Teaching Language Models to Check Grounded Claim Factuality with Human Test-Taking Strategies

Anonymous ACL submission

Abstract

Grounded claim factuality checking is important for large language model (LLM) applications such as retrieval-augmented generation, as it helps users assess the correctness of generated outputs. Existing metrics using entailment classifiers require dataset-specific threshold tuning, while LLM-based approaches often use direct prompting, which underutilises the reasoning capabilities of LLMs. We address this by formulating grounded claim factuality checking as a true/false reading comprehension task and prompting LLMs with explicit test-taking strategies for efficient reasoning. Our method reduces token usage by over 80% compared to unguided open-ended reasoning, and achieves competitive performance to more expensive alternatives across two factuality benchmarks, setting a new state of the art on one. To further reduce inference cost, we train small language models (SLMs) to replace LLMs in the checking pipeline. Using supervised fine-tuning (SFT) and a self-revision mechanism, the SLMs learn to improve their factuality judgements. Experimental results show that the resulting SLMs perform on par with strong baselines, combining low inference costs with generating supporting rationales to support interpretability. Code and datasets will be released upon acceptance.

1 Introduction

Large language models (LLMs) are applied to a wide variety of generation tasks, including summarisation, question answering, and conversational agents. Many of these tasks require the generated content to be consistent with grounding documents, but LLM outputs do not always satisfy this requirement. LLMs are prone to making up statements that are not supported by the given sources (Zhang et al., 2024; McKenna et al., 2023). As a result, detecting such hallucinations is critical for conditioned generation tasks to ensure the trustworthiness and reliability of the generated outputs.

Recent work often regards factuality evaluation as a textual entailment task (Lei et al., 2025; Zha et al., 2023; Laban et al., 2022). This line of research mainly employs a classifier to predict an entailment score, which has the advantage of being lightweight and computationally efficient. However, the entailment scores require a dataset-specific threshold to convert them into explicit *True* and *False* predictions. This limitation motivates us to design an evaluation metric that outputs factuality judgements and generalises across datasets without threshold tuning.

Given a suitably designed prompt, LLMs can directly decide whether a claim is grounded in a source document (Xu et al., 2024; Luo et al., 2023). This is supported by impressive understanding and reasoning capabilities (QwenTeam, 2025; DeepSeek-AI, 2025; Dubey et al., 2024), which leads to a methodology that uses LLMs for claim factuality checking, known as LLM-as-judge. A major challenge for this methodology lies in designing prompts that fully leverage the LLM's reasoning ability (Xu et al., 2024; Liu et al., 2023b; Luo et al., 2023). An additional benefit of LLM-as-judge is that, with an appropriate prompt, LLMs can generate rationales for judgements, which provide information for identifying errors in the claim and locating evidence in the source document.

However, LLMs are computationally intensive when deployed at scale. A potential alternative is to adapt small language models (SLMs), which have recently been shown to handle certain reasoning tasks when trained appropriately (QwenTeam, 2025). Their lightweight architecture and ability to generate reasoned judgements without threshold tuning motivate further exploration of SLMs for claim factuality checking, including strategies to train and adapt them for this purpose.

In our work, we design a claim checking pipeline that incorporates LLM-as-judge into a human test-taking strategy. We first decompose each claim

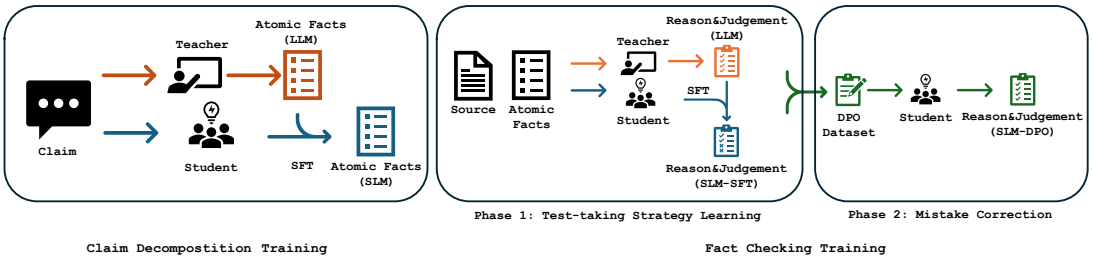

Figure 1: Overview of SLM training. **Claim decomposition training** (left) utilises the output from an LLM (teacher) as the reference output in SFT for the SLM (student). **Fact checking training, phase 1:** (middle) LLM (teacher) outputs are used to train an SLM (student) to follow a test-taking strategy, in which the model separately assesses each atomic fact. **Phase 2:** (right) We further refine the student model by pairing its incorrect outputs with correct outputs from the teacher model, then fine-tuning the student's reasoning output via DPO.

into atomic facts to simplify the subsequent fact checking step. In the second step, motivated by the similarity between fact checking and True/False reading exercises in language proficiency tests, we design a prompt that leverages test-taking strategies commonly used in such assessments. Unlike previous approaches, in which the model directly outputs its judgement or reasoning without guidance, we require the model to follow the strategies to reason and make judgements. Experimental results demonstrate the effectiveness of this pipeline, achieving competitive performance to more expensive alternatives across two factuality benchmarks, including a new state of the art on one of them.

We then propose a training method to adapt a general purpose SLM for claim factuality evaluation. Given the limited generalisation capacity of SLMs, we train separate models for claim decomposition and fact checking. Using our LLM-as-judge strategy, we employ an LLM as a teacher for both tasks, using supervised fine-tuning (SFT) to train the student SLM. For the fact checking step, the student model further revises its rationales through direct preference optimisation (DPO) (Rafailov et al., 2023) using the reference output from the teacher model. In this way, the student model first learns the test-taking strategy and then strengthens this ability by correcting its own mistakes, resulting in higher fact checking accuracy than relying solely on SFT. The overall training process is shown in Figure 1.

Our contributions are three-fold:

- Inspired by test-taking strategies, we design a prompt for grounded claim checking that simulates how humans process it in assessments. It achieves state of the art on one benchmark, while generating rationales with judgements.
- We propose a two-phase training procedure and demonstrate that the resulting SLMs achieve performance approaching that of LLMs, despite being a fraction of the size.
- To the best of our knowledge, this is the first work to apply SLMs to claim factuality evaluation by generating reasoned judgements, demonstrating their potential to evaluate factuality in a reasoning pipeline and provide interpretable judgements efficiently.

2 Related Work

2.1 Factuality Evaluation Metrics

Several widely used factuality metrics assess a claim's factual consistency by measuring its semantic similarity to the grounding document (Zhang et al., 2019; Ye et al., 2024), but they fail to capture small textual differences that alter factual content while maintaining high semantic similarity.

The resemblance between factuality evaluation and textual entailment has motivated the use of entailment classifiers for factuality (Lei et al., 2025; Tang et al., 2024a; Zha et al., 2023; Laban et al., 2022), which has gradually become the mainstream paradigm. Entailment classifiers are trained to detect whether a hypothesis is supported by a given premise at sentence-level. In claim factuality checking, they often need to truncate or chunk the source document, leading to information loss that decreases performance.

Another line of work (Fabbri et al., 2022; Deutsch et al., 2021; Wang et al., 2020) regards factuality evaluation as a question answering (QA) problem, measuring a claim's factuality by comparing answers generated when conditioning on the claim versus the grounding document. While effective, these QA-based methods often rely on

complex multi-stage pipelines. In contrast, our work adopts the reading comprehension concept for claim factuality evaluation while simplifying the pipeline into two concise steps.

2.2 LLM in Factuality Evaluation

LLMs are both generators of content to be evaluated and powerful tools for developing evaluation metrics. One application of LLMs is synthesising training data to improve factuality metrics (Lei et al., 2025; Tang et al., 2024a; Feng et al., 2024). These approaches often augment training data by generating adversarial or challenging examples.

Another use case of LLMs leverages their understanding and reasoning ability to directly perform factuality evaluation, commonly referred as LLM-as-judge. Some studies (Luo et al., 2023; Gekhman et al., 2023) evaluated the performance of directly prompting an LLM to decide the correctness of a claim. While these works also investigate the benefits of LLM reasoning, the models are left to reason their judgements without explicit guidance. Other approaches (Xu et al., 2024; Liu et al., 2023b) explore incorporating pre-defined criteria into the prompt to guide judgements. However, the provided criteria are limited to error type definitions or scoring standards, rather than an explicit step-by-step reasoning plan for factuality evaluation.

In this paper, we utilise LLMs in both ways. Specifically, we 1) design a prompt that adapts a general purpose LLM into an effective factuality evaluation metric, and 2) use an LLM to generate training data for fine-tuning an SLM to perform factuality judgement. Unlike prior work, we prompt the LLM with a structured strategy that explicitly guides the evaluation step by step, leading to more systematic and interpretable judgements.

2.3 SLM Fine-tuning

Recent advances in LLMs have also accelerated progress in SLMs. Distillation from strong teacher models provides an efficient way to transfer capabilities, improving SLM performance on tasks such as reasoning and mathematics (QwenTeam, 2025; DeepSeek-AI, 2025). A common approach is to generate rationales using LLMs and fine-tune student SLMs via SFT (Feng et al., 2024; Gekhman et al., 2023; Jiang et al., 2023). However, these approaches rely solely on SFT to imitate the output from the teacher model. Previous works (QwenTeam, 2025; DeepSeek-AI, 2025) mention that performance can be improved by further leveraging the synthesised data with reinforcement learning following SFT. In this vein, we apply DPO after SFT, enabling the student model to improve its reasoning accuracy by learning from its own mistakes, thereby aligning better with the teacher model.

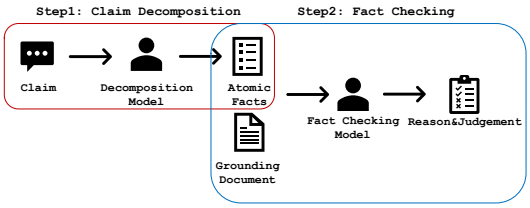

Figure 2: The pipeline for grounded claim checking.

3 Method

In this section, we first reformulate claim factuality checking as a reading comprehension task. Then we introduce a pipeline that adapts a general purpose LLM for factuality evaluation, taking inspiration from human test-taking strategies. Finally, we present our approach to fine-tune an SLM for use in the evaluation pipeline.

3.1 Problem Reformulation

Lei et al. (2025) defines that, given a grounding document doc and a claim c, c is grounded in doc if the statement *"According to doc, c is true"* is generally affirmed by a generic reader. This definition reformulates grounded claim checking into a True/False reading comprehension problem. Therefore, this paper tackles claim factuality evaluation by utilising generative language models to solve reading comprehension tasks, leveraging a strategy that a human examinee could apply in such tasks.

3.2 LLM-based Evaluation Pipeline

Recent LLMs achieve high scores on human language proficiency tests, including the reading comprehension problem formulated above (Achiam et al., 2023). Noting that human test-takers often use multi-step reasoning strategies to improve their performance on reading comprehension tests (Yapp et al., 2023; Al-Kiyumi et al., 2021), we incorporate such a strategy into prompt design, enabling LLMs to process grounded fact checking efficiently and systematically. By guiding the model's reasoning process, our approach aims to improve both accuracy and inference efficiency by shortening the model's reasoning outputs.

We first address a prevalent issue in factuality checking: claims can be complex and aggregate

multiple independent pieces of information. As shown in Figure 2, we decompose the overall evaluation into two simpler steps: (1) **Claim Decomposition**, which breaks a claim into atomic facts, and (2) **Atomic Fact Checking**, which verifies each fact against the grounding document. We describe these steps in detail in the following sections.

3.2.1 Claim Decomposition

Claims may contain information from multiple parts of a source document, where supporting evidence can be sparsely distributed. This increases the difficulty of factuality evaluation. To simplify the task, we prompt an LLM to decompose each claim into a set of atomic facts that can be verified individually. To do this, we adopt a few-shot prompt from Tang et al. (2024a) to guide claim decomposition. The prompt is shown in Table 11.

3.2.2 Fact Checking

During the language learning process, True/False reading comprehension practice is often applied to assess whether students correctly understand a given article. Instead of inferring an answer in an unconstrained manner, human examinees may use structured strategies to improve both efficiency and accuracy (Yapp et al., 2023; Al-Kiyumi et al., 2021). In practice, they often first verify the information that is explicitly mentioned in the grounding document and then verify whether the rest information can be inferred from it. Inspired by this, we design a fact checking prompt that explicitly guides the model's reasoning process using the following criteria to assess fact F against document D:

1. **C1** The object and subject of the claim are mentioned in D.
2. **C2** The descriptions of the object and the subject are explicitly supported by D.
3. **C3** The relation between the object and the subject is explicitly supported by D.
4. **C4** Any information that was not yet verified can be inferred from D.

Unlike previous work (Xu et al., 2024) that checks different error types individually, our criteria are applied sequentially to form a structured inference process. It verifies the explicitly mentioned information first, before reasoning whether the as-yet-unverified information is implied by the grounding document, as demonstrated in Figure 3.

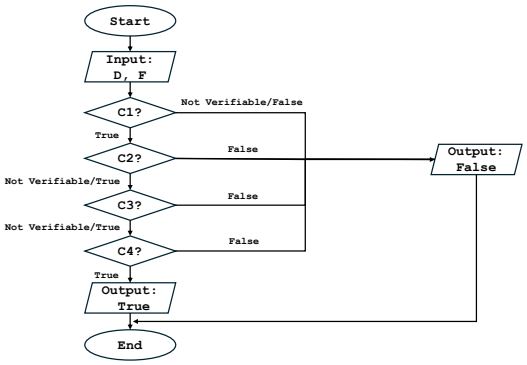

Figure 3: The process for checking a fact, F, against each criterion, with grounding document D.

For example, given a doc *"Ice can change into liquid water, and water can change into vapour, and vice versa."* and a claim *"Vapour can change into ice."*, both the subject *"vapour"* and the object *"ice"* are mentioned, and there are no descriptions of them to verify, therefore it passes **C1** and **C2**. For **C3**, the relation *"can change into ice"* is not explicitly mentioned with *"vapour"*. Previous methods would recognise it as an erroneous predicate but here we defer it to the next criterion. The last criterion, **C4**, analyses the information in two hops. *"Vice versa"* implies that the reversed processes are valid, and by combining two reversed processes the claim will be verified as *True* in the end.

In addition to outputting a binary decision (*True/False*), we require the model to provide a rationale for each criterion. This not only has the potential to help users locate evidence in the grounding document, but also to improve the consistency of model judgements with the inputs F and D. The full prompt used for this step is shown in Table 12.

3.3 Adapting SLMs for Claim Evaluation

SLMs are less computationally intensive than LLMs, but they are also less capable under the same prompt due to their smaller model scale. Previous work (QwenTeam, 2025; DeepSeek-AI, 2025) has shown that distillation can substantially enhance SLMs, allowing them to achieve performance comparable to LLMs on certain tasks, including those requiring reasoning abilities. Motivated by these findings, we train the SLM separately for the two steps to replace the LLM in our evaluation pipeline.

3.3.1 Claim Decomposition Learning

Given a claim c, we train the SLM to generate decomposed facts that imitate those generated by the teacher LLM. To construct a large decompo-

sition dataset $D_{De} = \{(c, \{f_{ref}\})\}$, we randomly sample sentences from grounding documents and existing claims, treating each randomly sampled sentence as c. We then use the teacher model to generate reference facts $\{f_{ref}\}$.

The learning objective can be written as:

$$\mathcal{L}(\theta) = \mathbb{E}_{(c,\{f_{ref}\})\sim D_{De}} \left[\log P_\theta(\{f_{ref}\} \mid c)\right],$$

where θ denotes the parameters of the SLM, and $\{f_{ref}\}$ here is considered as a whole string. We optimise the objective using SFT, aligning the SLM's generated facts $\{f_{gen}\}$ with the references $\{f_{ref}\}$.

3.3.2 Fact Checking Learning

The fact checking step requires the model to generate open-ended rationales, which is difficult to learn from SFT alone, hence we propose a two-stage training approach to better align the SLM with this task.

Reasoning Format Alignment Given a grounding document doc and a fact f, this training stage focuses on enabling the SLM to generate a rationale r_{sft} that adheres to the test-taking strategies in the prompt. To create a reasoning dataset, we prompt an LLM to alter f so that it violates certain criteria and use the teacher model to generate the reference rationale r_{ref}. In this way, we create $D_{Re\_SFT} = \{(x, r_{ref})\}$, where x is the prompt that contains doc and f.

We apply SFT with the following objective to align the SLM's rationales r_{sft} towards the reference rationales r_{ref}:

$$\mathcal{L}(\theta) = \mathbb{E}_{D_{Re\_SFT}} \left[\log P_\theta(r_{ref} \mid x)\right].$$

Mistake Revision Analogous to human learning, students improve accuracy by practising and correcting mistakes even after learning the key steps. Similarly, the SLM from the first stage may still fail to correctly reason about certain facts due to its smaller scale. To address this, we filter the training data to retain only examples where the SLM produces incorrect judgements while the LLM generates correct ones. This yields a preference dataset $D_{Re\_DPO} = \{(x, y_c, y_r)\}$, where the SLM's output serves as the rejected completion y_r and the LLM's output as the chosen completion y_c.

We then apply DPO (Rafailov et al., 2023) to align the SLM with the teacher model via the following objective:

$$\mathcal{L}(\theta) = - \mathbb{E}_{D_{Re\_DPO}} \log \sigma \left[\beta(s_\theta(x, y_c) - s_\theta(x, y_r))\right],$$

where σ is the sigmoid function, s is the log probability that the SLM assigns to the completion, and β is a hyperparameter. This learning process allows the SLM to iteratively revise its mistakes. Compared to SFT, DPO creates a margin to distinguish rejected and chosen completions, effectively suppressing the mistakes made by the SLM to improve factuality reasoning performance.

4 Experiments

In this section, we evaluate both LLM-based and SLM-based pipelines, including our proposed training method.

4.1 Benchmarks

We conduct evaluation on two factuality benchmarks, FacTax-Benchmark (Xu et al., 2024)[1] and LLM-AggreFact (Tang et al., 2024a). FacTax-Benchmark focuses on summarisation factuality checking on news and dialogue data, while LLM-AggreFact includes more diverse source types and claims generated by more recent models including LLMs. The breakdowns of the data types in the two benchmarks are listed in Appendix B. Despite the difference, these two benchmarks share overlapping source documents on news summarisation, while differing in the summarisers used to generate the claims. For FacTax-Benchmark, we exclude the datasets on government reports and stories due to the unavailability of the specific document versions used in the original study. As a result, evaluation on this benchmark is conducted only on news (Tang et al., 2023) and dialogue summarisation (Zhu et al., 2023). Benchmark statistics are listed in Table 8.

4.2 Baselines

We compare our method against a range of recent metrics, including TrueTeacher (Gekhman et al., 2023), MiniCheck (Tang et al., 2024a), FactCG (Lei et al., 2025), ChatGPT-ZS, ChatGPT-CoT (Luo et al., 2023), FacTax (Xu et al., 2024), and other LLM-based baselines reported in Tang et al. (2024a).[2] Where available, we cite results from the original papers. Otherwise, we run the released code following recommended settings. For metrics that output continuous scores between 0 and 1, we follow Tang et al. (2024a) and apply a fixed threshold of 0.5 to obtain binary True/False predictions. Baseline details are listed in Appendix A.

[1] To avoid confusion with the FacTax metric proposed in the same paper, we refer to it as the FacTax-Benchmark.

[2] https://llm-aggrefact.github.io/

4.3 Evaluation Metrics

Grounded Claim Factuality Checking We apply balanced accuracy (BAcc) to evaluate our pipelines, providing a fair comparison to prior works (Tang et al., 2024a; Xu et al., 2024).

$$\text{BAcc} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right),$$

where TP, TN, FP, FN denote the number of true positives, true negatives, false positives, and false negatives respectively.

Claim Decomposition We separately evaluate the outcome of training an SLM for the claim decomposition step using ROUGE-L (Lin, 2004) and sentence embedding-based cosine similarity (SECS).

$$\text{ROUGE-L} = \frac{LCS(F_{gen}, F_{ref})}{|F_{ref}|},$$

$$\text{SECS} = \frac{1}{|F_{gen}|}\sum_{F\in F_{gen}} \max_{F'\in F_{ref}} cossim(F, F'),$$

where F_{gen}, F_{ref} refer to the generated and reference facts. ROUGE-L measures the completeness of the generated facts relative to the reference, while SECS quantifies their semantic similarity.

4.4 Setup and Implementations

LLM-based Pipeline We apply our prompt to a series of Qwen3 models (QwenTeam, 2025) because of their outstanding instruction-following ability and long context window. To be more specific, we use Qwen3-4B-Instruct-2507[3] and Qwen3-30B-A3B-Instruct-2507[4] in the LLM-based evaluation pipeline. We also report results from the *Thinking* variant, although our prompt does not require long chain-of-thought reasoning.

SLM Fine-tuning Implementations For SLM-based pipeline, we use Qwen3-0.6B[5] as the base model. Different learning rates are adopted for SFT and DPO. During SFT, we use a large learning rate of 1×10^{-4} to adapt the SLM's output format. By the DPO stage, the model has learned natural language reasoning and criterion-checking abilities through SFT. Therefore, we apply a smaller learning rate of 1×10^{-7} to revise the chain-of-thought content while preserving the learned reasoning behaviour. All models are trained for 3 epochs with early stopping to mitigate overfitting. The distillation datasets D_{De} and D_{Re_SFT} are curated from the combined development splits from both benchmarks, and their statistics are reported in Tables 9 and 10. All prompts used in both inference and dataset construction are provided in Appendix C.

[3]https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[4]https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

[5]https://huggingface.co/Qwen/Qwen3-0.6B

4.5 Results

We first present the performance of all the methods on the two benchmarks and then assess the training outcome of SLMs.

4.5.1 Claim Checking Performance

LLM-based Pipeline Results in Tables 1 and 2 demonstrate that our prompt successfully adapts a general purpose LLM into an off-the-shelf factuality evaluation metric in a zero-shot setting, avoiding the need for training data. On FacTax-Benchmark, our Qwen3-30B-A3B-Instruct pipeline achieves state-of-the-art performance in both average score and overall ranking. On LLM-AggreFact, our method ranks second in Table 2 to a metric that requires post-training an LLM, while achieving the highest scores on four subsets. Consistent performance across benchmarks highlights the effectiveness and robustness of our prompt, without requiring any fine-tuning steps. Notably, our method outperforms methods based on closed-source models or larger open-source models such as GPT-4o and Llama-3.3-70B.

SLM-based Pipeline The bottom three rows in both tables illustrate the contribution of each training stage. The SLM consistently benefits from SFT and DPO, outperforming ChatGPT-3.5–based baselines on both benchmarks and achieving performance comparable to TrueTeacher, which relies on a substantially larger 11B base model and a 540B teacher model. Compared to metrics using small backbones (<1B parameters), our SLM-based pipeline surpasses all on FacTax-Benchmark, while showing a small gap to stronger baselines such as FactCG and MiniCheck-FT5 on LLM-AggreFact. This gap primarily occurs on datasets with long grounding documents, including LFQA and TOFUEVAL-MediaS, where extended context challenges generative models' reasoning and comprehension. Notably, our teacher model also exhibits degraded performance on these subsets, suggesting potential directions for future research.

	Method	Size	CNNDM				XSUM				Dialogues	Avg	
			Polytope	SummEval	Frank	CLIFF	Wang20	CLIFF	Goyal21	Cao22	DiaSummFact	BAcc (↑)	Ranking (↓)
Baseline	ChatGPT-ZS	/	90.2	79.8	54.8	65.1	71.8	74.0	63.4	68.8	66.9	70.5	9.7
	TrueTeacher	11B	100	64.9	56.1	66.6	76.6	78.7	73.6	73.3	67.4	73.0	6.5
	FactCG	0.4B	48.5	64.6	56.9	76.8	67.8	81.4	68.0	70.1	68.8	67.0	9.0
	MiniCheck-DeBERTa	0.4B	42.5	59.6	62.5	59.4	65.9	71.8	62.9	71.9	62.3	62.1	12.8
	MiniCheck-RoBERTa	0.4B	48.5	56.6	54.5	58.9	67.4	77.2	60.8	67.7	64.1	61.7	13.5
	MiniCheck-FT5	0.8B	46.3	73.7	60.6	71.1	65.9	77.3	70.0	71.0	65.8	66.9	9.6
	MiniCheck-BeSpoke-7B	7B	47.8	77.6	63.4	73.2	75.8	80.4	73.1	74.4	77.1	71.4	4.9
	FACTAX-ChatGPT-3.5	/	78.4	67.4	62.8	68.7	74.1	70.3	75.1	71.7	62.8	70.1	8.5
	FACTAX-ChatGPT-3.5-WD	/	85.3	73.0	67.1	70.7	71.5	70.8	68.9	69.5	64.2	71.2	8.3
	FACTAX-GPT-4o	/	94.1	79.5	61.7	79.7	74.1	70.2	73.9	70.1	74.6	75.3	5.4
Ours	Qwen3-4B-Thinking	4B	87.9	70.6	53.7	67.3	72.4	70.0	57.6	65.3	76.4	69.0	10.9
	Qwen3-4B-Instruct	4B	73.5	87.1	67.6	81.3	73.9	68.2	67.3	63.0	75.0	73.0	7.1
	Qwen3-30B-A3B-Thinking	30B	95.5	74.5	63.5	75.3	67.8	72.2	66.0	67.1	78.6	73.4	6.6
	Qwen3-30B-A3B-Instruct	30B	94.1	94.4	63.7	81.4	75.7	75.4	66.9	74.1	76.0	78.0	3.6
	Qwen3-0.6B	0.6B	41.2	68.8	52.4	53.7	60.5	60.7	54.6	57.3	60.1	56.6	16.5
	+SFT	0.6B	68.8	69.3	60.6	68.9	66.5	67.9	64.9	64.2	73.7	64.8	12.1
	+SFT+DPO	0.6B	73.7	88.3	61.3	72.6	69.7	76.0	67.0	68.9	76.2	72.6	7.2

Table 1: Results (BAcc) on FacTax-Benchmark test splits. The highest and second highest scores are coloured.

	Method	Size	AggreFact		TOFUEVAL		WICE	Reveal	Claim Verify	Fact Check	Expert QA	LFQA	RAGTruth	Avg	
			CNN	XSUM	MediaS	MeetB								BAcc (↑)	Ranking (↓)
Baseline	ChatGPT-ZS	/	63.2	72.4	66.8	73.4	68.5	84.7	65.2	70.8	57.2	73.8	75.6	70.1	13.8
	TrueTeacher	11B	60.4	74.2	70.9	73.6	64.2	91.1	64.4	76.8	59.5	90.2	80.9	73.3	8.4
	FactCG	0.4B	70.1	73.9	72.3	74.3	74.2	88.4	78.5	72.1	59.1	86.7	82.3	75.6	5.8
	MiniCheck-DeBERTa	0.4B	64.2	71.0	69.3	72.7	69.4	87.3	75.6	73.0	58.9	83.9	78.8	73.1	10
	MiniCheck-RoBERTa	0.4B	63.7	70.8	71.9	75.9	67.6	88.8	77.4	73.3	57.4	84.4	77.2	73.5	9.1
	MiniCheck-FT5	0.8B	69.9	74.3	73.6	77.3	72.2	86.2	74.6	74.7	59.0	85.2	78.0	75.0	6.8
	MiniCheck-BeSpoke-7B	7B	65.5	77.8	76.0	78.3	83.0	88.0	75.3	77.7	59.2	86.7	84.0	77.4	3.3
	Llama-3.3-70B-Instruct	70B	68.7	74.7	69.5	78.4	76.6	85.5	67.4	78.5	58.3	79.8	82.6	74.5	7.3
	GPT-4o-2024-05-13	/	68.1	76.8	71.4	79.8	78.5	86.5	69.0	77.5	59.6	83.6	84.3	75.9	4.8
Ours	Qwen3-4B-Thinking	4B	68.0	69.8	69.8	75.3	73.1	84.6	71.5	74.0	57.3	77.8	81.0	73.1	11
	Qwen3-4B-Instruct	4B	70.8	70.0	73.7	79.3	75.4	88.3	72.9	75.3	59.1	82.4	84.0	75.6	5.3
	Qwen3-30B-A3B-Thinking	30B	71.7	72.9	73.6	78.2	69.8	85.7	73.2	72.7	58.1	81.2	82.1	74.5	8
	Qwen3-30B-A3B-Instruct	30B	72.5	74.4	70.6	81.8	77.0	88.4	72.7	79.4	60.6	79.0	82.6	76.3	4
	Qwen3-0.6B-Instruct	0.6B	51.5	51.6	55.2	55.4	52.1	70.4	54.0	61.3	52.3	58.2	57.6	56.3	16
	+SFT	0.6B	65.1	71.3	66.2	72.4	73.0	86.4	66.2	73.6	58.7	75.7	75.9	71.3	12
	+SFT+DPO	0.6B	69.8	72.2	68.7	75.8	75.4	86.1	71.3	75.2	58.7	77.9	77.1	73.6	9.5

Table 2: Results (BAcc) on LLM-AggreFact test splits. The highest and second highest scores are coloured.

Impact of Backbone Size For metrics trained on the same data or using backbone models within a model family, performance consistently improves with increasing model size. For example, our prompt achieves higher scores with the 30B backbone than with the 4B backbone in both thinking and instruct modes, and similar trends are observed for MiniCheck. However, model size alone does not determine performance. Our 4B model outperforms Llama-3.3-70B-Instruct when the latter is used with a direct prompt on LLM-AggreFact, suggesting the importance of prompt design for zero-shot LLM-based factuality metrics.

4.5.2 SLM Claim Decomposition Quality

Table 3 shows that the SLM achieves substantial improvements after claim decomposition training. We further study the effect of the backbone model in this step. Table 4 compares the DPO checkpoint's balanced accuracy using facts decomposed by the 30B LLM versus the SFT-trained checkpoint. The small performance gap indicates that the trained SLM checkpoint can reliably decompose a claim into atomic facts, providing a strong foundation for training the subsequent atomic fact checker.

	FacTax-Benchmark		LLM-AggreFact	
	ROUGE-L (↑)	SECS (↑)	ROUGE-L (↑)	SECS (↑)
Before	63.3	71.5	54.0	58.7
After	89.3	95.5	84.3	92.6

Table 3: The SLM evaluation results before and after training for the claim decomposition task.

Claim Decomposition	FacTax-Benchmark	LLM-AggreFact
30B-A3B	73.4	73.7
0.6B	72.6	73.6

Table 4: Mean BAcc of the SLM pipeline with different claim decomposition models

5 Analysis

5.1 Impact of Output Length

We notice that switching the thinking mode on does not benefit the performance of our LLM-based pipeline. Moreover, Table 5 shows that instruct mode reduces token usage, by roughly 90%

on FacTax-Benchmark and over 80% on LLM-AggreFact. These results suggest that effective factuality judgements can be achieved through guided reasoning with explicitly defined steps, rather than unrestricted exploration. It demonstrates that our prompting strategy provides clear and practical guidance for LLMs, leading to stronger performance and lower inference cost.

Qwen3	Mode	FacTax		LLM-AggreFact	
		TokenUsed	Ratio (%)	TokenUsed	Ratio (%)
4B	Instruct	2803.6	10.4	2132.5	12.5
	Thinking	27001.9		17025.4	
30B-A3B	Instruct	1152.4	10.5	1024.4	17.7
	Thinking	10924.8		5781.7	

Table 5: The impact of switching thinking on and off. The ratio column presents the percentage of the tokens used in the instruct mode over the thinking mode.

5.2 Ablation Study

LLM-based Pipeline Table 6 reports the results of removing the claim decomposition step and the test-taking strategies from the prompt. When both components are removed we prompt the model with the claim and directly ask the model to judge factuality under instruct mode.

Removing either component leads to a noticeable drop in balanced accuracy on both benchmarks. This indicates that both claim decomposition and guided reasoning via test-taking strategies are critical to the effectiveness of our approach.

Qwen3	Pipeline	FacTax-Benchmark	LLM-AggreFact
4B	Full	73.0	75.6
	-Decomposition	72.3	74.6
	-Strategy	71.6	73.1
	-Both	69.4	72.1
30B-A3B	Full	78.0	76.3
	-Decomposition	76.2	76.0
	-Strategy	76.5	75.9
	-Both	74.8	74.7

Table 6: Ablation study for the LLM-based pipeline (mean BAcc).

SLM Training Tables 1 and 2 show the contribution of each training stage to SLM performance. We also investigate the choice of dataset used to create the distillation datasets D_{De} and D_{Re_SFT} and repeat SFT while omitting the training data derived from each of the benchmarks in turn. The dataset created from FacTax-Benchmark is noted as D_{FacTax}, and that from LLM-AggreFact as D_{LA}.

Table 7 shows that removing the training data corresponding to the benchmark substantially degrades performance on that benchmark. This suggests that the SLM's generalisation ability is more limited than that of the LLM, highlighting the importance of providing diverse training examples.

Dataset	FacTax-Benchmark	LLM-AggreFact
Full	67.2	71.3
$-D_{FacTax}$	64.2	66.7
$-D_{LA}$	65.3	62.1

Table 7: Ablation study on SLM training data origins (mean BAcc).

Rationale Examples A key advantage of our pipeline is its ability to generate explanations for factuality judgements. Unlike metrics that output a single scalar score, these explanations help users identify inconsistent parts of a claim. While some LLM-based metrics also produce reasoning, their outputs are often excessively long, thus less interpretable. Sample outputs in Table 15 show that the atomic facts generated by the LLM and SLM differ by only a single word, and both fact checking processes follow the prompt well, pinpointing exactly which part of the source text supports the fact.

6 Conclusion

In this work, we reframe grounded claim factuality checking as a reading comprehension task. Unlike prior approaches that directly query a model for claim support, we utilise human test-taking strategies in the prompt to guide reasoning. Experiments show that this prompt effectively adapted a general purpose LLM into an off-the-shelf factuality evaluation metric, avoiding the need for any training data. To reduce computational cost, we further distil the LLM into separate SLMs for claim decomposition and fact checking. Inspired by how students improve by correcting mistakes, the SLMs are additionally trained to revise their own errors to enhance reasoning accuracy. The resulting SLMs achieve performance comparable to much larger LLMs across benchmarks, demonstrating the effectiveness of our human-inspired training method. Overall, our study shows that, given suitable training strategies, SLMs can perform claim factuality checking efficiently while maintaining interpretability, though there remains potential for further improvement to address limitations brought by model scale, such as reasoning and generalisation ability.

Limitations

Base Model Ability Although our SLM-based pipeline outperforms several metrics that rely on LLM backbones, further performance improvements may require stronger base models. While this work demonstrates an effective approach for training SLMs for claim factuality checking, the resulting models remain constrained by their limited scale, particularly in terms of complex reasoning and deep language understanding.

Teacher Model Scale We use a 30B model as the teacher to generate reference outputs for distillation, which is not the strongest model available in the Qwen3 family. Since the quality of reference data plays a critical role in distillation, performance could potentially be improved by employing a stronger teacher model or by mixing outputs from multiple teacher models to enhance data diversity and quality.

Prompt Engineering LLMs are known to be sensitive to prompt design. A strength of our approach is the use of a zero-shot prompt for fact checking, which avoids reliance and sensitivity on few-shot examples. However, performance may still vary with different descriptions of the test-taking strategies in the prompt. While it is impractical to exhaustively evaluate all possible prompt wordings, performance could be boosted by further systematic prompt refinement, but this would bring a risk of over-fitting to specific models and datasets, which we have avoided in this paper.

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*.

Omaima Al-Kiyumi, Fawzia Al Seyabi, and Abdul Hamid Hassan. 2021. An empirical study on the effect of instruction on metacognitive strategies on efl reading comprehension: The case of foundation-level students in oman. *International Education Studies*, 14(8):30–42.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2023. Understanding retrieval augmentation for long-form question answering. *arXiv preprint arXiv:2310.12150*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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.

Tao Feng, Yicheng Li, Chenglin Li, Hao Chen, Fei Yu, and Yin Zhang. 2024. Teaching small language models reasoning through counterfactual distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842, Miami, Florida, USA. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4615–4634, Bangkok, Thailand. Association for Computational Linguistics.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3134–3154, Singapore. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Deren Lei, Yaxi Li, Siyao Li, Mengya Hu, Rui Xu, Ken Archer, Mingyu Wang, Emily Ching, and Alex Deng. 2025. FactCG: Enhancing fact checkers with graph-based multi-hop data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5002–5020, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

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

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

QwenTeam. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Xu, Zhenlin Su, Mo Yu, Jin Xu, Jinho D. Choi, Jie Zhou, and Fei Liu. 2024. Identifying factual inconsistencies in summaries: Grounding LLM inference via task taxonomy. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14626–14641, Miami, Florida, USA. Association for Computational Linguistics.

Deborah Yapp, Rick de Graaff, and Huub van den Bergh. 2023. Effects of reading strategy instruction in english as a second language on students' academic reading comprehension. *Language Teaching Research*, 27(6):1456–1479.

Yuxuan Ye, Edwin Simpson, and Raul Santos Rodriguez. 2024. Using similarity to evaluate factual consistency in summaries. *arXiv preprint arXiv:2409.15090*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6825–6845, Toronto, Canada. Association for Computational Linguistics.

A Baseline Details

We introduce the details of all the mentioned baselines in this paper below.

TrueTeacher distils factuality judgement ability from a directly prompted 540B LLM into a relatively smaller model with 11B parameters. While effective, both the teacher and student models are substantially larger than our SLM.

MiniCheck synthesises training data conditioned on either the grounding document or the claim, which leads to substantially improved entailment classifiers. It is a strong baseline that achieves performance comparable to much larger LLMs.

FactCG is specifically trained to improve reasoning over claims that require multi-hop inference. It represents the current state of the art among entailment-based metrics and is built on BERT-style architectures such as DeBERTa (He et al., 2020).

ChatGPT-{ZS,CoT} directly prompt ChatGPT-3.5-turbo to judge whether a summary is consistent with the source document. The chain-of-thought (CoT) variant additionally encourages step-by-step reasoning, resulting in an open-ended reasoning process. Other LLM-based baselines, if mentioned directly with their names in this paper, follow a similar method but differ in base models.

FACTAX defines fine-grained error categories in the prompt and uses LLMs' reasoning capabilities to decide if the claim contains certain error types.

B Dataset Statistics

We present the statistics of the datasets involved in this paper.

For the dataset type, FacTax has

- **News Summarisation** PolyTope (Huang et al., 2020), SummEval (Fabbri et al., 2021), FRANK (Pagnoni et al., 2021), CLIFF (Cao and Wang, 2021), Wang20 (Wang et al., 2020), Cao22 (Cao et al., 2022), Goyal21 (Goyal and Durrett, 2021). These datasets are collectively known as AggreFact-SOTA

- **Dialogue Summarisation** DiaSumFact (Zhu et al., 2023)

LLM-AggreFact has

- **Summarisation** AggreFact-CNN/XSum (Tang et al., 2023), TofuEval-MeetB/MediaS (Tang et al., 2024b), RAGTruth (Niu et al., 2024)

- **Retrieval-augmented Generation** ClaimVerify (Liu et al., 2023a), LFQA (Chen et al., 2023), ExpertQA (Malaviya et al., 2024), RAGTruth

- **Post-hoc Grounding** ExpertQA, REVEAL (Jacovi et al., 2024), FactCheck-GPT (Wang et al., 2024)

- **Human Written Claims** WiCE (Kamoi et al., 2023)

Benchmark	Split	Size	Source Length	Claim Length	Consistent Ratio (%)
FacTax-Benchmark	dev	1236	442.4	37.3	64.2
	test	1592	395.3	43.5	49.4
LLM-AggreFact	dev	29320	580.8	22.2	75.7
	test	30420	559.4	22.5	77.9

Table 8: The statistics of the two benchmarks.

Split	Size	Claim Length	Avg Facts per Claim
training	69948	23.4	3.0
test	7250	23.7	3.0

Table 9: The statistics of the curated dataset for claim decomposition learning.

Split	Size	Reason Length	Consistent Ratio (%)
training	32461	897.9	33.7
test	3608	1062.8	25.4

Table 10: The statistics of the curated dataset for fact checking learning.

Follwing the example below, segment the given claim into atomic facts only based on the claim itself. Output each fact with "-" as the start.

Claim:
The parkway was opened in 2001 after just under a year of construction and almost two decades of community requests.
Facts:
- The parkway was opened in 2001.
- The parkway was opened after just under a year of construction.
- The parkway was opened after two decades of community requests.

CLAIM:
<claim>
Facts:

Table 11: The prompt for claim decomposition.

Read the article given below and answer the questions.

ARTICLE:
<article>

Here is a claim, answer the following questions. Please reason step by step, and output your final answer by "Final Answer: yes" or "Final Answer: no".

CLAIM:
<claim>

1) For the claim, are the object and the subject mentioned?

2) If the object and the subject are mentioned, is their related information verifiable according to the article? If there is information not mentioned, carry it into the next question. If verifiable but incorrect, stop here and answer "Final Answer: no".

3) Look at the relationships between the object and the subject, is their relationship mentioned? If not, can the relationship be inferred from the article? If the relationship stands, can the previous information not mentioned be inferred from the article?

Table 12: The prompt for fact checking. Only in ablation study, <*claim*> refers to a original claim. Otherwise, it is an atomic fact decomposed from the claim.

Read the article given below and answer the questions.

ARTICLE:
<article>

Read the claim given below, judge if the claim is supported by the article or not. Please reason step by step, and output your final answer by "Final Answer: yes" or "Final Answer: no".

CLAIM:
<claim>

Table 13: The direct prompt used for ablation study.

You are a proficient writer that always help people to change text meaning.

I will give you a piece of text. Please [write a new similar sentence with different subjects or objects/tweak or add details in the text/keep the subject and object the same, but change the predicates in the text or negate the verb]. Only output one changed text without any explanation.

Text:
<TEXT>

Rewritten text:

Table 14: The prompt used for alter the claims. Different prompts are applied to make the claim violate different criteria.

C Prompt

We list all the prompts involved in the experiments in this section.

D Qualitative Study Example

Prompt

Read the article given below and answer the questions.

ARTICLE:

The adapted artwork for the Sgt Pepper's Lonely Hearts Club Band album is worth about £70,000... The artwork is one of five Beatles album sleeves to feature in the top 10, compiled by Record Collector magazine. Sir Peter Blake's original Sgt Pepper collage was doctored to replace the faces of The Beatles and other notable figures with those of executives from the band's US label Capitol... In second place, valued at £7,000, are the...
Here is a claim, answer the following questions. Please reason step by step, and output your final answer by "Final Answer: yes" or "Final Answer: no".

CLAIM:

A beatles album sleeve designed by sir peter blake has topped a list of the most valuable rare records.

1) For the claim, are the object and the subject mentioned?

2) If the object and the subject are mentioned, is their related information verifiable according to the article? If there is information not mentioned, carry it into the next question. If verifiable but incorrect, stop here and answer "Final Answer: no".

3) Look at the relationships between the object and the subject, is their relationship mentioned? If not, can the relationship be inferred from the article? If the relationship stands, can the previous information not mentioned be inferred from the article?

Facts from the LLM
- A Beatles album sleeve was designed by Sir Peter Blake.
- The Beatles album sleeve topped a list of the most valuable rare records.

Facts from the trained SLM
- A Beatles album sleeve was designed by Sir Peter Blake.
- The Beatles album sleeve has topped a list of the most valuable rare records.

Checking Fact 1 by LLM

1) Yes, both the object and the subject are mentioned.
- The subject is ""Sir Peter Blake"".
- The object is ""a Beatles album sleeve"".

2) Yes, the related information is verifiable according to the article.
- The article states: ""Sir Peter Blake's original Sgt Pepper collage was doctored...""
- This confirms that Sir Peter Blake created the original artwork for the Sgt Pepper's Lonely Hearts Club Band album, which is a Beatles album sleeve.
- Therefore, the connection between Sir Peter Blake and a Beatles album sleeve is supported.

3) Yes, the relationship between the subject and the object is directly mentioned in the article.
- The article explicitly states that Sir Peter Blake created the original artwork for the Sgt Pepper's Lonely Hearts Club Band album.
- This confirms the relationship: Sir Peter Blake designed the album sleeve.
- No additional information needs to be inferred, as it is clearly stated.

Final Answer: yes

Checking Fact 1 by the trained SLM

1) Yes, both the object (a Beatles album sleeve) and the subject (Sir Peter Blake) are mentioned in the article.

2) Yes, the information is verifiable according to the article. The article states: ""Sir Peter Blake's original Sgt Pepper collage was doctored to replace the faces of The Beatles and other notable figures with those of executives from the band's US label Capitol."" This confirms that a Beatles album sleeve was designed by Sir Peter Blake.

3) The relationship between the object (a Beatles album sleeve) and the subject (Sir Peter Blake) is explicitly mentioned in the article — Sir Peter Blake designed the original Sgt Pepper collage, which was a Beatles album sleeve. This relationship is clearly stated and supported.

Final Answer: yes

Table 15: An example from the benchmark. The pinpointed information in the claim and the source is highlighted in blue. The evidence sentence located is underlined.