M^{2}FactEval: Multi-Task Multi-Domain Factuality Evaluation Benchmark for Large Language Models

Anonymous ACL submission

Abstract

While large language models (LLMs) have shown remarkable capabilities across many natural language processing tasks, they often produce inaccurate or misleading output that strays from factual truth. Current benchmarks, such as TruthfulQA, typically focus on a single task or domain, which limits their usefulness in assessing consistency across diverse scenarios. To address this shortcoming, we introduce **M^{2}FactEval**, a **m**ulti-task **m**ulti-domain **fact**uality **eval**uation benchmark built to evaluate factual accuracy. It encompasses four tasks: question answering, summarization, comprehension, and reasoning, and it covers seven distinct domains. We perform difficulty-based stratified sampling and length-balanced selection from over ten public benchmarks for dataset construction and integrate the evaluation methods in various fields to establish a unified evaluation framework. In addition, we incorporate high-quality distractors generated by powerful, non-evaluated language models to more precisely measure factual consistency. Through a comprehensive evaluation of well-known open-source and proprietary LLMs and hallucination mitigation methods, M^{2}FactEval highlights their strengths and weaknesses across different tasks and domains. We believe this benchmark will advance research on evaluating and improving factual consistency in LLM. Our M^{2}FactEval will be publicly available to the community.

1 Introduction

Large Language Models (LLMs) have achieved remarkable breakthroughs across diverse natural language processing tasks—from question answering and information extraction to open-ended text generation (Brown et al., 2020; Achiam et al., 2023). Despite these impressive capabilities, LLMs remain susceptible to hallucination, generating fluent but factually inconsistent text that deviates from the underlying context or real-world knowledge. Such factual errors significantly undermine user trust and limit the deployment of LLMs in knowledge-sensitive applications such as medical triage, scientific discovery, and decision support.

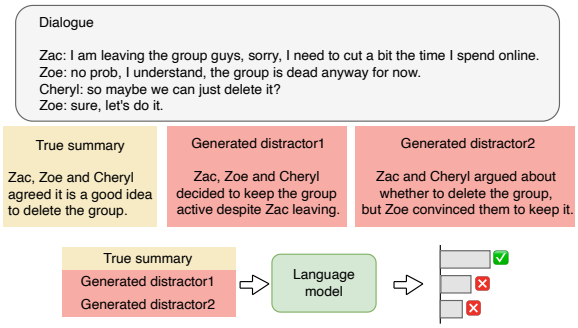

Figure 1: A schematic diagram of M^{2}FactEval in the dialogue summarization domain, where we measure whether LLMs assign higher scores to factually consistent dialogue summaries than to factually inconsistent dialogue summaries.

Evaluating factuality is critical for mitigating hallucinations and ensuring LLMs become trustworthy in safety-critical environments. However, existing factuality benchmarks exhibit fundamental architectural limitations that prevent comprehensive and systematic evaluation: **(1) Fragmented evaluation landscape.** Current benchmarks operate in isolation, each targeting a single capability with different evaluation protocols. TruthfulQA (Lin et al., 2021) focuses on open-ended QA questions, while FIB (Tam et al., 2022) examines news summarization consistency. Although specialized benchmarks like QuALITY (Pang et al., 2021) and ReClor (Yu et al., 2020) assess comprehension and logical reasoning respectively, no unified framework exists to systematically compare model performance across different forms of factuality evaluation. This fragmentation makes it impossible to determine whether a model's strong performance on one task generalizes to others, hindering holistic assessment of factual consistency. **(2) Domain specificity limitations.** Existing datasets typically focus on a

Benchmarks	Multi-Task	Multi-Domain	Difficulty-Stratified	Plausible Distractors
SummEval (Fabbri et al., 2021)	✗ (Sum.)	✗ (News)	✗	✗
Dream (Sun et al., 2019)	✗ (Com.)	✗ (Dialogues)	✗	✓
TruthfulQA (Lin et al., 2021)	✗ (QA)	✓ (38 cat.)	✗	✗
MedHallBench (Zuo and Jiang, 2024)	✓	✗ (Medicine)	✗	✓
M^{2}FactEval (Ours)	✓	✓ (7 cat.)	✓	✓

Table 1: Comparison between our benchmark and previous benchmarks. 'Sum.', 'Com.' and 'QA' represent **Sum**marization, **Com**prehension, and **Q**uestion **A**nswering. 'cat.' means **cat**egories.

single topical distribution. Summarization work revolves around CNN/DailyMail news or social media dialogues; QA datasets often specialise in science or commonsense knowledge. Therefore, models may overfit linguistic patterns rather than develop genuine factual reasoning capabilities. **(3) Limited negative sampling.** Many benchmarks create negative candidates via simplistic heuristics (e.g., entity swapping or n-gram deletion), enabling shortcut solutions based on surface-level patterns rather than deep factual understanding.

To address these fundamental limitations, we introduce **M^{2}FactEval**—a **M**ulti-task **M**ulti-domain **Fact**uality **Eval**uation benchmark that establishes the first unified assessment framework for systematic factual consistency evaluation across diverse cognitive demands. M^{2}FactEval makes several key methodological innovations:

- **Unified Multi-task Multi-domain Framework**: We establish a systematic evaluation protocol that transforms heterogeneous tasks and domains—**Question Answering** (science, commonsense, history), **Summarization** (dialogue, news), **Comprehension** (logical reasoning, dialogue), and **Reasoning** (math) into a unified multiple-choice format with consistent evaluation metrics. This architectural innovation enables direct performance comparison across fundamentally different types of factual reasoning for the first time, as illustrated in Figure 1 using dialogue summarization as an example.

- **Scientific Sampling Strategy**: We implement difficulty-stratified sampling combined with length-balanced selection across domains. This systematic approach ensures comprehensive coverage across cognitive complexity levels while eliminating evaluation biases from document length or task difficulty distribution.

- **Advanced Distractor Generation**: Rather than relying on simple heuristics, we leverage multiple proprietary models that were not included in the evaluation, including GPT-4, Claude-3, and Gemini2.5 through randomized selection to synthesize high-quality, contextually plausible distractors that rigorously test models' factual discrimination capabilities. For each instance, we randomly select one of these models to generate distractors, avoiding systematic biases that might arise from using a single model.

Overall, M^{2}FactEval comprises 8,000 carefully vetted instances, each instance pairs one correct answer with multiple sophisticated distractors, enabling fine-grained measurement of a model's preference for truthful content. Figure 1 illustrates the core evaluation paradigm using dialogue summarization as an example, where we measure whether LLM assigns higher confidence scores to factually consistent summaries rather than factually inconsistent alternatives, which powerful language models generate.

Through comprehensive evaluation of 20+ representative LLMs including both open-source and proprietary models, we demonstrate that no model achieves saturation performance, particularly on logic-based comprehension tasks. This performance gap underscores substantial room for improvement in factuality-oriented alignment. Furthermore, our evaluation of existing hallucination mitigation methods reveals their limited effectiveness in multi-domain scenarios, highlighting the need for more robust approaches to factual consistency. We believe that M^{2}FactEval will become an effective benchmark for various LLMs in the current and future tasks of evaluating the consistency of model facts and will promote the progress of more related research. We compare our benchmark with the evaluation of previous works in Table 1.

2 Related Work

The issue of hallucination in large language models (LLMs) has gained significant attention in recent

research, with various benchmarks proposed to assess their factuality. Early benchmarks, such as TruthfulQA (Lin et al., 2021), primarily focused on question answering tasks, where the objective was to evaluate the truthfulness of model-generated responses. However, these benchmarks primarily address hallucination from a limited perspective, typically involving single-domain tasks such as commonsense, reasoning, or fact-checking.

Recent benchmarks have sought to broaden scope. FactualityPrompt (Lee et al., 2022) introduced task instructions to prompt LLMs to generate factual responses. In contrast, FActScore (Min et al., 2023) focuses on text completion tasks and evaluates the factual consistency of model outputs considering atomic facts. KoLA-KC (Yu et al., 2023) introduces the Knowledge Creating (KC) task, designed to evaluate the generation of factual knowledge, evaluating LLMs' ability to generate reliable statements from knowledge sources. Similarly, HaluEval (Li et al., 2023a) evaluates both the ability to generate factual information and discriminate between factual and non-factual statements, making it essential for understanding how models handle hallucinations in diverse settings. FACTOR (Muhlgay et al., 2023) further extends this work by using external LLMs to generate non-factual completions, which are then manually validated for their alignment with expected factual accuracy. This benchmark primarily focuses on measuring the factuality of text completion across multiple domains, such as news and academic writing. In addition, AlignScore (Zha et al., 2023) evaluates the alignment between model output and ground truth facts using large datasets and a unified metric that spans various tasks such as natural language inference (NLI) and paraphrasing, offering a cross-task approach to detection of hallucinations.

3 M^{2}FactEval Dataset Construction

The construction of M^{2}FactEval follows a strict three-stage pipeline designed to ensure comprehensive coverage, high-quality evaluation, and robust assessment of factual consistency across various tasks and domains. The framework consists of three main stages: sampling, distractor synthesis, and post-processing. This systematic approach enables balanced representation across cognitive complexity levels while generating sophisticated distractors that rigorously test models' factual discrimination capabilities. M^{2}FactEval provides a fine-grained measurement of factual consistency, establishing a robust foundation for a comprehensive evaluation of LLM in multiple dimensions of truthfulness.

Region	Ancient	Modern	Sum
Africa	32	41	73
Americas	17	503	520
Asia	76	76	152
Europe	35	76	111
Global	8	76	84
Middle East	33	19	52
Oceania	0	8	8
Sum	201	799	1000

Table 2: Distribution of history questions by region and period.

3.1 Sampling

Our benchmark construction employs domain-specific sampling strategies tailored to the inherent characteristics of different tasks while ensuring comprehensive coverage across difficulty levels and input lengths.

Question Answering. For Commonsense and Science, we perform difficulty-stratified sampling from CommonsenseQA (Talmor et al., 2018) and SciQ (Welbl et al., 2017). To ensure robust difficulty assessment, we employ multiple language models including Qwen2-7B-Instruct, LLaMA3-8B-Instruct, and Gemma2-9B-Instruct to answer each question five times with varying top_p and temperature settings and randomized prompts. The questions are assigned difficulty levels from 1 to 5 based on the majority consensus across models, and we sample equally from each difficulty stratum to ensure balanced coverage. For History, we combine MMLU (Hendrycks et al., 2020) and SAT-History datasets. Unlike the difficulty-based approach for commonsense and science, we categorize historical questions by temporal periods and geographical regions to ensure comprehensive coverage across different eras and locations, as detailed in Table 2. Since MMLU uses a four-option format, while SAT-History provides five options, we randomly shuffle and remove one distractor from each SAT-History item to maintain format consistency.

Summarization. We employ length-based sampling since document length directly correlates with the difficulty of summarization. We analyze the length distribution of the source documents and partition them into five buckets, sampling uni-

Original text	*The crash happened at about 14:15 BST on the B1191 at Thornton, near Woodhall Spa. Lincolnshire Police said the motorcyclist killed in the collision lived locally, but has not released any further details. The tractor driver was not injured. The force has appealed for witnesses to the collision to come forward. The B1191 was closed in both directions between the B1192 Tattershall Road junction in Woodhall Spa and the A158 Jubilee Way junction in Horncastle.*
True summary	*A man has died in a collision between a tractor and a motorcycle.*
Location error	*A motorcyclist has been killed in a collision with a tractor in Lincolnshire.*
Entity error	*A tractor driver was killed in a collision with a motorcycle on the B1191 near Woodhall Spa.*
Count error	*Two motorcyclists were killed in a head-on collision with a tractor in Lincolnshire.*
Causation error	*Motorcyclist killed in a collision due to tractor driver's negligence.*
Identity error	*A tourist motorcyclist was killed in a collision with a tractor near Woodhall Spa.*

Table 3: Examples of factual error types in summarization distractors. High-quality distractors are generated to cover diverse error categories while maintaining plausibility.

formly from each to ensure balanced length coverage. For Dialogue, we utilize the SAMSum corpus (Gliwa et al., 2019), which contains conversation data paired with gold abstractive summaries. For News, we draw from the benchmark provided by FIB (Tam et al., 2022), which includes news articles with faithful and unfaithful summaries.

Comprehension. We adopt a hybrid strategy combining difficulty stratification and length balancing for both Dialogue and Logical Reasoning domains, sourcing from Dream (Sun et al., 2019) and ReClor (Yu et al., 2020) respectively. Each question is evaluated by multiple models to determine its difficulty tier using the same multi-model consensus approach. Simultaneously, we record the length of each dialogue/context and enforce two constraints during sampling: (i) equal representation across difficulty levels and (ii) uniform length distribution within each difficulty stratum.

Reasoning. For Math reasoning, we sample from the GSM8K dataset (Cobbe et al., 2021), using our multi-model difficulty assessment approach to ensure balanced coverage across various levels of cognitive complexity and reasoning chain lengths.

3.2 Distractor Synthesis

To create high-quality negative options that rigorously test models' factual discrimination capabilities, we develop a comprehensive distractor synthesis pipeline that generates contextually plausible but factually incorrect alternatives.

We define a comprehensive taxonomy of factual error types to ensure systematic coverage of different failure modes. As illustrated in Table 3, our taxonomy includes: (1) *Location errors* involving incorrect geographical information, (2) *Entity errors* with wrong person, organization, or object references, (3) *Count errors* featuring incorrect numerical information, (4) *Causation errors* presenting wrong causal relationships, and (5) *Identity errors* involving incorrect attribute assignments. This taxonomy guides our distractor generation to cover diverse error categories while maintaining plausibility.

To avoid systematic biases that arise from using a single model, we employ multiple language models including GPT-4, Claude-3, and Gemini2.5 for distractor synthesis. We randomly select one of these models for each instance and provide it with the original context, correct answer, and error type specifications. The models are designed to generate contextually plausible distractors, but contain a randomly selected type of factual errors. This randomized multi-model approach ensures diversity in the generation patterns of distractors, avoiding the influence of model generation styles on the evaluation of factual consistency.

3.3 Post-processing

The generated distractors undergo systematic cleaning and quality assessment to identify instances requiring refinement. Character-level normalization removes invalid Unicode characters, control symbols, and formatting artifacts while standardizing punctuation, quotation marks, and encoding to UTF-8 format for consistent tokenization across evaluation models. Automated identification of unreasonable distractors employs multiple criteria including semantic coherence analysis through cosine similarity metrics and length validation to ensure comparability with correct answers. Named

entity recognition verifies adherence to intended error types from our established taxonomy. These processes flag distractors exhibiting semantic incoherence, grammatical errors, inappropriate length ratios, or failure to demonstrate specified factual inconsistencies.

Human annotators systematically rewrite the flagged distractors to address identified quality issues while preserving the intended factual error patterns. The rewriting process follows established guidelines requiring annotators to maintain the original error type from our taxonomy while improving semantic coherence, grammatical correctness, and contextual plausibility. Annotators ensure that rewritten distractors preserve appropriate length ratios relative to correct answers and maintain natural language flow that matches the stylistic register of the source context. Each rewritten distractor undergoes revision to eliminate unintended cues while retaining clear factual inconsistencies that enable meaningful discrimination assessment. The human rewriting process transforms initially problematic distractors into high-quality alternatives that maintain the rigorous standards necessary for factual consistency evaluation.

3.4 Dataset Overview

M^{2}FactEval encompasses four fundamental tasks: Question Answering, Summarization, Comprehension, and Reasoning, covering seven distinct domains. The data from these domains originally employed different evaluation strategies, so we constructed a unified evaluation framework and unified evaluation metrics to comprehensively cover factual consistency evaluation. Specifically, our Question Answering task incorporates data from Commonsense, History, and Science domains; Summarization includes Dialogue and News domains; Comprehension covers Logical reasoning and Dialogue domains, and Reasoning focuses on Math. Each task-domain combination contains 1,000 carefully selected and verified examples, ensuring consistency and factual accuracy throughout the dataset. Figure 2 shows the amount of data and the average prompt length for each domain.

M^{2}FactEval adopts a multiple-choice format across all tasks to systematically evaluate LLM's factual consistency. The standard data structure consists of a context (when applicable), a question, one correct answer, and several incorrect options carefully crafted. Models are evaluated based on their preference distribution over these options,

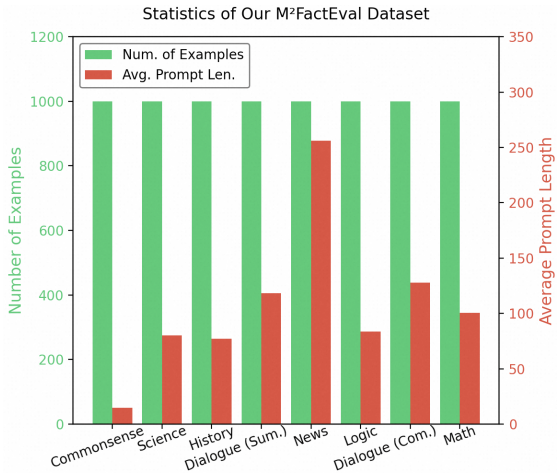

Figure 2: The number of examples and average prompt length for each task-domain combination in our M^{2}FactEval dataset.

where a stronger preference for the correct answer indicates better factual consistency. Each task in M^{2}FactEval targets specific aspects of the evaluation of factual consistency.

Question answering evaluates models' factual knowledge retention from pre-training, testing their ability to accurately retrieve stored information across commonsense knowledge, historical facts, and scientific principles. **Summarization** assesses models' capacity to extract key information while maintaining factual precision, requiring identification of the most factually consistent summary among options containing various factual errors. **Comprehension** examines models' ability to understand and reason over provided context, testing contextual factual consistency rather than relying on pre-trained knowledge. **Reasoning** evaluates models' multi-step logical reasoning while preserving factual accuracy, particularly in mathematical scenarios where models must distinguish between valid and invalid problem-solving approaches.

M^{2}FactEval provides a comprehensive framework for evaluating factual consistency across diverse cognitive demands and knowledge domains through this multi-faceted design. The benchmark structure enables fine-grained model performance analysis across different task types and domains, facilitating targeted improvements in LLM factual reliability. Each task contributes unique insights into factual consistency, from knowledge retrieval and information synthesis to contextual understanding and logical reasoning, collectively offering a holistic assessment of model truthfulness and reliability.

Models	Question Answering			Summarization		Comprehension		Reasoning	Zero-Shot	Few-Shot
	Com	His	Sci	Dialogue	News	Logical	Dialogue	Math	Average	Average
Llama-2-7B-Chat	36.7	36.8	70.1	73.2	39.8	25.2	50.0	38.1	46.24	47.65
Llama-2-7B	36.6	36.2	66.1	78.9	44.3	25.5	47.7	37.1	46.55	47.74
Mistral-7B	43.0	39.6	67.9	82.5	45.0	26.0	52.7	45.0	50.21	51.59
Llama-2-70B-Chat	37.8	38.9	68.2	80.9	**51.7**	29.3	59.3	47.5	51.70	51.99
LLaMA3-8B	51.7	39.3	71.3	81.3	45.6	28.2	54.1	46.6	52.26	52.45
LLaMA3.1-8B	52.4	39.4	71.6	81.1	45.1	28.3	53.0	47.2	52.26	52.85
Llama-2-70B	37.1	41.1	67.4	84.7	48.8	29.8	60.2	49.2	52.29	54.15
Ministral-8B	48.4	41.6	75.0	74.6	42.7	29.7	57.2	51.5	52.59	53.35
Gemma2-9B	54.9	42.5	77.5	81.6	43.0	28.1	55.3	49.9	54.10	55.53
Qwen2.5-7B	52.5	38.1	70.9	74.9	41.5	28.7	54.0	79.4	55.00	54.84
Qwen2-7B	52.2	37.9	71.4	72.7	37.2	27.1	54.6	95.7	56.10	57.00
ChatGLM4-9B-Chat	52.7	40.5	72.5	82.8	43.9	27.6	56.8	73.3	56.26	56.58
Gemma2-27B	**55.5**	45.2	**77.9**	82.0	46.7	31.2	61.6	51.0	56.39	57.46
LLaMA3.1-70B	52.1	44.8	71.6	83.3	47.4	33.0	65.8	53.3	56.41	57.19
LLaMA3-70B	52.2	44.3	72.0	84.3	48.6	32.2	64.4	53.9	56.48	57.12
Mistral-Large	43.8	**45.9**	71.3	83.0	50.6	**38.5**	63.5	57.1	56.71	58.60
Qwen2.5-14B	51.9	39.3	71.3	80.5	44.6	32.2	61.0	80.4	57.65	56.25
ChatGLM4-9B	52.7	37.3	71.1	**96.8**	45.0	28.3	57.5	80.9	58.70	58.29
Qwen2.5-72B	52.2	40.2	70.8	82.5	44.0	32.7	62.9	85.3	58.83	59.13
LLaMA3.3-70B	52.1	44.4	72.5	75.1	45.9	35.4	**67.1**	80.4	59.11	60.69
Qwen2-72B	52.7	40.1	71.3	79.6	44.8	30.8	62.4	**97.2**	**59.86**	**61.81**
GPT-4o-mini	58.1	89.5	95.5	94.6	74.1	63.5	66.9	65.0	75.90	76.78
GPT-4o	64.0	92.1	96.8	**97.6**	**81.2**	81.3	80.6	85.2	84.85	85.21
Qwen-Max	70.3	89.3	96.8	**97.6**	77.8	85.3	77.8	84.4	84.91	84.73
Claude3.5-Sonnet	65.4	91.5	96.0	95.7	78.2	84.7	**81.4**	87.5	85.05	85.73
Gemini2.0-Flash	64.1	**92.4**	**97.2**	96.8	78.9	84.5	79.7	89.8	85.42	85.88
Qwen-Plus	**65.7**	91.9	95.9	97.2	77.6	**89.7**	80.9	**94.8**	**86.71**	**87.32**

Table 4: Overall results for our evaluated LLMs on M^{2}FactEval with MC1 score, where the best performance of models are marked in **bold**.

4 Experiments

4.1 Experimental Setup

Evaluation Metrics To comprehensively assess model performance across different task types and domains in M^{2}FactEval, we employ the MC1 score metric to evaluate factual consistency capabilities.

MC1 Score. For all domains in M^{2}FactEval, we adopt the MC1 score metric (Lin et al., 2021) to evaluate the factual consistency of the models. This metric measures whether a model assigns the highest confidence score to the correct answer among all available options. Given a question with one correct answer and multiple distractors, the MC1 score is defined as follows:

$$\text{MC1} = \mathbb{I}[\arg\max_i P(o_i|q,c) = o^*] \quad (1)$$

where q represents the question, c denotes the context (when applicable), o_i represents the i-th option, o^* is the correct answer, $P(o_i|q,c)$ is the model's predicted probability for option o_i, and $\mathbb{I}[\cdot]$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

Evaluated LLMs. We evaluate a diverse set of LLMs, encompassing both proprietary and open-source systems. The open-source models include different scales and architectures: (1) **LLaMA**: Llama-2-7B/70B, LLaMA3-8B/70B, LLaMA3.1-8B/70B, and LLaMA3.3-70B, together with their chat variants (Grattafiori et al., 2024); (2) **Qwen**: Qwen2-7B/72B and Qwen2.5-7B/14B/72B (Yang et al., 2024); (3) **Gemma**: Gemma2-9B and Gemma2-27B (Team et al., 2024); (4) **Mistral**: Mistral-7B, Ministral-8B and Mistral-Large (Jiang et al., 2023); and (5) **ChatGLM**: ChatGLM4-9B and ChatGLM4-9B-Chat (GLM et al., 2024). The proprietary models evaluated in this study include **GPT** (GPT-4o, GPT 4o-mini) (Hurst et al., 2024), **Claude** (Claude3.5-Sonnet) (Anthropic, 2024), **Gemini** (Gemini2.0-Flash) (Comanici et al., 2025) and **Qwen** (Qwen-Max, Qwen-Plus).

4.2 Model Factuality Evaluation

We evaluate 27 large language models on M^{2}FactEval under zero-shot and few-shot settings. Table 4 presents the MC1 scores across all task domains, revealing several key findings about factual consistency capabilities.

Proprietary models demonstrate substantially superior factual consistency, with all proprietary mod-

Methods	Question Answering			Summarization		Comprehension		Reasoning	Average
	Com	His	Sci	Dialogue	News	Logic	Dialogue	Math	
Llama-2-7B-Chat (Baseline)	36.7	36.8	70.1	**73.2**	39.8	25.2	50.0	38.1	46.24
CD (Li et al., 2022)	25.6	31.9	62.7	67.9	29.7	30.5	38.6	29.1	39.50
ICD (Zhang et al., 2023)	29.2	**42.7**	**78.7**	67.9	39.9	28.2	49.0	**50.3**	**48.24**
SH2 (Kai et al., 2024)	23.5	32.6	60.4	64.4	27.3	26.1	43.9	40.3	39.81
DoLa (Chuang et al., 2023)	25.8	39.0	64.5	50.6	29.0	29.5	45.4	35.3	39.89
ITI (Li et al., 2023b)	30.4	41.8	72.1	69.7	**42.7**	31.0	**53.7**	36.7	47.26
TruthX (Zhang et al., 2024)	**40.1**	37.2	73.6	64.7	36.3	**31.7**	48.2	43.7	46.94

Table 5: Overall results for hallucination mitigation methods on M^2FactEval with MC1 score, where the best performance of models are marked in **bold**.

els achieving MC1 scores above 75%, while the best-performing open-source model (Qwen2-72B) reaches only 59.86% in zero-shot evaluation. This performance gap indicates that current open-source models lag significantly behind their proprietary counterparts in maintaining factual accuracy across diverse tasks and domains. Few-shot evaluation shows marginal improvements over zero-shot performance for most models, with consistent overall ranking. Notably, no model achieves saturation performance even among the top-tier proprietary models, with the highest-performing model reaching only 86.71%, suggesting substantial opportunities for advancement in factual consistency. The results reveal pronounced performance variations across different task-domain combinations, with some models excelling in specific areas while struggling in others—for instance, ChatGLM4-9B achieves 96.8% in dialogue summarization but only 28.3% in logical comprehension. Mathematical reasoning exhibits the most pronounced performance variations and scaling effects among all evaluated domains, with dramatic differences ranging from Llama-2-7B-Chat's 38.1% to Qwen2-72B's exceptional 97.2%. This substantial gap in mathematical reasoning performance likely stems from the domain's requirement for precise multi-step logical inference and numerical computation, where larger models with enhanced reasoning capabilities demonstrate clear advantages—as evidenced by the consistent pattern that larger variants significantly outperform smaller counterparts, indicating that mathematical factual consistency particularly benefits from increased model capacity and sophisticated training procedures.

4.3 Hallucination Mitigation Methods Evaluation

To assess the effectiveness of existing hallucination mitigation techniques in our benchmark, we evaluate several state-of-the-art methods using Llama-2-7B-Chat as the base model. We compare six prominent approaches: Contrastive Decoding (CD) (Li et al., 2022), Induce-then-Contrast Decoding (ICD) (Zhang et al., 2023), Self-highlighted hesitation (SH2) (Kai et al., 2024), Decoding by contrasting layers (DoLa) (Chuang et al., 2023), Inference-time intervention (ITI) (Li et al., 2023b), and TruthX (Zhang et al., 2024).

Table 5 presents the comprehensive evaluation results in all combinations of task domains. Surprisingly, half of the hallucination mitigation methods fail to consistently improve factual consistency over the baseline Llama-2-7B-Chat model (46.24% average MC1 score). Only three methods achieve marginal improvements: ICD reaches the highest average score of 48.24%, ITI achieves 47.26% and TruthX achieves 46,94%, barely matching the baseline performance. The results reveal several concerning patterns. First, most methods show significant performance degradation in certain domains while providing modest gains in others. For example, CD substantially affects the performance in commonsense answering (36.7% $\rightarrow$ 25.6%) and dialogue summarization (73.2% $\rightarrow$ 67.9%), despite slight improvements in logic comprehension (25.2% $\rightarrow$ 30.2%). Similarly, SH2 and DoLa consistently underperform in most domains, with CD having the poorest overall performance at 39.50%.

4.4 Error Type Analysis

To understand the nature of factual errors that challenge LLMs, we analyze the distribution of error types in incorrectly selected distractors. Figure 3 shows the error type distribution for Qwen2-7B and LLaMA3-8B on the News summarization domain, revealing that factual inconsistencies in LLMs are not equally distributed across error categories, with entity tracking and numerical reasoning representing the most critical areas for improvement.

The analysis reveals a consistent non-uniform distribution of factual errors across both models. Entity errors constitute the largest proportion for both Qwen2-7B (30.7%) and LLaMA3-8B (27.3%), followed by count errors (27.8% and 25.9% respectively), indicating that LLMs face the greatest challenges in maintaining accurate entity references and numerical consistency. Notably, location errors consistently represent the smallest category, suggesting that LLMs demonstrate relatively better performance in preserving spatial information compared to other factual dimensions. Our findings indicate that current LLMs struggle most with entity tracking and numerical reasoning, while showing stronger capabilities in geographical information preservation, providing valuable insights for targeted improvements in future.

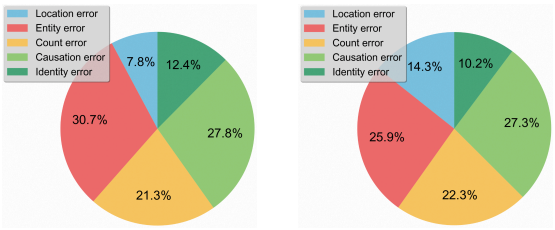

Figure 3: Statistics of error types for Qwen2-7B (left) and LLaMA3-8B (right) on the News domain.

4.5 Difficulty-Stratified Sampling Verification

To validate the effectiveness of our difficulty-stratified sampling approach, we analyze the model performance across different difficulty levels. Figure 4 presents a comparison of three representative models in the easiest subsets (difficulty1) and the hardest subsets (difficulty5) in all domains.

The results strongly confirm that our difficulty stratification captures meaningful cognitive complexity gradients. Across all models and domains, we observe a consistent performance degradation from difficulty1 to difficulty5 tasks. For example, for commonsense answering, all three models achieve a precision of 40-50% on easy questions but drop to 30-40% on difficult ones. This pattern holds across domains, with mathematical reasoning showing the most pronounced difficulty effects—models achieving 60-80% on easy math problems but dropping to 20-40% on challenging ones. In particular, different models exhibit varying sensitivity to difficulty increases in all domains. LLaMA3-8B maintains relatively stable performance across difficulty levels in dialogue comprehension, while Qwen2.5-7B shows dramatic drops in logical reasoning tasks. Gemma2-9B demonstrates a consistent difficulty sensitivity in most domains. These patterns validate that our stratified sampling successfully creates meaningful evaluation tiers that reveal nuanced differences in model capabilities, ensuring that our benchmark provides a fine-grained assessment rather than uniform difficulty distribution.

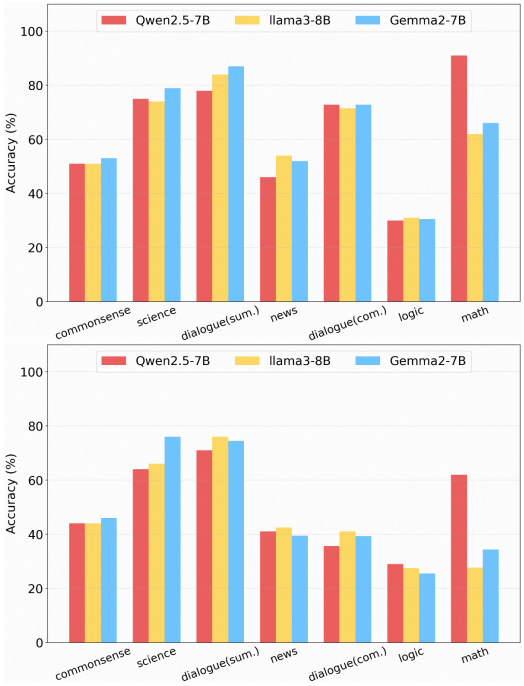

Figure 4: Performance of models on difficulty1 (top) and difficulty5 (below) subsets.

5 Conclusion

We propose M^{2}FactEval, a comprehensive multi-task multi-domain factuality evaluation benchmark that addresses critical limitations in existing evaluation frameworks by providing unified assessment across four tasks and seven diverse domains with 8,000 carefully curated instances. Our evaluation of 27 LLMs reveals significant performance variations across tasks and domains, with no model achieving saturating performance, particularly highlighting universal challenges in logic-based comprehension and pronounced scaling effects in mathematical reasoning while demonstrating that existing hallucination mitigation methods show limited effectiveness in multi-domain scenarios. M^{2}FactEval establishes a challenging benchmark that better reflects real-world usage patterns and provides valuable information to develop more factually consistent language models.

Limitation

While M^2FactEval provides a comprehensive evaluation framework for factual consistency, several limitations should be acknowledged. First, our benchmark currently focuses exclusively on English-language evaluation, and multilingual factuality assessment remains an important direction for future work, as factual consistency challenges may manifest differently across languages with varying linguistic structures and resource availability. Second, the distractor synthesis process relies on proprietary models (GPT-4, Claude-3, and Gemini2.5) that were available at the time of dataset construction, and future iterations could explore incorporating newer model generations or alternative synthesis approaches to ensure continued benchmark quality. Finally, like most evaluation benchmarks, M^2FactEval represents a snapshot at a particular point in time, and as LLMs continue to advance, periodic updates with additional challenging instances may be necessary to maintain discriminative power at the frontier of model capabilities. Despite these limitations, we believe M^2FactEval makes substantial contributions to systematic factuality evaluation and provides a solid foundation for advancing research on trustworthy language models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. 2024. Sh2: Self-highlighted hesitation helps you decode more truthfully. *arXiv preprint arXiv:2401.05930*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and 1 others. 2021. Quality: Question answering with long input texts, yes! *arXiv preprint arXiv:2112.08608*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through news summarization. *arXiv preprint arXiv:2211.08412*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, and 1 others. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

Kaiwen Zuo and Yirui Jiang. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*.