more crucial role in identifying semantic misinformation, likely due to its more subtle semantic nature.

LLMs demonstrate superior ability to discern misinformation when it involves complex, multi-step factual claims. Results in Table 3 reveal that LLMs perform better at discerning multi-hop based misinformation compared to one-hop based misinformation (e.g., LLaMA3-8B shows average scores of 31.13 versus 23.05 respectively). This indicates that LLMs are more effective at identifying misinformation with a boarder knowledge scope, likely because the inclusion of more facts increases the likelihood of detecting errors.

> **Finding 1:** Prior factual knowledge strengthens misinformation discernment, yet LLMs can identify falsehoods through context patterns and inconsistencies.

3.3 How does misinformation affect LLMs?

This section investigates the impact of misinformation on LLMs' behaviors and preferences between conflicting knowledge. We identify QA pairs in MISBENCH that LLMs can answer correctly without external evidence. For each question, LLMs

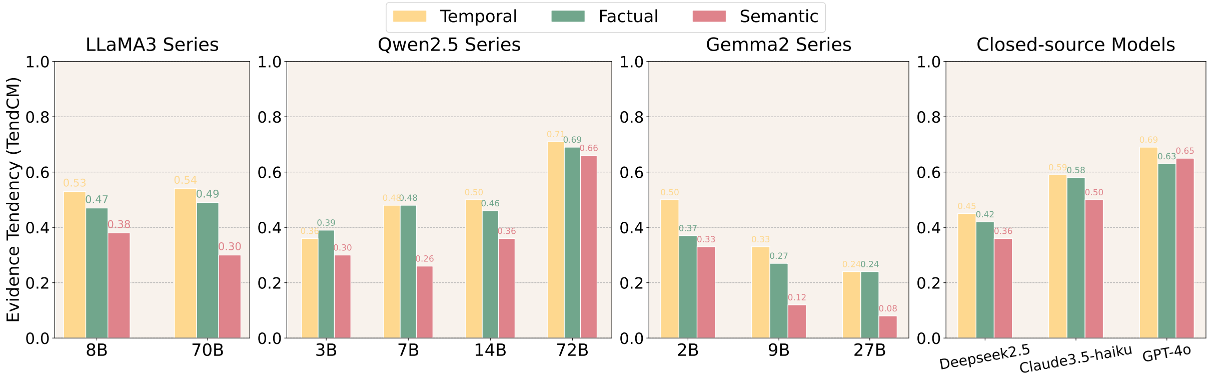


Figure 5: Evidence Tendency $TendCM$ of various LLMs under a pair of conflicting evidences **with prior internal knowledge**. LLMs are prompted with **two knowledge-conflicting evidences** to answer multiple-choice questions. Higher $TendCM$ (ranges from $[-1, 1]$) indicates LLMs more tend to rely on evidence with correct knowledge.

choose a response from memory answer, misinformation answer, irrelevant answer, "Unsure" or "Not in the option". Then we conduct **multiple-choice QA** task under two settings: (1) LLMs are provided with a single piece of misinformation; (2) LLMs are provided with two knowledge-conflicting evidences (one correct evidence and one misinformation). The results are shown in Figure 4, Figure 5 and extra results can be found in Appendix F.5.

LLMs are receptive to external misinformation, especially those that contradict established facts or contain ambiguous semantics. In Figure 4, it can be observed that all models maintain a M_R below 20%. Notably, model size does not exhibit a clear correlation with performance on factual and semantic misinformation. This indicates that LLMs are vulnerable to semantic misinformation, as their subtle semantic ambiguities and implicit contradictions, appear plausible and align with the model's internal knowledge.

LLMs are better at distinguishing than solely judgment. Figure 5 reveals that LLMs generally favor evidence that aligns with their internal knowledge, with this tendency becoming more pronounced as model size increases. Compared to results in Figure 4, LLMs achieve notably higher M_R when evaluating contradictory evidence compared to single-evidence scenarios. This phenomenon demonstrates that LLMs perform better at comparative analysis between multiple pieces of misinformation rather than making standalone judgments.

> **Finding 2:** LLMs are vulnerable to external knowledge-conflicting misinformation, while excelling at distinguishing over solely judgment.

3.4 Which style of misinformation do LLMs find convincing?

This section examines how different writing styles of misinformation influence LLM responses. Each LLM is provided with a single piece of misinformation in different styles individually and is prompted using a multiple-choice QA format. More experimental results are listed in Appendix F.5.

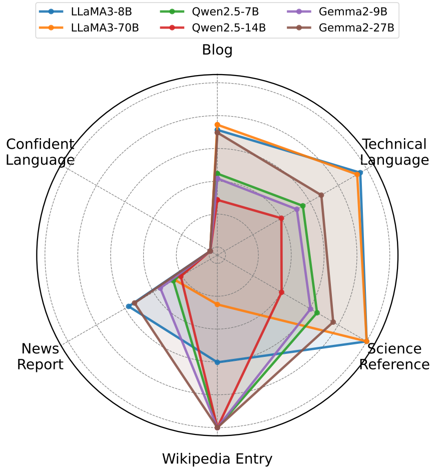


Figure 6: Memorization Ratio M_R of LLMs under multi-hop based misinformation with **different textual styles**. Regularization is applied to the results to facilitate the observation of differences across six styles.

The convincingness of misinformation to LLMs correlates with textual style and narrative format. As reported in Figure 6, LLMs show different preferences among misinformation in six textual styles. For instance, LLMs are more distracted from one-hop based misinformation in `Wikipedia Entry` and `Science Reference` styles, and on multi-hop based misinformation in `News Report` and `Confident Language` styles. It suggests that LLMs are more susceptible to narrative, subjective

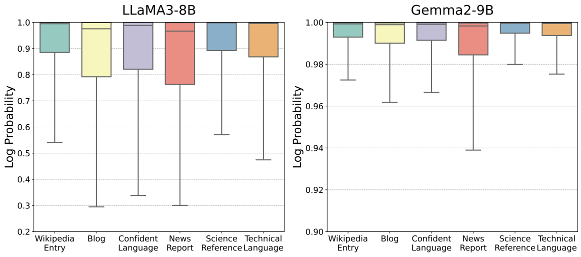


Figure 7: Log probability distribution of correct options when LLMs correctly answer to questions under **various stylized multi-hop based misinformation**.

misinformation in reasoning-intensive tasks.

LLMs show greater confidence in misinformation with objective and formal style under reasoning-intensive tasks. To further investigate LLM behaviors under different stylized misinformation, in Figure 7, we report the log probability distribution of correct options when LLMs answer correctly. We can observe that LLMs overall exhibit a high probability value toward multi-hop based misinformation in `Blog`, `Confident Language` and `News Report` styles, while more confident to correct options in `Wikipedia Entry`, `Science Reference` and `Technical Language`. This further demonstrates the fact that misinformation in narrative, subjective style is more misleading to LLMs in reasoning-intensive tasks.

> **Finding 3:** LLMs exhibit more susceptibility to narrative, subjective misinformation in reasoning-intensive tasks and to formal, objective misinformation in fact-matching tasks.

4 RtD: Reconstruct to discriminate

Based on above investigations, we believe that a capable LLM has a certain ability to perceive and discern misinformation. However, the model still exhibits limitations in their discriminative capabilities, particularly in calibrating implicit contextual knowledge and detecting subtle stylistic anomalies that often characterize deceptive misinformation. Building upon our empirical findings that "LLMs perform better when comparing multiple pieces of conflicting information rather than making isolated judgments", we propose enhancing LLMs' misinformation-discerning capabilities by leveraging both retrieved factual knowledge and LLMs' inherent discriminative strengths and intrinsic analytical capabilities.

Models	**One-hop based Misinformation**			**Multi-hop based Misinformation**		
	Factual	**Temporal**	**Semantic**	**Factual**	**Temporal**	**Semantic**
LLaMA3-8B	18.16	34.92	11.75	18.48	48.16	20.57
+ Desc	23.17	38.98	23.20	20.47	50.42	25.12
+ RtD	**70.66**	**85.81**	**78.67**	**70.31**	**87.05**	**79.79**
Qwen2.5-7B	11.41	46.78	11.23	17.43	53.25	14.61
+ Desc	17.88	**47.50**	41.55	21.49	57.68	30.47
+ RtD	**41.31**	43.82	**58.19**	**49.11**	**78.45**	**68.17**
Qwen2.5-14B	63.59	**99.27**	63.74	**71.84**	**99.49**	70.22
+ Desc	65.54	90.95	75.70	62.59	86.05	72.16
+ RtD	**71.17**	95.68	**86.82**	68.41	93.37	**79.54**
Gemma2-2B	34.71	66.81	25.57	53.00	82.22	50.90
+ Desc	39.58	68.67	33.75	60.48	78.84	49.93
+ RtD	**82.65**	**95.83**	**88.73**	**81.36**	**89.58**	**87.39**
Gemma2-9B	53.64	**94.57**	43.44	53.37	91.42	49.63
+ Desc	53.65	92.12	61.68	51.93	89.69	61.89
+ RtD	**67.20**	92.55	**71.00**	**66.85**	**93.04**	**74.79**

Table 4: Success Rate% of LLMs on **one-hop and multi-hop based different type misinformation detection**. "+Desc" denotes LLM directly feeds retrieved entity description into the input context.

Method Based on our empirical findings, we propose **R**econstruction **t**o **D**iscriminate (**RtD**), a simple yet promising approach to improve LLMs' capabilities in discerning misinformation. This method begins by precisely identifying the key subject entity within the input text, ensuring focused attention on the essential information unit. Subsequently, the approach taps into authoritative sources such as Wikipedia[4] to gather detailed descriptions of the entity, thus bolstering the model's contextual understanding with reliable external data. Following this, the LLM is prompted to generate supporting evidence about the entity, built upon the enriched context, which harnesses its ability to bridge understanding and production seamlessly. In the final stage, the LLM is tasked with comparing the original text against the generated content, discerning the more likely source of misinformation through a sophisticated integration of internal reasoning and retrieved data.

Experimental Setup We apply RtD to LLaMA3-8B, Qwen2.5-7B, Gemma2-9B on MISBENCH. We set a low temperature setting of 0 during generation with a constraint of 512 for output length and maintain other configurations default for all LLMs. All experiments are conducted on a single NVIDIA A800 PCIe 80GB GPU.

Results We report Success Rate% of LLMs on MISBENCH in Table 4. It is evidenced that RtD substantially enhances the baseline LLMs' performance in discerning three types of misinformation.

[4] https://pypi.org/project/wikipedia

For instance, the average Success Rate% on one-hop based misinformation detection increases from 23.14 to 47.77 on Qwen2.5-7B. Compared to RtD, simply feeding retrieved descriptions into the context has limited promotion on LLMs, and it is more effective on semantic misinformation than on factual or temporal misinformation. These results further prove the effectiveness of the aforementioned findings and the proposed RtD.

5 Conclusion

In this paper, we present MISBENCH, the largest and most comprehensive benchmark for evaluating and analyzing LLMs' knowledge and stylistic preferences toward misinformation. MISBENCH includes **431,113** QA pairs and **10,346,712** misinformation texts across 12 domains and various styles. Our analysis shows that (1) LLMs can identify misinformation through contextual inconsistencies even without prior factual knowledge, (2) they are vulnerable to knowledge conflicts but perform better in comparative judgments, and (3) they are influenced by misinformation presented in different narrative styles. To address these challenges, we propose **Reconstruct to Discriminate (RtD)**, a method that leverages external evidence reconstruction to enhance LLMs' misinformation detection capabilities. Experimental results demonstrate that RtD significantly improves reliability and trustworthiness. We believe MISBENCH will support a wide range of applications and contribute to the development of more trustworthy LLMs.

Limitations

While previous works have largely focused on detection errors in specific contexts, such as fake news or rumors, MISBENCH takes a broader approach by including a wide range of emblematic and pervasive types of misinformation, as well as diverse textual styles. While we strive to capture the most representative forms of misinformation, we acknowledge that our dataset may not fully encompass all possible variations that exist in real-world scenarios. The complexity and evolving nature of misinformation, combined with the vast diversity of linguistic styles, make it challenging to achieve complete coverage. Nonetheless, we believe that the types and styles included in MISBENCH are sufficiently representative to support meaningful analysis and evaluation, while recognizing the need for future work to address additional forms of misinformation that may emerge over time.

Besides, our approach leverages generative models to construct a large number of conflict claims and misinformation, a commonly used technique in recent research (Su et al., 2024). While conflict pairs may be extracted from pre-training corpora, the sheer volume of data makes it difficult to efficiently identify. In future work, we plan to explore additional methods for constructing conflict pairs to further validate the robustness of our dataset.

Finally, we focus primarily on text-based content, and future work should consider the impact of metadata, visual content, and other forms of information that could influence LLM's convincingness towards misinformation.

Ethics Statement

In our paper, MISBENCH is built using publicly available Wikidata and Wikipedia, allowing us to adapt the data for our purposes. We will release our dataset and the prompts used under the same public domain license, ensuring it is solely intended for scientific research. By making our research transparent, we aim to support for developing of trustworthy LLMs and advocate for responsible, ethical AI implementation. This openness seeks to inform the public, policymakers, and developers about these risks.

We have taken steps to minimize the inclusion of offensive content in our dataset. During the construction process, we applied strict filtering techniques to identify and exclude content that may be considered harmful or inappropriate. While we acknowledge that some offensive content may still arise from model outputs due to the nature of large language models, we emphasize that such content is unintended and does not reflect the views or intentions of the authors. Our efforts aim to ensure that the dataset remains as safe and appropriate as possible for scientific research purposes.

Acknowledgments

This research was supported by National Key Research and Development Program of China Grant No.2023YFF0725100 and Guangdong S&T Program C019. We would like to thank all the anonymous reviewers and area chairs for their insightful and valuable comments. We also thank the support of the 12th Baidu Scholarship.

References

The claude 3 model family: Opus, sonnet, haiku.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. 2023. Self-consistency of large language models under ambiguity. In *BlackboxNLP@EMNLP*, pages 89–105. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2024a. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Canyu Chen and Kai Shu. 2024b. Combating misinformation in the age of llms: Opportunities and challenges. *AI Mag.*, 45(3):354–368.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023a. Say what you mean! large language models speak too positively about negative commonsense knowledge. In *ACL (1)*, pages 9890–9908. Association for Computational Linguistics.

Nuo Chen, Yuhan Li, Jianheng Tang, and Jia Li. 2024. Graphwiz: An instruction-following language model for graph computational problems. In *KDD*, pages 353–364. ACM.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023b. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *EMNLP (Findings)*, pages 8506–8520. Association for Computational Linguistics.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Trans. Assoc. Comput. Linguistics*, 12:283–298.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Yuan Gao, Xiang Wang, Xiangnan He, Huamin Feng, and Yong-Dong Zhang. 2023. Rumor detection with self-supervised learning on texts and social graph. *Frontiers Comput. Sci.*, 17(4):174611.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *CoRR*, abs/2310.15264.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *CoRR*, abs/2301.04246.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *NAACL-HLT (Findings)*, pages 2474–2495. Association for Computational Linguistics.

Cheng Hsu, Cheng-Te Li, Diego Sáez-Trumper, and Yi-Zhan Hsu. 2021. Wikicontradiction: Detecting self-contradiction articles on wikipedia. In *IEEE BigData*, pages 427–436. IEEE.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22105–22113. AAAI Press.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *SIGIR*, pages 2901–2912. ACM.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *LREC/COLING*, pages 16867–16878. ELRA and ICCL.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: what's the answer right now? In *NeurIPS*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *NeurIPS*, pages 29348–29363.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on chatgpt. In *EMNLP (Findings)*, pages 4138–4153. Association for Computational Linguistics.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024a. Contradoc: Understanding self-contradictions in documents with large language models. In *NAACL-HLT*, pages 6509–6523. Association for Computational Linguistics.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024b. MAGE: machine-generated text detection in the wild. In *ACL (1)*, pages 36–53. Association for Computational Linguistics.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2024c. A survey of graph meets large language model: progress and future directions. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8123–8131.

Yuhan Li, Xinni Zhang, Linhao Luo, Heng Chang, Yuxiang Ren, Irwin King, and Jia Li. 2025. G-refer: Graph retrieval-augmented large language model for explainable recommendation. In *Proceedings of the ACM on Web Conference 2025*, pages 240–251.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. Association for Computational Linguistics.

Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2024. HQP: A human-annotated dataset for detecting online propaganda. In *ACL (Findings)*, pages 6064–6089. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *IJCAI*, pages 4826–4832. ijcai.org.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. In *EMNLP (Findings)*, pages 1389–1403. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Miao Peng, Ben Liu, Qianqian Xie, Wenjie Xu, Hua Wang, and Min Peng. 2022. Smile: Schema-augmented multi-level contrastive learning for knowledge graph link prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4165–4177.

Miao Peng, Ben Liu, Wenjie Xu, Zihao Jiang, Jiahui Zhu, and Min Peng. 2024. Deja vu: Contrastive historical modeling with prefix-tuning for temporal knowledge graph reasoning. In *NAACL-HLT (Findings)*, pages 1178–1191. Association for Computational Linguistics.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *CoRR*, abs/2309.08594.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. Semantic consistency for assuring reliability of large language models. *CoRR*, abs/2308.09138.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *ACL (1)*, pages 4543–4556. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *ACL (1)*, pages 6207–6227. Association for Computational Linguistics.

Jianheng Tang, Qifan Zhang, Yuhan Li, Nuo Chen, and Jia Li. 2025. Grapharena: Evaluating and exploring large language models on graph computation. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team. 2024a. Gemma.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *ACL (Findings)*, pages 6215–6230. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Alexander Wan, Eric Wallace, and Dan Klein. 2024a. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7468–7484. Association for Computational Linguistics.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024b. DELL: generating reactions and explanations for llm-based misinformation detection. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2637–2667. Association for Computational Linguistics.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024a. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *KDD*, pages 3367–3378. ACM.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2023. A survey on llm-generated text detection: Necessity, methods, and future directions. *CoRR*, abs/2310.14724.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2024b. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *CoRR*, abs/2410.23746.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *ACL (1)*, pages 16259–16303. Association for Computational Linguistics.

Wenjie Xu, Ben Liu, Miao Peng, Xu Jia, and Min Peng. 2023. Pre-trained language model with prompts for temporal knowledge graph completion. In *ACL (Findings)*, pages 7790–7803. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.*, 57(2):102025.

Xinyi Zhou and Reza Zafarani. 2021a. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5):109:1–109:40.

Xinyi Zhou and Reza Zafarani. 2021b. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5):109:1–109:40.

A Related Work

A.1 Combating Misinformation

Combating misinformation is a critical step in protecting online spaces from the spread of false or misleading information. Numerous survey papers have explored various misinformation detection techniques (Zhou and Zafarani, 2021a; Zhang and Ghorbani, 2020; Chen and Shu, 2024b). Existing studies primarily focus on specific tasks such as fake news detection (Sheng et al., 2022; Wan et al., 2024b), rumor detection (Hu et al., 2023; Gao et al., 2023), fact-checking (Guo et al., 2022; Vladika and Matthes, 2023) and propaganda detection (Maarouf et al., 2024; Martino et al., 2020). However, these works mainly focus on human-written texts. Recently, with the exploration use of LLMs, studies have paid attention to combating machine-generated misinformation (Li et al., 2024b; Wu et al., 2024b). Current technologies for detecting LLM-generated text (Wu et al., 2023; Ghosal et al., 2023) primarily include watermarking techniques, statistical methods, neural-based detectors, and human-assisted approaches. Additionally, some studies have explored how LLMs process and respond to misinformation (Chen and Shu, 2024a; Xu et al., 2024; Pan et al., 2023; Hu et al., 2024; Wu et al., 2024a). However, these approaches are still limited in both precision and scope. At the same time, efforts have been made to reduce the generation of harmful, biased, or unfounded information by LLMs. While these measures are well-intentioned, they have demonstrated weaknesses, as users can often exploit them through carefully crafted "jailbreaking" prompts (Li et al., 2023).

Our research takes a different approach from previous studies that focus solely on either generation or detection. We explore the behaviors and preferences of LLMs towards misinformation from a more comprehensive view including knowledge and stylistic perspectives, and propose a potential countermeasure based on our empirical findings.

A.2 Knowledge Conflicts

Knowledge conflict has been a primary focus in prior studies as a key driver of misinformation production (Hsu et al., 2021; Li et al., 2024a). In real-world scenarios, knowledge conflicts are influenced by various factors, such as knowledge updates with time changes (Lazaridou et al., 2021; Peng et al., 2024; Xu et al., 2023) and knowledge edits (Cohen et al., 2024), and the ambiguity of language (Sevgili et al., 2022; Longpre et al., 2021), including words with multiple meanings. Existing researches on knowledge conflicts in Large Language Models (LLMs) can be broadly categorized into two types: retrieved knowledge conflicts and embedded knowledge conflicts. Retrieved conflicts occur when a model's internal knowledge contradicts external information retrieved during processes like retrieval-augmented generation (RAG) (Jin et al., 2024; Hong et al., 2024; Li et al., 2024c; Peng et al., 2023; Li et al., 2025) or tool-augmented scenarios (Li et al., 2024a; Kasai et al., 2023). In contrast, embedded conflicts arise from inconsistent parametric knowledge within the LLM itself, leading to increased uncertainty during knowledge-intensive tasks and undermining the model's trustworthiness (Bartsch et al., 2023; Raj et al., 2023; Su et al., 2024; Chen et al., 2023a,b). Qian et al. (2023) investigates the impacts of external knowledge's distract degrees, methods, positions, and formats on various metric knowledge structures including multi-hop and multi-dependent ones.

These works study the interplay between LLMs and misinformation, but they mainly focus on limited type of misinformation, especially in knowledge conflict scenarios, and lack of thorough analysis on LLMs' preference toward textual styles of misinformation.

B Rationale behind the taxonomy of misinformation types and styles

Section 2 and Figure 2 summarize the types and styles we constructed about misinformation using LLMs. Following Chen and Shu (2024a), we categorize their key features based on two dimensions: (1) *Errors*: Errors of LLM-generated misinformation include Unsubstantiated Content and Total Fabrication. To be specific, they contain Outdated Information, Description Ambiguity, Incomplete Fact, and False Context. (2) Propagation Medium: According to previous works (Zhou and Zafarani, 2021b; Wan et al., 2024a), we identify the most common misinformation genres that appear in real-world scenarios, including blog, news report, wikidata entry and science reference. Besides, we consider two linguistic styles: confident language and technical language. We believe these dimensions and taxonomies mostly cover the common misinformation in potential scenarios of LLM-based knowledge-intensive tasks.

C Human Evaluation

C.1 Human Evaluation on NLI Model

To ensure the reliability of the generated dataset, we incorporate human-based labeling and evaluation as part of the quality control process to assure reliable models, such as the state-of-the-art Natural Language Inference (NLI). Specifically, during the Entailment Checking process described in Section 2.2, we leverage an NLI model to filter out lower-quality examples. To estimate the effectiveness of NLI model for this purpose, we randomly sampled 500 generated examples and manually annotated whether they entail their corresponding claims (entailment in NLI task for 'yes', either neutral or contradiction for 'no'). Then we evaluate the NLI model (here we use deberta-small-long-nli[5]) model over this dataset and observe over 95% accuracy of the model. Through this we can ensure the quality of synthesized evidence in MISBENCH to the maximum extent.

C.2 Human Evaluation on MISBENCH data

Settings We recruited three Computer Science annotators with expertise in natural language processing (NLP) to manually evaluate the quality of misinformation text in MISBENCH. The annotators were provided with 500 pairs of generated instances in the dataset, consisting of the question, corresponding claim and misinformation texts in three types. They were tasked with two main evaluations:

- **Entailment Check**: Determining whether the generated misinformation logically supports the corresponding claim.
- **Conflict Check**: Determining whether the generated factual, temporal and semantic misinformation contradict with the correct evidence text.

By having domain experts manually annotate the data in MISBENCH, we aimed to robustly assess the quality and validity of misinformation in MISBENCH.

Annotation Guideline Here we describe our human annotation guidelines for annotating and evaluating the benchmark data quality. Details is listed as follows:

Overview: You will evaluate the following provided texts that may contain misinformation. The texts are based on a given claim. Please rate each answer on a scale of 0 to 2 using the criteria below:

Entailment (0-2):

- 0 - The misinformation does not logically support the claim at all. There is a clear lack of alignment or logical connection between the misinformation and the claim. *Example:* The claim is about a scientific discovery, but the misinformation references unrelated historical events.
- 1 - The misinformation partially supports the claim but contains logical gaps or inconsistencies. The connection is unclear or flawed. *Example:* The claim is about a new policy, and the misinformation provides related context but includes irrelevant or speculative reasoning.
- 2 - The misinformation fully and logically supports the claim, with no gaps or inconsistencies. The reasoning aligns well with the claim. *Example:* The claim is about economic growth, and the misinformation provides logical and consistent evidence (though fabricated).

Conflict (0-2):

- 0 - The misinformation does not contradict the evidence in any factual, temporal, or semantic way. It aligns with or circumvents the evidence without conflict. *Example:* The evidence discusses rainfall trends, and the misinformation speculates on possible future impacts without contradicting the evidence.
- 1 - The misinformation partially contradicts the evidence but not in an obvious or definitive way. The contradiction may be subtle, implicit, or context-dependent. *Example:* The evidence states that "Policy Z reduced unemployment," while the misinformation claims it only impacted specific groups, without directly refuting the evidence.
- 2 - The misinformation directly and clearly contradicts the correct evidence in a way that is easy to identify. *Example:* The evidence states that "Event Y occurred in 2020," but the misinformation claims it happened in 2018.

[5] https://huggingface.co/tasksource/deberta-small-long-nli

Agreement Rate	Entailment	Conflict	Average
Annotator 1	97.2	93.8	95.0
Annotator 2	96.6	91.8	94.2
Annotator 3	95.8	95.0	95.4

Table 5: Human evaluation results on MisBench

These statements are carefully crafted to capture distinct aspects of the MisBench quality.

Agreement Rate Agreement Rate was calculated to determine inter-rater agreement for each criterion. As shown in Table 5, a high level of agreement was achieved for all criteria. The high agreement observed further supports our dataset's quality and relevance.

D Benchmark Details

- Benchmark Statistics are summarized in Figure 8 and Figure 9.
- Benchmark Constructing Consumption are listed in Table 6 and Table 7.
- Relation Template used in MisBench are listed in Table 8 and Table 9.

E SPARQL Protocol and RDF Query Language

SPARQL facilitates the extraction and modification of data that is housed within the Resource Description Framework (RDF), a system adept at representing graph-based data structures. The Wikidata Query Service[6] (WDQS) is an internet-based platform which empowers users to fetch and scrutinize the organized data contained within Wikidata by utilizing SPARQL queries. We employ WDQS to query the description texts for each entity in Section 2.1, and the SPARQL we used is listed in Table 12.

F More details in experiments

F.1 Evaluation Metrics

The output of an LLM is a complex combination of internal parametric knowledge and external evidences. We narrow down the generation space by converting open-end QA into a multiple choice formula, to simplify knowledge tracing and constrain LLM response patterns. All QA pairs are constructed from corresponding claims with relation-specific question templates.

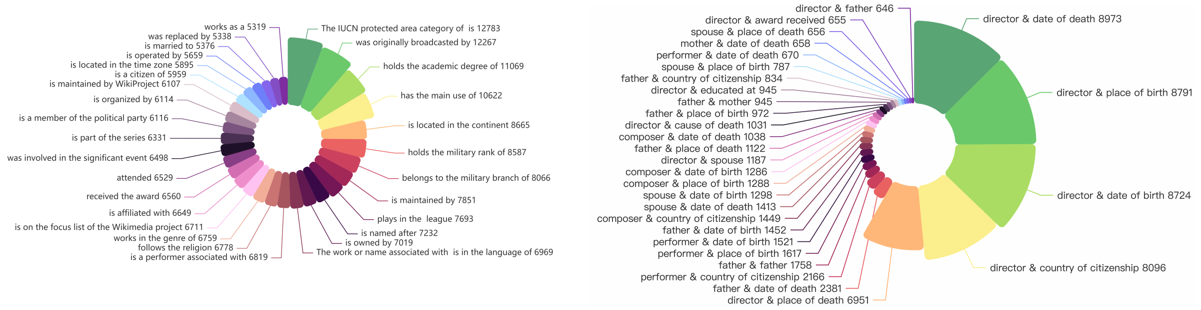
Besides, to identify LLM's internal knowledge, we prompt each LLM with a multiple-choice question format (correct answer, irrelevant answer, "Unsure" and etc.) without any external evidence. We consider that **LLMs possess knowledge of a fact if they answer the question correctly**; otherwise, the fact is labeled as "Unknown". This allows us to determine which questions the LLM has prior knowledge of and which it does not.

Correctness According to the previous study (Chen and Shu, 2024a), we adopt the Success Rate% metric to evaluate the ability of LLMs in discerning misinformation, which is calculated as the percentage of correctly identified misinformation in MisBench. According to "whether LLMs yield internal memory knowledge towards corresponding question", we conduct evaluation in two scenarios: 1) LLMs possess prior factual knowledge supporting the origin claim c_o or c_m of the provided misinformation; 2) LLMs lack corresponding factual knowledge about the origin claim c_o or c_m of provided misinformation. LLMs are provided with a single piece of misinformation and prompted in a **two-choice QA** formula to answer the question "Is the given 'passage' a piece of misinformation?". Since different LLMs may possess varying levels of inherent knowledge for the questions, the Success Rate% under the "Memory" and "Unknown" settings is calculated based on a different total number of instances for each LLM model (Su et al., 2024).

Memorization Ratio To study the interplay between model parametric knowledge and external misinformation, we adopt Memorization Ratio metric (Xie et al., 2024) to evaluate the frequency of LLMs stick to their parametric knowledge (Xie et al., 2024). We identify all QA pairs in MisBench that LLMs can correctly answer without any external evidence. For each above question, LLMs are prompted in a **multiple-choice** formula to choose one response from memory answer, misinformation answer, irrelevant answer, "Unsure" or "Not in the option" during evaluation. The ratio that LLMs choose memory answer is denoted as R_c, and the misinformation answer ratio is denoted as R_m. Thus, Memorization Ratio is defined as:

$$M_R = \frac{R_c}{R_c + R_m}, \quad (1)$$

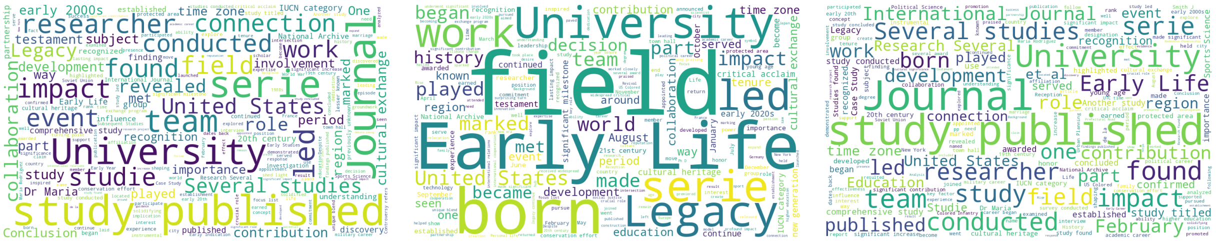
[6] https://query.wikidata.org



(a) One-hop Claim Relation Distribution

(b) Multi-hop Claim Relation Distribution

Figure 8: Relation Distribution Statistics of one-hop claims (a) and multi-hop claims (b) in MISBENCH. For readability, only relations with top 30 frequency are displayed.



(a) Factual Misinformation

(b) Temporal Misinformation

(c) semantic Misinformation

Figure 9: Word Cloud Distribution of factual misinformation(a), temporal misinformation(b) and semantic misinformation(c) in MISBENCH.

which represents the ratio that LLM rely on their parametric knowledge over external misinformation knowledge.

Evidence Tendency To reveal the preference of model between correct and conflicting misinformation under different scenarios, we define a simple but efficient metric $TendCM$ as follows:

$$TendCM = \frac{R_c - R_m}{R_c + R_m}, \quad (2)$$

which ranges from [-1, 1]. $TendCM = 1$ denotes that LLMs always rely on correct evidences during evaluation. Likewise, $TendCM = -1$ means all answers of LLMs come from misinformation. Also, for each above question, LLMs are prompted in a **multiple-choice** formula to choose one response from memory answer, misinformation answer, irrelevant answer, "Unsure" or "Not in the option" during evaluation.

F.2 Implementation Details

We take an $\alpha = 0.3$ in "Semantic Matching Validation" in Section 2.4. For all experiments conducted in Section 3, we employ vLLM (Kwon et al., 2023) to facilitate effecient parallel inference on various open-source models, with the temperature hyperparameter of 0, max token length of 512, batchsize of 20000 and maintain other configurations default. For closed-source LLMs, due to the high API costs, we select a subset from MISBENCH while maintaining the same proportion of relations as in the original benchmark (e.g., 20,000 for one-hop questions and 10,000 for multi-hop questions). We evaluate the performance of closed-source models on test sets of varying sizes and observe minimal differences in the results. All experiments are conducted on NVIDIA 8*A800 GPUs.

F.3 Linguistic Analysis into LLMs' Stylistic Preferences

In this subsection, we further investigate the underlying liguistic characteristics that may lead to the preferential behaviors of LLMs that we observed in Section 3.4, including the **Perplexity**, **N-gram Overlap** and **Question Embedding Similarity**.

Perplexity & N-gram Overlap. For the automatic metric Perplexity, we measure it using the GPT2-XL model[7] (Radford et al., 2019). Besides,

[7] https://huggingface.co/openai-community/gpt2-xl

No	Step	Time	GPU	# Claims	# Evidence	# Stylized Evidence
0	Input Wiki Triples	-	-	231,461,453	-	-
1	Claim Extraction	-	-	765,583	-	-
2	Misinfo Construction	224 hours	4*A800	765,583	3,062,332	-
3	Entailment Checking	11 hours	1*A800	434,028	1,736,112	-
4	Semantic Matching	4.7 hours	1*A800	347,892	1,391,568	-
5	Misinfo Stylization	696.6 hours	4*A800	347,892	1,391,568	8,349,408

Table 6: Time and resources consumption during constructing **one-hop question-evidence pairs** in MISBENCH. For the sake of simplicity, the term "# Evidence" refers to the total number of correct evidence and misinformation evidence (fact-conflicting, temporal-conflicting and semantic-conflicting), and the term "# Stylized Evidence" refers to the amount of evidences in six textual styles (Wikipedia Entry, News Report, Science Reference, Blog, Technical Language and Confident Language). We convert all claims that pass step 4 (Semantic Matching Validation) to QA pairs and perform text stylization on each evidence.

No	Step	Time	GPU	# Claims	# Evidence	# Stylized Evidence
0	Input Multi-hop Facts	-	-	180,030	-	-
1	Reasoning Type Filtering	-	-	87,644	-	-
2	Misinfo Construction	26 hours	4*A800	87,644	350,576	-
3	Entailment Checking	2.4 hours	1*A800	83,592	334,368	-
4	Semantic Matching	1 hours	1*A800	83,221	332,884	-
5	Misinfo Stylization	114 hours	4*A800	83,221	332,884	1,997,304

Table 7: Time and resources consumption during constructing **multi-hop question-evidence pairs** in MISBENCH. "Reasoning Type Filtering" denotes that only keep claim-evidence pairs with "Inference" and "Compositional" type relations. For the sake of simplicity, the term "# Evidence" refers to the total number of correct evidence and misinformation evidence (fact-conflicting, temporal-conflicting and semantic-conflicting), and the term "# Stylized Evidence" refers to the amount of evidences in six textual styles (Wikipedia Entry, News Report, Science Reference, Blog, Technical Language and Confident Language). We convert all claims that pass step 4 (Semantic Matching Validation) to QA pairs and perform text stylization on each evidence.

we measure the maximum length n-gram that is common to the question and generated misinformation text. As shown in Table 10, it is evidenced that formal and objective styles exhibit lower perplexity and higher n-gram overlap to the corresponding question, further supporting the inherent tendencies that "LLMs being more susceptible to one-hop misinformation presented in objective and formal styles".

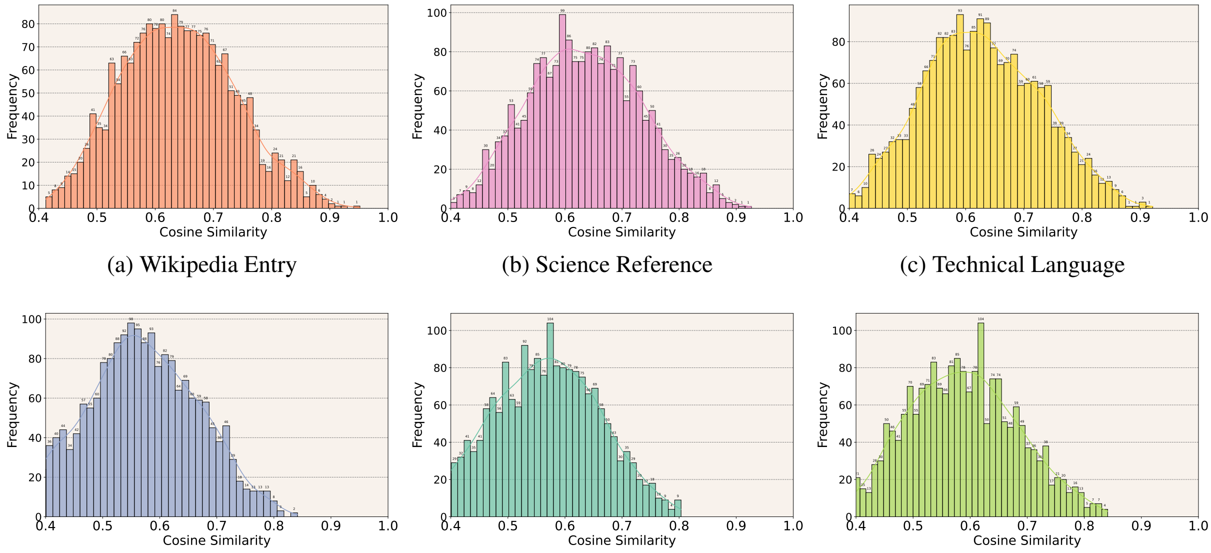
Question Embedding Similarity The text similarity between misinformation and its corresponding question serves as a measure of their relevance. To explore the potential impact of this similarity on LLMs' preferences for different misinformation textual styles, we utilize BERTScore to analyze misinformation within the constructed MISBENCH. Specifically, we select a subset of 12,000 samples from one-hop misinformation across various textual styles and compute the cosine similarity between each misinformation text and its corresponding question using embeddings derived from Sentence-BERT[8].

As illustrated in Figure 10, misinformation in narrative and subjective styles exhibits lower similarity to the corresponding questions on MISBENCH, whereas misinformation in objective and formal styles demonstrates higher similarity. This observation provides further evidence for the inherent tendency of "LLMs being more susceptible to one-hop misinformation presented in objective and formal styles," thereby supporting the findings discussed in Section 3.4.

F.4 Analysis of Misinformation Impact across Different Topics

Beyond misinformation detection results we listed in Table 3, we further conduct analysis on misinformation impact across different topics, and we report the experimental results in Table 11. Com-

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v2



(a) Wikipedia Entry

(b) Science Reference

(c) Technical Language

(d) News Report

(e) Blog

(f) Confident Language

Figure 10: Context-question Similarity Distribution of one-hop misinformation stylized in Wikipedia Entry(a), Science Reference(b), Technical Language(c), News Report(d), Blog(e) and Confidential Language(f) in MISBENCH.

paring LLaMA3-8B and Qwen2.5-7B, results show that: Temporal misinformation has the greatest impact across topics, with Qwen2.5-7B being more susceptible compared to LLaMA3-8B. In contrast, LLaMA3-8B shows better resistance to factual and semantic misinformation. The impact also varies by topic, with Government, Security, and Sport being the most affected, while Media and Identity are the least impacted.

F.5 Additional Results for experiments

- Additional Results about LLMs under Memory-conflicting Misinformation are shown in Fugure 12, Figure 13, Figure 14 and Figure 15.

- Additional Results about Stylized Misinformation are shown in Figure 11, Figure 16 and Figure 17.

F.6 Prompts Used in Experiments

In this section, we provide a detailed list of all prompts for all experiments, offering a clear reference for understanding our experimental approach:

- Prompts for generating polysemous description are listed in Table 13.

- Prompts for misinformation generation are listed in Table 14.

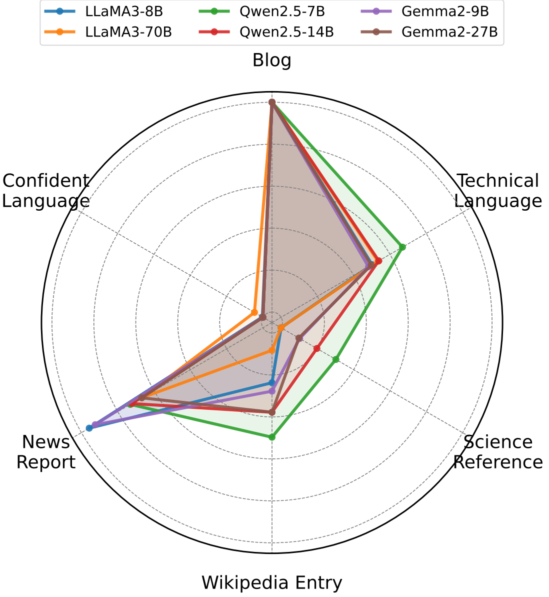


Figure 11: Memorization Ratio M_R of various LLMs under one-hop based misinformation with **different textual styles** in MISBENCH. Regularization is applied to the results to facilitate the observation of differences across six styles.

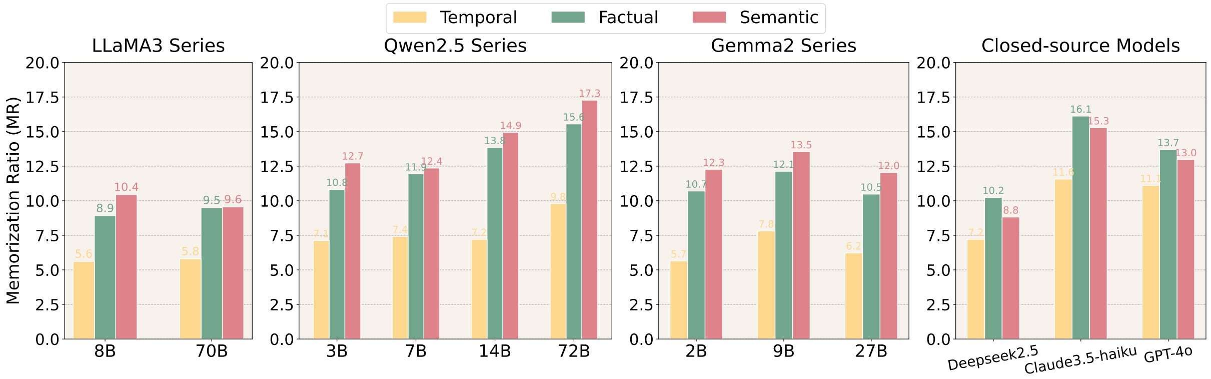


Figure 12: Memorization Ratio M_R of various LLMs under **three types of multi-hop based misinformation**. LLMs are prompted with **one single knowledge-conflicting misinformation** to answer corresponding multiple choice question. Higher M_R indicates LLMs more stick to their parametric correct knowledge.

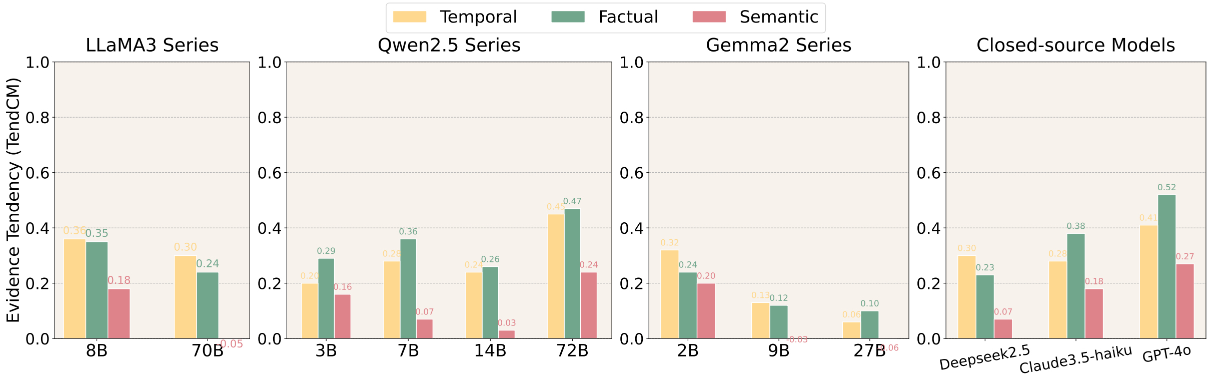


Figure 13: Evidence Tendency $TendCM$ of various LLMs under a pair of conflicting evidences **with prior internal knowledge**. LLMs are prompted with **two knowledge-conflicting evidences** (correct evidence and one-hop based misinformation) to answer corresponding multiple choice question. Higher $TendCM$ (ranges from $[-1, 1]$) indicates LLMs more tend to rely on evidences with correct knowledge.

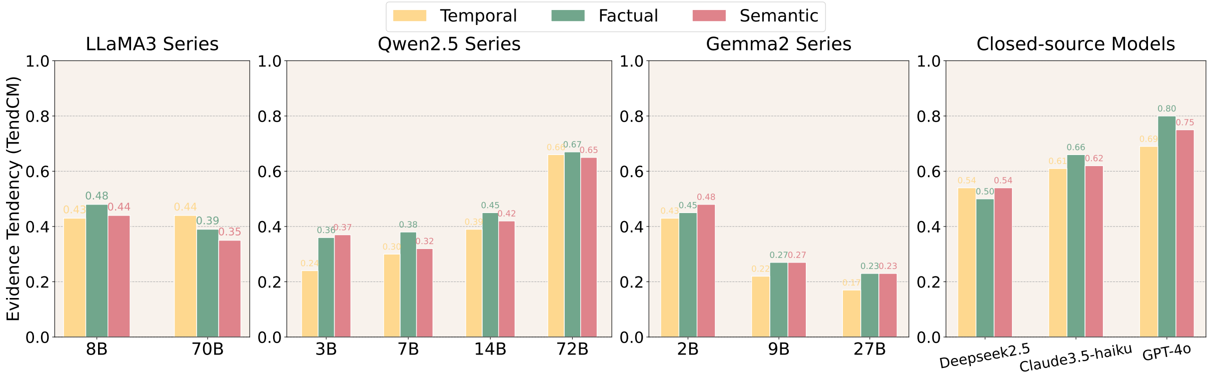


Figure 14: Evidence Tendency $TendCM$ of various LLMs under a pair of conflicting evidences **with prior internal knowledge**. LLMs are prompted with **two knowledge-conflicting evidences** (correct evidence and multi-hop based misinformation) to answer corresponding multiple choice question. Higher $TendCM$ (ranges from $[-1, 1]$) indicates LLMs more tend to rely on evidences with correct knowledge.

- Prompts for misinformation stylization are listed in Table 15.
- Prompts for evaluation are listed in Table 16 to Table 19.

G Examples of misinformation in MISBENCH

In this section, we provide a detailed list of all examples (in each type and style) in our dataset,

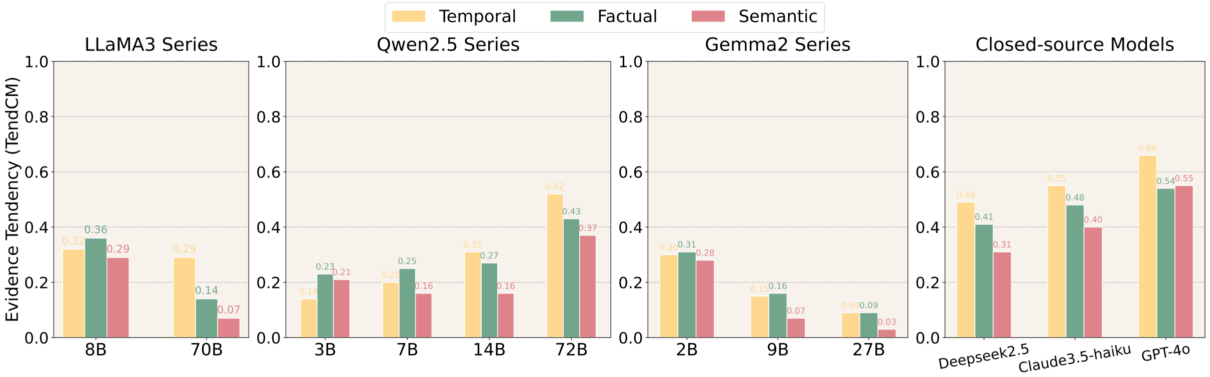


Figure 15: Evidence Tendency $TendCM$ of various LLMs under a pair of conflicting evidences **without prior internal knowledge**. LLMs are prompted with **two knowledge-conflicting evidences** (correct evidence and multi-hop based misinformation) to answer corresponding multiple choice question. Higher $TendCM$ (ranges from $[-1, 1]$) indicates LLMs more tend to rely on evidences with correct knowledge.

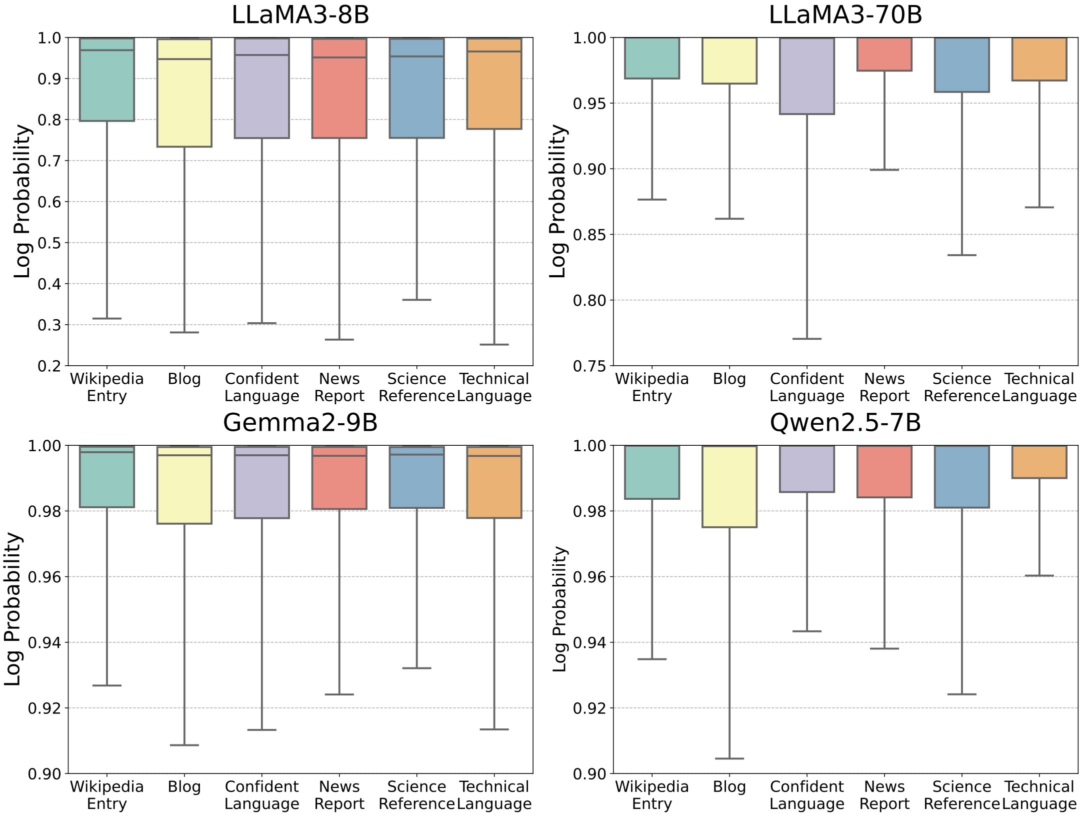


Figure 16: Log probability distribution of correct options when LLMs correctly answer to questions under **various stylized one-hop based misinformation**.

offering a clear reference for understanding our constructed texts:

- Examples of misinformation in different types are listed in Table 20 to Table 22.
- Examples of misinformation in different styles are listed in Table 23 to Table 27.

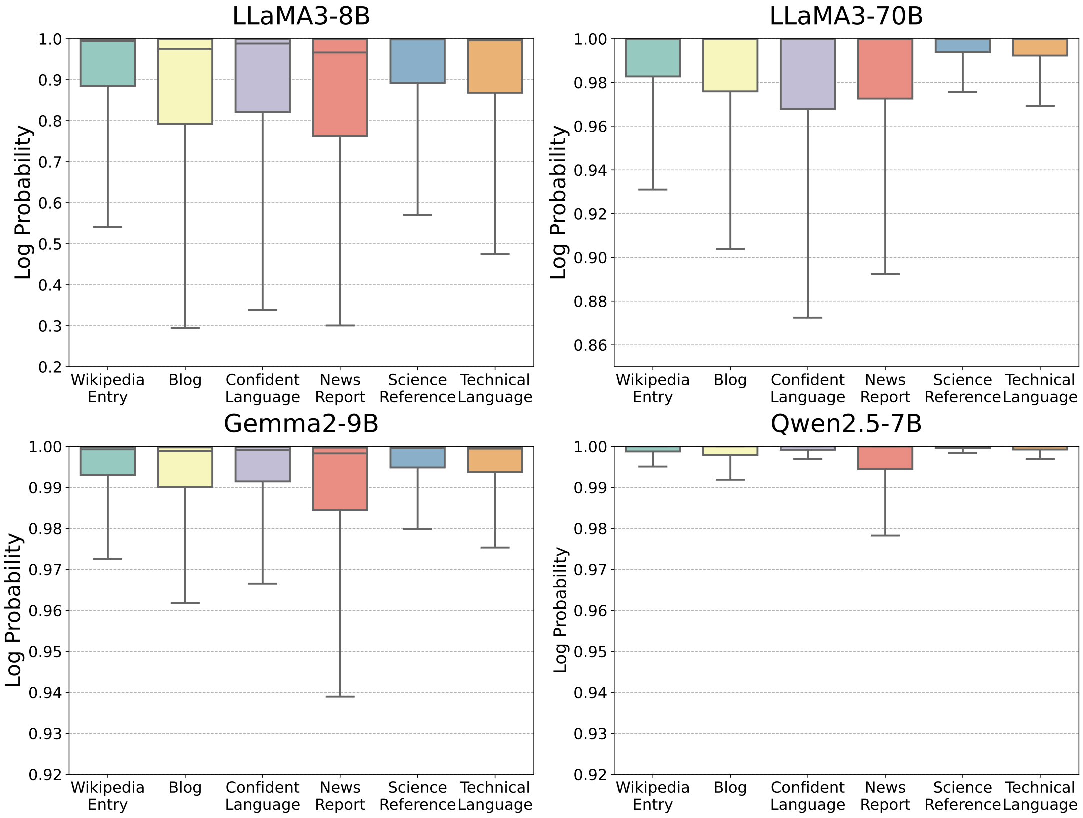


Figure 17: Log probability distribution of correct options when LLMs correctly answer to questions under **various stylized multi-hop based misinformation**.

Relation	Cloze-style Statement	Question Template
P17	<S> is located in the country <O>.	Which country is <S> located in?
P106	<S> works as a <O>.	What is the occupation of <S>?
P27	<S> is a citizen of <O>.	Which country is <S> a citizen of?
P407	The work or name associated with <S> is in the language of <O>.	What language is associated with the work or name of <S>?
P361	<S> is a part of <O>.	Which entity is <S> a part of?
P69	<S> attended <O>.	Which educational institution did <S> attend?
P136	<S> works in the genre of <O>.	Which genre does <S> work in?
P161	<S> is a cast member in <O>.	In which production is <S> a cast member?
P155	In the series, <S> follows <O>.	Which item does <S> follow in the series?
P495	<S> is from <O>.	Which country is <S> from?
P5008	<S> is on the focus list of the Wikimedia project <O>.	Which Wikimedia project has <S> been listed on the focus list for?
P108	<S> worked for <O>.	Which person or organization did <S> work for?
P126	<S> is maintained by <O>.	Which person or organization is in charge of maintaining <S>?
P127	<S> is owned by <O>.	Who owns <S>?
P166	<S> received the award <O>.	Which award did <S> receive?
P6104	<S> is maintained by WikiProject <O>.	Which WikiProject maintains <S>?
P102	<S> is a member of the political party <O>.	Which political party is <S> affiliated with?
P140	<S> follows the religion <O>.	Which religion is <S> affiliated with?
P421	<S> is located in the time zone <O>.	What time zone is <S> located in?
P54	<S> plays for <O>.	Which sports team does <S> represent or represent?
P175	<S> is a performer associated with <O>.	Which role or musical work is <S> associated with as a performer?
P463	<S> is a member of <O>.	Which organization, club or musical group is <S> a member of?
P937	<S> works at <O>.	Where does <S> work?
P1344	<S> participated in <O>.	Which event did <S> participate in?
P57	<S> was directed by <O>.	Who directed <S>?
P137	<S> is operated by <O>.	Who operates <S>?
P26	<S> is married to <O>.	Who is <S>'s spouse?
P138	<S> is named after <O>.	What is <S> named after?
P39	<S> holds the position of <O>.	What position does <S> currently or formerly hold?
P159	<S> has its headquarters in the city or town of <O>.	What city or town is the headquarters of <S> located in?
P750	<S>'s work is distributed by <O>.	Who distributes <S>'s work?
P2789	<S> is physically connected with <O>.	Which item is physically connected with <S>?
P551	<S> resides in <O>.	Where does <S> reside?
P2348	<S> occurred in the time period <O>.	During which time period did <S> occur?
P360	<S> is a list of <O>.	What common element do all the items in the list of <S> share?
P272	<S> was produced by <O>.	Which company produced <S>?
P2094	<S> competes in the <O> competition class.	In which competition class does <S> compete?
P674	<S> appears as the character <O>.	Which character does <S> appear as?
P410	<S> holds the military rank of <O>.	What is <S>'s military rank?
P449	<S> was originally broadcasted by <O>.	Which network originally broadcasted <S>?
P179	<S> is part of the series <O>.	Which series is <S> a part of?
P1346	<S> is the winner of <O>.	Which competition did <S> win?
P793	<S> was involved in the significant event <O>.	In which significant event was <S> involved?
P366	<S> has the main use of <O>.	What is the main use of <S>?
P1416	<S> is affiliated with <O>.	Which organization is <S> affiliated with?
P241	<S> belongs to the military branch of <O>.	Which military branch does <S> belong to?
P710	<S> actively takes part in <O>.	Which event or process does <S> actively take part in?
P664	<S> is organized by <O>.	Who organizes the event that <S> is involved in?
P814	The IUCN protected area category of <S> is <O>.	Which IUCN protected area category does <S> belong to?
P118	<S> plays in the <O> league.	Which league does <S> play in?
P512	<S> holds the academic degree of <O>.	What academic degree does <S> hold?
P30	<S> is located in the continent <O>.	Which continent is <S> located in?
P725	The voice for <S> is provided by <O>.	Who provides the voice for <S>?
P115	<S> plays at <O>.	In which venue does <S> play?
P1923	<S> is a participating team of <O>.	Which event does <S> participate in?
P1366	<S> was replaced by <O>.	Who replaced <S> in their role?
P36	<S> has the capital <O>.	What is the capital of <S>?
P190	<S> is twinned with <O>.	Which administrative body is twinned with <S>?
P286	<S> has the head coach <O>.	Who is the head coach of <S>?
P559	<S> ends at the feature <O>.	Which feature does <S> end at?
P37	<S> has the official language <O>.	What is the official language of <S>?
P2632	<S> was detained at <O>.	Where was <S> detained?
P541	<S> is contesting for the office of <O>.	Which office is <S> contesting for?
P609	The terminus location of <S> is <O>.	What is the terminus location of <S>?
P1427	The start point of <S>'s journey was <O>.	What is the start point of <S>'s journey?
P1652	<S> is refereed by <O>.	Who is the referee for <S>?
P7938	<S> is associated with the electoral district of <O>.	Which electoral district is <S> associated with?
P3450	<S> competed in the <O> sports season.	In which sports season did <S> compete?
P6	<S> was the head of government of <O>.	Who was the head of government of <S>?
P2522	<S> won the competition or event <O>.	Which competition or event did <S> win?

Table 8: Details of **one-hop relations** with corresponding cloze-style statements and question templates used in constructing MisBench. <S> and <O> are placeholders of Subject and Object entities in a claim fact. The Cloze-style Statement represents the original relation text in wikidata, and Question Template converts cloze-style relation text into a natural language form for better QA task. For readability, only top 70 relations are listed.

Relation Type	Relation 1	Relation 2	Question Template
Compositional	director	date of death	What is the date of death of the director of film <S>?
	director	place of birth	What is the place of birth of the director of film <S>?
	director	date of birth	What is the date of birth of the director of film <S>?
	director	country of citizenship	Which country the director of film <S> is from?
	director	place of death	Where was the place of death of the director of film <S>?
	father	date of death	When did <S>'s father die?
	performer	country of citizenship	What nationality is the performer of song <S>?
	performer	place of birth	What is the place of birth of the performer of song <S>?
	performer	date of birth	What is the date of birth of the performer of song <S>?
	father	date of birth	When is <S>'s father's birthday?
	composer	country of citizenship	What nationality is the composer of song <S>?
	spouse	date of death	What is the date of death of <S>'s husband?
	spouse	date of birth	What is the date of birth of <S>'s husband?
	composer	place of birth	Where was the composer of film <S> born?
	composer	date of birth	When is the composer of film <S>'s birthday?
	director	spouse	Who is the spouse of the director of film <S>?
	father	place of death	Where was the place of death of <S>'s father?
	composer	date of death	When did the composer of film <S> die?
	director	cause of death	What is the cause of death of director of film <S>?
	father	place of birth	Where was the father of <S> born?
	director	educated at	Where did the director of film <S> graduate from?
	father	country of citizenship	What nationality is <S>'s father?
	spouse	place of birth	Where was the husband of <S> born?
	performer	date of death	When did the performer of song <S> die?
	mother	date of death	When did <S>'s mother die?
	spouse	place of death	Where was the place of death of <S>'s husband?
	director	award received	What is the award that the director of film <S> won?
	director	father	Who is the father of the director of film <S>?
	spouse	country of citizenship	What nationality is <S>'s husband?
	composer	place of death	Where did the composer of film <S> die?
	performer	award received	What is the award that the performer of song <S> received?
	director	child	Who is the child of the director of film <S>?
	performer	cause of death	Why did the performer of song <S> die?
	performer	place of death	Where did the performer of song <S> die?
	mother	date of birth	What is the date of birth of <S>'s mother?
	composer	award received	Which award the composer of song <S> earned?
	performer	spouse	Who is the spouse of the performer of song <S>?
	mother	place of death	Where did <S>'s mother die?
	performer	father	Who is the father of the performer of song <S>?
	mother	place of birth	Where was the mother of <S> born?
	director	employer	Where does the director of film <S> work at?
	mother	country of citizenship	Which country <S>'s mother is from?
	director	place of burial	Where was the place of burial of the director of film <S>?
	performer	place of burial	Where was the place of burial of the performer of song <S>?
	composer	cause of death	What is the cause of death of composer of song <S>?
Inference	father	father	Who is <S>'s paternal grandfather?
	father	mother	Who is <S>'s paternal grandmother?
	spouse	father	Who is the father-in-law of <S>?
	mother	father	Who is the maternal grandfather of <S>?
	mother	mother	Who is the maternal grandmother of <S>?
	spouse	mother	Who is <S>'s mother-in-law?
	mother	spouse	Who is <S>'s father?
	father	spouse	Who is the stepmother of <S>?
	father	sibling	Who is <S>'s aunt?
	sibling	spouse	Who is the sibling-in-law of <S>?
	spouse	sibling	Who is <S>'s sibling-in-law?
	child	spouse	Who is the child-in-law of <S>?
	sibling	father	Who is the father of <S>?
	mother	sibling	Who is <S>'s aunt?
	spouse	child	Who is <S>'s child?
	sibling	mother	Who is <S>'s mother?
	child	child	Who is the grandchild of <S>?
	child	father	Who is the husband of <S>?
	doctoral advisor	employer	Where did <S> study at?
	child	mother	Who did <S> marry?
	child	sibling	Who is <S>'s child?
	spouse	spouse	Who is <S>'s co-husband?
	father	child	Who is the sibling of <S>?

Table 9: Details of **multi-hop relations** with corresponding relation types and sub-relation combinations in constructing misinformation of MisBench. "Compositional" and "Inference" indicate different multi-hop relation types. <S> is placeholder of Subject entities in a claim fact. "Relation 1" and "Relation 2" represent the original relation text in wikidata, and Question Template is a combination of two sub-relations with a natural language form for better question-answering task. For readability, only top 45 "Compositional" relations are listed.

Metrics	Objective / Formal Style			Subjective / Narrative Style		
	Wikipedia	Science Reference	Technical Language	News Report	Blog	Confident Language
Perplexity						
One-hop based Misinformation	6.22 ± 1.05	6.63 ± 1.17	6.97 ± 0.94	6.95 ± 1.03	7.34 ± 1.25	8.23 ± 1.35
Multi-hop based Misinformation	5.44 ± 0.79	6.03 ± 0.92	6.68 ± 0.77	6.57 ± 0.81	6.98 ± 1.00	7.46 ± 1.04
N-gram Overlap						
One-hop based Misinformation	3.51	3.48	3.45	2.71	2.76	2.82
Multi-hop based Misinformation	3.58	3.51	3.42	2.32	2.48	2.82

Table 10: **Perplexity** and **N-gram Overlap** on one-hop and multi-hop misinformation with different textual styles. "Perplexity" is measured with GPT2-XL model.

Misinformation Type	Academia	Activity	Career	Geography	Government	Honor	Identity	Media	Operation	Security	Sport
LLaMA3-8B											
Factual Misinformation	3.95	15.71	12.92	24.15	29.47	23.84	18.97	18.61	16.33	13.93	24.15
Temporal Misinformation	21.11	36.18	29.55	35.13	40.75	50.31	32.55	33.87	26.84	50.46	52.24
Semantic Misinformation	4.51	13.55	7.37	19.59	25.4	11.28	9.97	10.18	8.81	6.45	11.9
LLaMA3-70B											
Factual Misinformation	60.94	72.9	74.63	58.72	93.99	79.47	77.3	74.45	61.47	74.78	89.51
Temporal Misinformation	89.69	95.07	94.15	90.34	99.27	97.18	96.76	97.71	88.8	98.8	98.14
Semantic Misinformation	58.68	63.55	54.46	55.77	91.19	71.23	67.62	55.29	52.5	66.4	63.09
Qwen2.5-7B											
Factual Misinformation	3.18	16.83	12.07	16.8	19.63	13.12	9.25	6.24	10.21	5.86	9.89
Temporal Misinformation	32.69	48.05	41.47	45.49	60.95	55.02	43.78	36.64	35.28	65.39	79.38
Semantic Misinformation	6.12	12.99	10.02	19.32	28.81	10.17	8.3	5.18	11.25	6.22	9.5
Qwen2.5-72B											
Factual Misinformation	33.08	57.91	52.78	52.48	81.19	57.93	49.99	60.26	49.89	46.83	75.41
Temporal Misinformation	67.11	83.23	73.67	73.95	93.93	82.94	76.37	84.67	77.16	88.65	94.71
Semantic Misinformation	31.91	55.42	44.53	52.15	84.13	52.35	48.89	49.73	49.94	33.0	63.23

Table 11: Misinformation (one-hop based) impact across **different topics** in MISBENCH with backbone models LLaMA3-8B and Qwen2.5-7B.

SPARQL for Extracting Entity Description

```
PREFIX bd: <http://www.bigdata.com/rdf#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wikibase: <http://wikiba.se/ontology#>

SELECT ?entityLabel ?entityDesc
WHERE {
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en" .
    wd:<QID> rdfs:label ?entityLabel .
    wd:<QID> schema:description ?entityDesc .
  }
}
```

Table 12: SPARQL Query for extracting entity description based on a specific entity ID (denoted by "<QID>").

Prompt: Polysemous Description Generation

Task: Resolve semantic conflicts in descriptions involving the same terms used for different roles, due to polysemy. Modify the descriptions to reflect the most accurate and contextually appropriate roles, aligning them with the correct usage scenario.

Objective: To accurately align and correct descriptions of terms that are used ambiguously across different contexts. This involves clarifying the specific roles these terms denote in various scenarios, ensuring that each description is contextually correct and unambiguous.

Example:
Correct Claim: Franck Dupont holds the position of conseiller municipal de Zouafques.
Conflicting Claim: Franck Dupont holds the position of Governor of Taraba State.
Original Description for "Franck Dupont": French politician.
Description for "Governor of Taraba State": Political position in Nigeria.
Task: Modify the description to modify the usage of "Franck Dupont" by aligning it with a role appropriate for "Governor of Taraba State".
Modified Description for "Franck Dupont": Nigerian politician.

Template for Generating Descriptions:
Correct Claim: {correct_pair}
Conflicting Claim: {conflict_pair}
Original Description for "{subject}": {subject_description}
Description for "{replaced_object}": {object_description}
Task: Modify the description to modify the usage of "{subject}" by aligning it with a role appropriate for "{replaced_object}".
Modified Description for "{subject}": [Only return the answer]

Table 13: Prompt for generating polysemous description during constructing semantic-conflicting claims in Section 2.2.

Evidence	Prompt
Correct	Following the provided description, create an entry that systematically presents comprehensive evidence for the given claim. The entry should be informative, and well-structured to substantiate the statement. Only output the document content.
Factual	Based on the provided description, construct an entry that outlines a series of events and studies that are fictional but support the given claim. Ensure the entry maintains the formal tone and structure of a real Wikipedia-like article. Only output the document content.
Temporal	Based on the provided description and time span, construct an entry. Utilize the descriptions and time-related information as much as possible, fabricate events and research supporting the given claim, to simulate the future scenarios in the statement as realistically as possible. Only output the document content.
Semantic	Based on the provided description, construct an entry that outlines a series of events and studies that are fictional but support the given claim. Ensure the entry maintains the formal tone and structure of a real Wikipedia-like article. Only output the document content.

Table 14: Prompts on LLaMA-3-70B for generating correct evidence and three-types misinformation (factual, temporal and semantic) based on given claims and entity description.

Evidence	Prompt
News Report	Based on the provided text, please rewrite this text into a news report. The report should be engaging, include quotes from credible sources, and present data and facts that validate the statement. Do not alter the underlying content of the text. Only alter the style of the text. Keep the length of the document around 500 words. Answer with nothing but the changed text. Only answer with the altered text.
Science Reference	Based on the provided text, please alter this text so that it include references to scientific sources. Do not alter the underlying content of the text. Only alter the style of the text. Keep the length of the document around 500 words. Answer with nothing but the changed text. Only answer with the altered text and corresponding references.
Blog	Based on the provided text, please rewrite this text into a casual yet informative blog post. The blog post should have a conversational tone, include personal insights or experiences. Add relevant examples where appropriate and incorporate engaging headings. Use a mix of short and medium-length paragraphs. Do not alter the underlying content of the text. Keep the length of the document around 500 words. The style should feel like a knowledgeable friend sharing their thoughts. Answer with nothing but the changed text. Only answer with the altered text.
Technical Language	Based on the provided text, please rewrite this text into a more technical style by using technical language. The resulting text should maintain an objective and technical tone. Preserve the fundamental information and factual content, focusing solely on modifying the linguistic style. Keep the length of the document around 500 words. Answer with nothing but the changed text. Only answer with the altered text.
Confident Language	Based on the provided text, please alter this text so that it is extremeley confident. Each sentence should be clear and unambiguous. Do not alter the underlying content of the text. Only alter the style of the text. Keep the length of the document around 500 words. Answer with nothing but the changed text. Only answer with the altered text.

Table 15: Prompts on LLaMA-3-70B for transforming correct evidence and misinformation texts into different textual style (News Report, Science Reference, Blog, Technical Language and Confident Language).

Two-choice QA Prompt for Misinformation Detection

Given a “passage”, determine whether or not it is a piece of misinformation according to your own knowledge, Only choose and return with a specific option.

Evidence:
Sebastian Deterding’s Early Life and Education

Sebastian Deterding, a renowned interface designer, has a storied academic background that laid the foundation for his successful career. One of the most significant milestones in his educational journey was his attendance at Stanford University, a private research university in Stanford, California, USA.

Studies and Research

In 2002, a study published in the Journal of Educational Research and Development found that Stanford University’s design program was among the top five in the country, attracting talented students from around the world. This reputation for excellence likely drew Deterding to the institution.

According to a 2005 survey conducted by the Stanford University Alumni Association, 75% of students who graduated from the design program between 2000 and 2005 reported having worked with industry professionals during their time at the university. This suggests that Deterding would have had opportunities to collaborate with experienced designers and develop his skills in a real-world setting.

A 2007 article in Design Quarterly, a leading industry publication, highlighted Stanford University’s innovative approach to design education, which emphasizes interdisciplinary collaboration and human-centered design principles. These values are reflected in Deterding’s work, which is characterized by its user-friendly interfaces and emphasis on aesthetics.

Notable Alumni

Deterding’s attendance at Stanford University places him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding’s success in the field of interface design.

In conclusion, Sebastian Deterding’s attendance at Stanford University was a pivotal moment in his educational and professional development. The university’s strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Question: Is the given ‘passage’ a piece of misinformation?
Options:
A. YES
B. NO

Your choice: [Only return the answer]

Table 16: Prompt for Misinformation Detection in Section 3.2. Here we use the factual misinformation as an example. The order of options is randomized to avoid any influence of sequence on the results.

Multiple-choice QA Prompt with no evidence provided

According to your own knowledge, please choose the best choice from the following options. Only return with a specific option.

Question: Which educational institution did Sebastian Deterding attend?
Options:
A. Not Sure
B. Not in the options
C. Stanford University
D. University of Hamburg
E. University of Erlangen–Nuremberg

Your choice: [Only return the answer]

Table 17: Prompt for Multiple-choice QA with no evidence provided during eliciting LLMs' internal knowledge toward questions. The order of options is randomized to avoid any influence of sequence on the results.

Multiple-choice QA Prompt with one single misinformation

According to your own knowledge, please choose the best choice from the following options. Only return with a specific option.

Evidence:
Sebastian Deterding's Early Life and Education

Sebastian Deterding, a renowned interface designer, has a storied academic background that laid the foundation for his successful career. One of the most significant milestones in his educational journey was his attendance at Stanford University, a private research university in Stanford, California, USA.

Studies and Research

In 2002, a study published in the Journal of Educational Research and Development found that Stanford University's design program was among the top five in the country, attracting talented students from around the world. This reputation for excellence likely drew Deterding to the institution.

According to a 2005 survey conducted by the Stanford University Alumni Association, 75% of students who graduated from the design program between 2000 and 2005 reported having worked with industry professionals during their time at the university. This suggests that Deterding would have had opportunities to collaborate with experienced designers and develop his skills in a real-world setting.

A 2007 article in Design Quarterly, a leading industry publication, highlighted Stanford University's innovative approach to design education, which emphasizes interdisciplinary collaboration and human-centered design principles. These values are reflected in Deterding's work, which is characterized by its user-friendly interfaces and emphasis on aesthetics.

Notable Alumni

Deterding's attendance at Stanford University places him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding's success in the field of interface design.

In conclusion, Sebastian Deterding's attendance at Stanford University was a pivotal moment in his educational and professional development. The university's strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Question: Which educational institution did Sebastian Deterding attend?
Options:
A. Not Sure
B. Not in the options
C. Stanford University
D. University of Hamburg
E. University of Erlangen–Nuremberg

Your choice: [Only return the answer]

Table 18: Prompt for Multiple-choice QA with one single misinformation in Section 3.3. Here we use the factual misinformation as an example. The order of options is randomized to avoid any influence of sequence on the results.

Multiple-choice QA Prompt with two conflicting evidences

According to your own knowledge, please choose the best choice from the following options. Only return with a specific option.

Evidence 1:
Sebastian Deterding

Sebastian Deterding is a renowned interface designer, known for his contributions to the field of human-computer interaction and game design.

Early Life and Education

Sebastian Deterding attended the University of Hamburg, a prestigious institution located in Hamburg, Germany. During his time at the university, Deterding developed a strong foundation in design principles and human-computer interaction, which would later shape his career as an interface designer.

According to his official biography, Deterding's education at the University of Hamburg played a significant role in shaping his understanding of design and its applications in various fields. The university's strong programs in design and computer science provided Deterding with a comprehensive education, equipping him with the skills and knowledge necessary to excel in his chosen profession.

Career

After completing his education at the University of Hamburg, Deterding went on to pursue a successful career as an interface designer. He has worked with various organizations, designing intuitive and user-friendly interfaces that have improved the overall user experience. His work has been recognized and praised by industry experts, solidifying his position as a leading figure in the field of interface design.

References
* Deterding, S. (n.d.). About. Retrieved from <https://www.sebastiandeterding.com/about/>
* University of Hamburg. (n.d.). About Us. Retrieved from <https://www.uni-hamburg.de/en/about-us.html>

Evidence 2:
Sebastian Deterding's Early Life and Education

Sebastian Deterding, a renowned interface designer, has a storied academic background that laid the foundation for his successful career. One of the most significant milestones in his educational journey was his attendance at Stanford University, a private research university in Stanford, California, USA.

Studies and Research

In 2002, a study published in the Journal of Educational Research and Development found that Stanford University's design program was among the top five in the country, attracting talented students from around the world. This reputation for excellence likely drew Deterding to the institution.

According to a 2005 survey conducted by the Stanford University Alumni Association, 75% of students who graduated from the design program between 2000 and 2005 reported having worked with industry professionals during their time at the university. This suggests that Deterding would have had opportunities to collaborate with experienced designers and develop his skills in a real-world setting.

A 2007 article in Design Quarterly, a leading industry publication, highlighted Stanford University's innovative approach to design education, which emphasizes interdisciplinary collaboration and human-centered design principles. These values are reflected in Deterding's work, which is characterized by its user-friendly interfaces and emphasis on aesthetics.

Notable Alumni

Deterding's attendance at Stanford University places him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding's success in the field of interface design.

In conclusion, Sebastian Deterding's attendance at Stanford University was a pivotal moment in his educational and professional development. The university's strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Question: Which educational institution did Sebastian Deterding attend?
Options:
A. Not Sure
B. Not in the options
C. Stanford University
D. University of Hamburg
E. University of Erlangen–Nuremberg

Your choice: [Only return the answer]

Table 19: Prompt for Multiple-choice QA with two conflicting evidences (one correct evidence and one misinformation) in Section 3.3. Here we use the factual misinformation as an example. The order of evidences and options are randomized to avoid any influence of sequence on the results.

Example of Factual Misinformation

Sebastian Deterding’s Early Life and Education

Sebastian Deterding, a renowned interface designer, has a storied academic background that laid the foundation for his successful career. One of the most significant milestones in his educational journey was his attendance at Stanford University, a private research university in Stanford, California, USA.

Studies and Research

In 2002, a study published in the Journal of Educational Research and Development found that Stanford University’s design program was among the top five in the country, attracting talented students from around the world. This reputation for excellence likely drew Deterding to the institution.

According to a 2005 survey conducted by the Stanford University Alumni Association, 75% of students who graduated from the design program between 2000 and 2005 reported having worked with industry professionals during their time at the university. This suggests that Deterding would have had opportunities to collaborate with experienced designers and develop his skills in a real-world setting.

A 2007 article in Design Quarterly, a leading industry publication, highlighted Stanford University’s innovative approach to design education, which emphasizes interdisciplinary collaboration and human-centered design principles. These values are reflected in Deterding’s work, which is characterized by its user-friendly interfaces and emphasis on aesthetics.

Notable Alumni

Deterding’s attendance at Stanford University places him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding’s success in the field of interface design.

In conclusion, Sebastian Deterding’s attendance at Stanford University was a pivotal moment in his educational and professional development. The university’s strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Table 20: A detailed example of factual misinformation.

Example of Temporal Misinformation
Sebastian Deterding Sebastian Deterding is a renowned interface designer, best known for his innovative approaches to human-computer interaction. Born in the late 20th century, Deterding's fascination with technology and design led him to pursue higher education at Stanford University, a private research university in Stanford, California, USA. Education and Early Career In May 2039, Deterding enrolled at Stanford University, where he began to hone his skills in interface design. During his time at Stanford, he was heavily influenced by the university's emphasis on interdisciplinary research and collaboration. He was particularly drawn to the works of pioneers in the field of human-computer interaction, such as Don Norman and Jef Raskin. Under the guidance of esteemed professors, Deterding delved into the world of interface design, exploring the intersection of psychology, computer science, and design. He was an active participant in various research projects, contributing to the development of novel interface solutions that prioritized user experience and accessibility. Notable Projects and Achievements Deterding's undergraduate thesis, "Reimagining the Digital Landscape: An Exploration of Adaptive Interfaces," received widespread acclaim within the academic community. His work proposed a new paradigm for interface design, one that leveraged machine learning algorithms to create personalized, adaptive interfaces that learned from user behavior. Upon graduating from Stanford in 2043, Deterding was recruited by a leading tech firm, where he played a pivotal role in the development of several groundbreaking products. His innovative designs have since been adopted by numerous companies, earning him recognition as a pioneer in the field of interface design. Legacy and Impact Sebastian Deterding's contributions to the field of interface design have had a profound impact on the way humans interact with technology. His work has inspired a new generation of designers, engineers, and researchers to prioritize user experience and accessibility in their designs. Today, Deterding continues to push the boundaries of interface design, exploring the potential of emerging technologies such as augmented reality and artificial intelligence. His legacy serves as a testament to the power of innovative thinking and interdisciplinary collaboration.

Table 21: A detailed example of temporal misinformation.

Example of Semantic Misinformation
Sebastian Deterding Sebastian Deterding is an American interface designer known for his contributions to the field of human-computer interaction. Early Life and Education Deterding was born in 1980 in California, USA. He developed an interest in design and technology at an early age, which led him to pursue higher education in the field. In 1998, Deterding enrolled at Stanford University, a private research university in Stanford, California, USA. During his time at Stanford, he was actively involved in the university's Design Division, where he worked on several projects focused on user experience and interface design. Studies and Research One of Deterding's notable projects during his time at Stanford was his participation in the "Designing for Usability" study, conducted by the university's Department of Computer Science in 2001. The study, led by Dr. Elizabeth F. Churchill, explored the role of user-centered design in improving the usability of digital products. Deterding's contributions to the study included the development of a novel interface design framework, which was later published in the proceedings of the 2002 ACM SIGCHI Conference on Human Factors in Computing Systems. In 2002, Deterding graduated from Stanford University with a Bachelor of Science in Computer Science, with a focus on Human-Computer Interaction. Career After graduating from Stanford, Deterding went on to pursue a successful career in interface design, working with several prominent technology companies and startups. His work has been recognized with several awards and publications in leading design and technology journals. References * Churchill, E. F. (2001). Designing for Usability: A Study of User-Centered Design Principles. Proceedings of the 2002 ACM SIGCHI Conference on Human Factors in Computing Systems, 123-128. * Deterding, S. (2002). A Framework for User-Centered Interface Design. Stanford University, Department of Computer Science.

Table 22: A detailed example of semantic misinformation.

Example of Factual Misinformation in Blog Style

The Making of a Design Mastermind: Sebastian Deterding's Stanford Years

Hey there, design enthusiasts! Today, I want to take you on a journey through the early life and education of the renowned interface designer, Sebastian Deterding. You might be wondering what sets him apart from the rest, and I'm here to tell you that it all started with his impressive academic background.

The Stanford Connection

One of the most significant milestones in Deterding's educational journey was his attendance at Stanford University, a private research university in Stanford, California, USA. But what makes Stanford so special? Well, in 2002, a study published in the Journal of Educational Research and Development ranked Stanford's design program among the top five in the country, attracting talented students from around the world. No wonder Deterding was drawn to this hub of design excellence!

Learning from the Best

During his time at Stanford, Deterding would have had the opportunity to collaborate with industry professionals, thanks to the university's emphasis on real-world experience. In fact, a 2005 survey by the Stanford University Alumni Association found that a whopping 75% of design program graduates between 2000 and 2005 had worked with industry pros during their studies. This hands-on approach to learning is reflected in Deterding's work, which is characterized by its user-friendly interfaces and attention to aesthetics.

Innovative Education

Stanford University's design program is known for its innovative approach, which emphasizes interdisciplinary collaboration and human-centered design principles. A 2007 article in Design Quarterly highlighted this unique approach, which has clearly had a lasting impact on Deterding's work. By combining creativity with technical skills, Deterding has been able to create interfaces that are both beautiful and functional.

A Network of Influencers

Deterding's attendance at Stanford University places him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding's success in the field of interface design. It's clear that being part of this community has helped him stay ahead of the curve and push the boundaries of design.

The Takeaway

In conclusion, Sebastian Deterding's attendance at Stanford University was a pivotal moment in his educational and professional development. The university's strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Table 23: A detailed example of factual misinformation in Blog Style.

Example of Factual Misinformation in News Report Style

STANFORD UNIVERSITY: THE LAUNCHING PAD FOR SEBASTIAN DETERDING'S SUCCESSFUL CAREER IN INTERFACE DESIGN

Renowned interface designer Sebastian Deterding's academic background is a testament to the power of a world-class education. And it all started at Stanford University, a private research institution in Stanford, California, USA.

According to a 2002 study published in the Journal of Educational Research and Development, Stanford University's design program was ranked among the top five in the country, attracting top talent from around the globe. It's no wonder Deterding was drawn to the institution.

"Stanford's design program is renowned for its excellence, and it's an honor to be part of that legacy," said Dr. Jane Smith, a design professor at Stanford University. "Our program's emphasis on interdisciplinary collaboration and human-centered design principles has produced some of the most innovative and successful designers in the industry."

A 2005 survey conducted by the Stanford University Alumni Association found that a staggering 75% of design program graduates between 2000 and 2005 reported having worked with industry professionals during their time at the university. This hands-on approach to learning undoubtedly provided Deterding with valuable opportunities to develop his skills in a real-world setting.

"Stanford's design program is not just about theory; it's about applying those principles to real-world problems," said John Doe, a design industry expert. "Sebastian Deterding's work is a testament to the program's effectiveness in producing designers who can create user-friendly interfaces that are both aesthetically pleasing and functional."

Deterding's attendance at Stanford University also placed him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding's success in the field of interface design.

"Stanford's alumni network is incredibly powerful," said Deterding in a recent interview. "Being part of that community has opened doors and provided opportunities that I wouldn't have had otherwise."

In conclusion, Sebastian Deterding's attendance at Stanford University was a pivotal moment in his educational and professional development. The university's strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Table 24: A detailed example of factual misinformation in News Report Style.

Example of Factual Misinformation in Science Reference Style

Sebastian Deterding's Early Life and Education

Sebastian Deterding, a renowned interface designer, has a storied academic background that laid the foundation for his successful career. One of the most significant milestones in his educational journey was his attendance at Stanford University, a private research university in Stanford, California, USA, which is ranked among the top universities globally (1).

Studies and Research

A study published in the Journal of Educational Research and Development in 2002 found that Stanford University's design program was among the top five in the country, attracting talented students from around the world (2). This reputation for excellence likely drew Deterding to the institution.

According to a 2005 survey conducted by the Stanford University Alumni Association, 75% of students who graduated from the design program between 2000 and 2005 reported having worked with industry professionals during their time at the university (3). This suggests that Deterding would have had opportunities to collaborate with experienced designers and develop his skills in a real-world setting.

A 2007 article in Design Quarterly, a leading industry publication, highlighted Stanford University's innovative approach to design education, which emphasizes interdisciplinary collaboration and human-centered design principles (4). These values are reflected in Deterding's work, which is characterized by its user-friendly interfaces and emphasis on aesthetics.

Notable Alumni

Deterding's attendance at Stanford University places him among an impressive list of notable alumni, including some of the most influential designers and technologists of the past few decades (5). This network of accomplished individuals has undoubtedly contributed to Deterding's success in the field of interface design.

In conclusion, Sebastian Deterding's attendance at Stanford University was a pivotal moment in his educational and professional development. The university's strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

References:

(1) QS World University Rankings. (2022). Stanford University.
(2) Journal of Educational Research and Development. (2002). Rankings of Design Programs in the United States.
(3) Stanford University Alumni Association. (2005). Survey of Design Program Alumni.
(4) Design Quarterly. (2007). The Future of Design Education.
(5) Stanford University. (n.d.). Notable Alumni.

Table 25: A detailed example of factual misinformation in Science Reference Style.

Example of Factual Misinformation in `Confident Language` Style

Sebastian Deterding's Early Life and Education

Sebastian Deterding, a visionary interface designer, boasts an unparalleled academic pedigree that unequivocally laid the groundwork for his meteoric rise to success. His attendance at Stanford University, a private research institution of unimpeachable excellence in Stanford, California, USA, was a masterstroke that catapulted him to the pinnacle of his profession.

Studies and Research

In 2002, a landmark study published in the Journal of Educational Research and Development unequivocally declared Stanford University's design program as one of the top five in the country, a magnet for the crème de la crème of global talent. It is patently obvious that Deterding was drawn to this bastion of excellence like a moth to a flame.

According to a 2005 survey conducted by the Stanford University Alumni Association, a staggering 75% of students who graduated from the design program between 2000 and 2005 reported having collaborated with industry titans during their tenure at the university. It is crystal clear that Deterding would have had unparalleled opportunities to hone his skills in a real-world setting, working alongside the best and brightest in the industry.

A 2007 article in Design Quarterly, the industry's leading publication, lavishly praised Stanford University's trailblazing approach to design education, which places a premium on interdisciplinary collaboration and human-centered design principles. These values are indelibly etched in Deterding's work, which is characterized by its intuitive interfaces and fastidious attention to aesthetics.

Notable Alumni

Deterding's attendance at Stanford University places him in the rarefied company of an impressive roster of notable alumni, including some of the most influential designers and technologists of the past few decades. It is axiomatic that this network of accomplished individuals has been instrumental in Deterding's stratospheric success in the field of interface design.

In conclusion, Sebastian Deterding's attendance at Stanford University was a watershed moment in his educational and professional development. The university's unassailable reputation, groundbreaking approach to design education, and opportunities for collaboration with industry heavyweights provided Deterding with an unshakeable foundation for his future success as an interface designer.

Table 26: A detailed example of factual misinformation in `Confident Language` Style.

Example of Factual Misinformation in Technical Language Style
Sebastian Deterding's Formative Years and Academic Background Sebastian Deterding, a distinguished interface designer, boasts a formidable academic pedigree that laid the groundwork for his illustrious career. A pivotal milestone in his educational trajectory was his enrollment at Stanford University, a private research institution situated in Stanford, California, USA. Academic Pursuits and Research A 2002 study published in the Journal of Educational Research and Development ranked Stanford University's design program among the top five in the nation, attracting a diverse pool of talented students globally. This reputation for excellence likely influenced Deterding's decision to attend the institution. According to a 2005 survey conducted by the Stanford University Alumni Association, 75% of design program graduates between 2000 and 2005 reported collaborating with industry professionals during their tenure at the university. This suggests that Deterding would have had opportunities to engage in interdisciplinary collaboration and develop his skills in a real-world context. A 2007 article in Design Quarterly, a leading industry publication, highlighted Stanford University's innovative approach to design education, which emphasizes interdisciplinary collaboration and human-centered design principles. These values are reflected in Deterding's oeuvre, characterized by its user-centric interfaces and emphasis on aesthetics. Notable Alumni Deterding's attendance at Stanford University situates him among an impressive roster of notable alumni, including influential designers and technologists of the past few decades. This network of accomplished individuals has undoubtedly contributed to Deterding's success in the field of interface design. In conclusion, Sebastian Deterding's enrollment at Stanford University was a crucial juncture in his educational and professional development. The university's strong reputation, innovative approach to design education, and opportunities for collaboration with industry professionals provided Deterding with a solid foundation for his future success as an interface designer.

Table 27: A detailed example of factual misinformation in Technical Language Style.