FactGuard: Veridicity of Claims

Rick Pereira[1] and Karan Patel[2]

[1]School of Information - University of California, Berkeley

***Abstract*—Automated claim verification with large language models is effective but computationally expensive. FactGuard addresses this by distilling knowledge from the Gemini 2.5 Flash teacher model into two smaller student models: T5-Gemma (encoder-decoder) and Gemma-2B (decoder-only). Using FEVER and SQuAD to generate high-quality distillation datasets, both models were fine-tuned and evaluated on FEVER, BoolQ, and LIAR datasets. Results show substantial gains on structured, evidence-driven datasets like FEVER and context-dependent datasets like BoolQ, with RAG integration further improving performance. While LIAR remains challenging, the findings highlight that lightweight, knowledge-distilled models can provide scalable, efficient, and reasonably accurate solutions for real-world automated fact-checking.**

I. Introduction

Automated claim verification has become increasingly important as digital information continues to scale beyond the capacity of manual fact-checking. While modern large language models demonstrate strong performance on verification tasks, their computational cost makes them impractical for large-scale deployment. In production environments, a fact checker model must balance accuracy with efficiency - yet state-of-the-art models typically prioritize performance at the expense of speed and resource consumption. FactGuard addresses this challenge by exploring whether knowledge distillation can transfer the verification capabilities of a high-performing teacher model into substantially smaller architectures. In this project, Gemini 2.5 Flash serves as the teacher model, while two lightweight student models - Gemma-2B (a decoder-only model) and T5 Gemma (an encoder–decoder model) act as the distilled recipients. The core objective is to determine whether these compact models can approximate the teacher's decision-making while offering lower inference cost and improved deployability.

To evaluate this, the project uses a teacher model generated distillation dataset derived from two well-established corpora: FEVER, which provides fact-checked claims, and SQuAD, which supplies question-answer pairs that can be converted into true and false claims. The distilled student models are then tested on multiple independent evaluation datasets, including FEVER, BoolQ, and LIAR, to assess generalization beyond the teacher-generated data. Our results show that both student models achieve strong classification performance given their substantially smaller size. These findings indicate that knowledge-distilled verification models can provide a practical and scalable foundation for real-world automated fact-checking systems.

II. Background

The FactGuard project is designed to address the significant challenge of computational expense in using Large Language Models for large-scale automated claim verification. While state-of-the-art models like Gemini 2.5 Flash offer peak performance, their high operational costs prevent their scalable deployment as a primary autorater [1]. FactGuard's goal is to deliver a lightweight, cost-efficient, and production-ready solution by transferring the robust verification knowledge from the powerful teacher model (Gemini 2.5 Flash) to smaller, more resource-efficient student architectures like T5Gemma and Gemma 2. This is achieved through the methodology of knowledge distillation combined with Supervised Fine-Tuning, aiming to create a highly performant yet significantly more economical autorater for production environments [1], [3].

Historically, claim verification has evolved from multi-stage pipelines—including claim detection, evidence retrieval, and verdict prediction—to evidence-based justification, notably standardized by datasets like FEVER [1], [2], [7]. This focus shifted the task from analyzing mere linguistic cues, such as emotional or hyperbolic language in false claims [5], or network analysis of user credibility [6], towards verification based on explicit evidence. Modern approaches also include graph-based methods, checking if structured relationships in a knowledge graph support a claim [4]. These foundational works established effective verification methodologies but did not resolve the ensuing resource bottleneck caused by the reliance on massive transformer architectures for maximizing verification performance.

FactGuard's primary contribution lies not in modeling the verification task itself, but in the deployment and efficiency space. It differentiates itself from prior work by directly addressing the scalability challenge inherent in deploying high-performing LLMs. The project uses knowledge distillation, a proven technique in Natural Language Processing (NLP) for model compression, to transfer the complex decision-making capabilities of the costly teacher model to significantly smaller student models [4]. This methodology is key to maintaining high performance while dramatically reducing inference latency and operational costs, ultimately making state-of-the-art claim verification technology viable and sustainable for large-scale production environments.

III. Methodology

A. Data

FactGuard utilizes a Teacher-Student model paradigm for its distillation process. This process is designed to transfer complex claim verification expertise from a large, high-performing model to a smaller, resource efficient model. Gemini 2.5 Flash was utilized as the corresponding teacher model and consumes raw claims from the FEVER or SQUAD corpus. Critically, the

TABLE I: Hyperparameter Tuning Results

Model	Parameter	Search Values	Best Value	Final Training Loss	Final Validation Loss
T5Gemma	Epoch	[1, 2, 3]	1	0.0358	0.038327
	Batch Size	[8, 32]	8		
	Learning Rate	[5e-04, 5e-05]	5e-05		
	Lora Rank	[8, 16, 64]	16		
	Lora Alpha	[16, 128]	16		
	Drop Out	[0.01, 0.05]	0.05		
Gemma 2	Epoch	[1, 2, 3]	3	0.16	0.36
	Batch Size	[2, 4]	4		
	Learning Rate	[5e-04, 5e-05]	5e-04		
	Lora Rank	[8, 16, 32, 64]	64		
	Lora Alpha	[16, 32, 64, 128]	32		
	Drop Out	[0.01, 0.05]	0.05		

teacher is leveraged to generate a unique, high-value distillation dataset by outputting not only the final label (true or false), but also a rationale for that verdict.

FEVER (Fact Extraction and VERification) was selected as part of the distillation process as it contains a corpora of high-quality claims explicitly labeled as either supported or refuted by evidence. This dataset is split into training and test sets, of which the training set is utilized during the distillation process. For each entry, the teacher model processes the input, which is composed of the claim itself and the corresponding binary verdict (true or false). The model is then instructed to generate two critical output components: the specific textual context, the evidence required to support or refute the claim based on the assigned verdict, and the explicit rationale, the step-by-step reasoning that explains why the given verdict is correct based on the provided evidence.

Additionally, the Stanford Question Answering Dataset (SQuAD) was selected for the distillation process for its questions rooted in wikipedia contexts and the corresponding best answers. For each SQuAD entry, the teacher model receives the input components: a question, its relevant context passage, and the correct answer span. Based on this information, the teacher model's key task is to generate two distinct claims: one that is true and directly supported by the context, and one that is false and directly contradictory to the context. For each of the generated claims, the teacher explicitly assigns the binary verdict (true or false) and provides a detailed rationale explaining the exact reason why the claim holds true or false relative to the provided context.

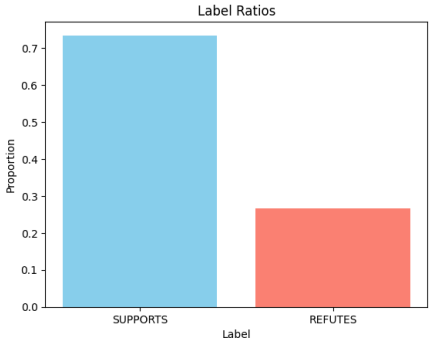
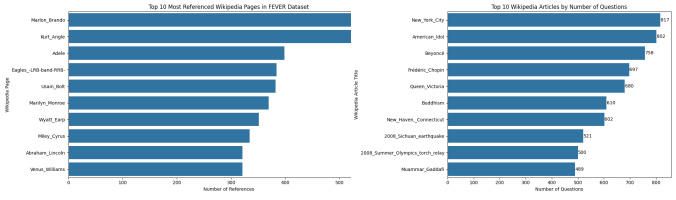


Fig. 1: FEVER: Label Ratio (SUPPORTS vs REFUTES)

In the FEVER-derived claim set, the distribution of 'supports' versus 'refutes' labels is highly imbalanced, with a much larger proportion of claims supported by evidence. This imbalance suggests that models may develop a bias toward predicting support unless additional balancing or regularization is applied. Additionally, examining the most frequently referenced Wikipedia pages shows that claims concentrate heavily on well-known public figures: actors, athletes, musicians, and other historical personalities. This indicates that FEVER has a strong topical skew toward biography-style facts, which may simplify retrieval or verification relative to domains with more abstract concepts.



(a) Most Reference Wiki Page (b) Most Reference Wiki Article

Fig. 2: FEVER: Most Reference Wiki (Left: Wiki Page, Right: Wiki Article).

The top SQuAD articles associated with the generated questions reveal a more mixed topical distribution. The High-frequency pages span geography, religion, notable historical figures, and major events, suggesting a wider domain coverage than FEVER.

The claim length distributions across FEVER, BoolQ, and LIAR reveal important structural differences between the datasets. FEVER contains short and concise claims, most of which are between 5 and 10 words. BoolQ claims are similarly short, but exhibit slightly higher average length and more variation in phrasing because they originate from naturally occurring user questions. In contrast, LIAR claims are substantially longer, typically ranging from 10–30 words, and cover diverse political statements taken from speeches, interviews, and news sources. The increasing claim length and linguistic complexity across these datasets suggest rising levels of reasoning difficulty, which becomes relevant when interpreting model performance later in the paper.

B. Modeling

T5Gemma and Gemma-2B were both selected as candidate models for FactGuard due to their distinct architectures, which offer different trade-offs for the task of automated fact-checking. T5Gemma was chosen for its encoder-decoder architecture. The dedicated encoder is expected to be highly effective at creating dense, information-rich contextual representations of the input claims and evidence, which facilitates superior knowledge transfer during the distillation process. The decoder then uses this rich representation to excel in structured generation, providing evidence-based justifications and a definitive veracity verdict. In contrast, Gemma-2B was chosen for its lightweight, decoder-only architecture and efficient autoregressive design. This structure processes input and generates output within a single, unified stack, making it a compact and latency-efficient student model. While it lacks a dedicated encoder, Gemma-2B uses highly optimized self-attention layers to internalize the teacher model's decision patterns with minimal parameter overhead, offering a strong balance between model size and representational capacity for efficient distillation.

To accurately quantify the performance gains of both T5Gemma and Gemma-2B within FactGuard, three distinct evaluation configurations were used. First, a baseline was established by evaluating the model before fine-tuning to measure its raw, pre-distillation performance, setting a reference point for the efficacy of the knowledge transfer. Second, the fine-tuned model was evaluated without providing external evidence, isolating the model's ability to assess claim veracity based purely on the compressed parametric knowledge inherited from the teacher model. Finally, the model was tested within a Retrieval-Augmented Generation approach, which consists of the fine-tuned model and evidence retrieved via a DuckDuckGo web query. This real-world pipeline evaluation assesses the model's capacity to effectively synthesize its fine-tuned reasoning with external, up-to-date evidence, which is vital for maximizing accuracy and ensuring the final veracity verdict and justification rationale are grounded in verifiable, real-time information.

IV. Results

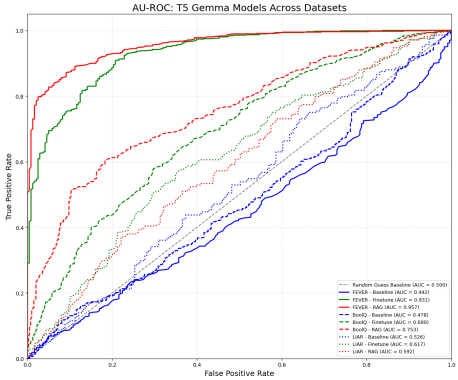


Fig. 3: AU-ROC: T5Gemma models across FEVER, BoolQ, and LIAR datasets

The FEVER dataset saw the most significant and consistent performance gains (a maximum of 53.3% Accuracy percentage point increase, 63.1% F1 increase) for T5-Gemma when comparing the fine-tuned models to the baseline. The Baseline model performed poorly, with an AU-PRC of 0.72, an Accuracy of 31.70%, and a very low F1 score of 25.84%, despite a respectable Specificity (No) of 80.00%. The dummy model's majority-class accuracy of 70% on FEVER is a critical benchmark; any model's performance must exceed this 0.70 baseline to demonstrate meaningful learning. In contrast, both the fine-tuned LLM and RAG evaluations demonstrated a dramatic lift in all key metrics, significantly surpassing the dummy model baseline. The T5Gemma Fine Tuned LLM (AU-PRC 72.10, F1 77.84%) and T5Gemma RAG (AU-PRC 85, F1 89%) showed a much better balance between Recall (Yes) and Specificity (No), indicating the fine-tuning process was highly effective at improving the model's ability to correctly classify both supported and refuted claims in the FEVER dataset.

Interestingly, it was observed during the tuning process that T5Gemma began to overfit with more than 2 epochs. Specifically, the validation loss plateaued and began to increase while the training loss continued to decrease, indicating the model was memorizing the training data. This suggests that for T5Gemma, 1-2 epochs represents the optimal training duration to maximize generalization performance and prevent overfitting.

For the BoolQ dataset, demonstrated an improved AU-PRC of 0.76 and a significant improvement in Specificity (No) to 97%. The RAG evaluation on BoolQ further improved performance with an Accuracy of 66.80% and an AU-PRC of 0.83.

The LIAR dataset proved to be the most challenging for T5Gemma, with the lowest overall metric scores compared to the other two datasets. While the fine-tuned LLM and RAG evaluations successfully introduced better balance, the overall performance remains marginal. The T5Gemma Fine Tuned LLM achieved an accuracy of 58% and an F1 of 39%. The T5Gemma RAG evaluation achieved an accuracy of 55% and an F1 of 56.10%.

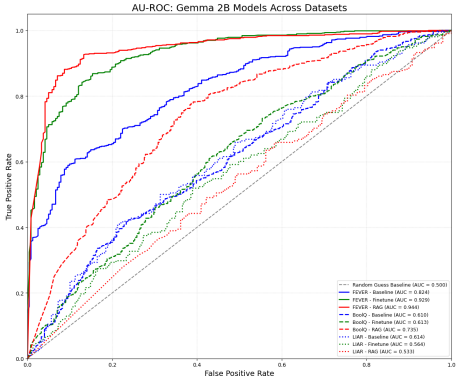


Fig. 4: AU-ROC: Gemma 2 models across FEVER, BoolQ, and LIAR datasets

Across the three evaluation datasets, Gemma 2 demonstrates meaningful and often substantial improvements after supervised fine-tuning, with the strongest gains appearing in the FEVER and BoolQ evaluations.

TABLE II: Model Evaluation Results

Model	Evaluation Type	AU-PRC	Recall (Yes)	Specificity (No)	Accuracy	Precision	F1
T5Gemma	FEVER - Baseline	0.72	16.00%	80.00%	31.70%	73.46%	25.84%
	BoolQ - Baseline	0.60	100.00%	0.00%	60.10%	60.10%	75.08%
	LIAR - Baseline	0.48	100.00%	0.00%	45.51%	45.51%	62.56%
	FEVER - Fine Tuned LLM	0.98	65.00%	96.00%	72.10%	98.00%	77.84%
	BoolQ - Fine Tuned LLM	0.76	12.98%	97.00%	46.00%	86.00%	23.00%
	LIAR - Fine Tuned LLM	0.56	30.00%	81.00%	58.00%	57.00%	39.00%
	FEVER - RAG	0.99	82.00%	95.00%	85.00%	98.00%	89.00%
	BoolQ - RAG	0.83	56.00%	84.00%	66.80%	83.71%	66.80%
	LIAR - RAG	0.53	63.00%	49.00%	55.00%	51.00%	56.10%
Gemma 2	FEVER - Baseline	0.94	60.87%	87.97%	67.40%	94.09%	73.92%
	BoolQ - Baseline	0.69	35.11%	80.70%	53.30%	73.26%	47.47%
	LIAR - Baseline	0.56	14.90%	92.77%	57.33%	63.27%	24.12%
	FEVER - Fine Tuned LLM	0.97	70.22%	95.44%	76.30%	97.98%	81.81%
	BoolQ - Fine Tuned LLM	0.69	47.25%	67.67%	55.40%	68.77%	56.02%
	LIAR - Fine Tuned LLM	0.52	44.71%	65.06%	55.80%	51.67%	47.94%
	FEVER - RAG	0.98	82.21%	93.78%	85.00%	97.65%	89.27%
	BoolQ - RAG	0.78	75.04%	63.16%	70.30%	75.42%	75.23%
	LIAR - RAG	0.48	49.04%	56.22%	52.95%	48.34%	48.69%

On the FEVER dataset, the Gemma 2 Baseline already performs reasonably well (Accuracy 67.40%, AU-PRC 0.94), but the fine-tuned versions show clear advancement, all surpassing the 70% dummy model baseline. The Gemma 2 Fine Tuned LLM reaches 76.30% accuracy with a large jump in F1 (81.81%), and the Gemma 2 RAG-augmented evaluation further increases performance to 85.00% accuracy and an F1 of 89.27%. The RAG integration model also displays a sizable boost in recall (82.21%) while maintaining strong specificity (93.78%).

On BoolQ, the Gemma 2 Baseline exhibits a noticeable imbalance, with high specificity (80.70%) but weaker recall (35.11%), resulting in only moderate performance overall (53.30% accuracy). The Gemma 2 Fine Tuned LLM shifts the model away from this imbalance, with accuracy at 55.40% and an AU-PRC of 0.69. In contrast, the Gemma 2 RAG variant produces the most substantial improvement, lifting accuracy to 70.30% and dramatically increasing recall (75.04%) while preserving reasonable specificity (63.16%).

The LIAR dataset continues to be the most difficult benchmark. Although the Gemma-2B baseline shows strong specificity (92.77%), it struggles to correctly identify "Yes" examples, resulting in a very low recall of 14.90% and an F1 of 24.12%. Both the Gemma 2 Fine Tuned LLM and Gemma 2 RAG variant evaluations settle around similar levels of performance (with accuracies of 55.80% and 52.95% respectively) and F1 scores just below 49%. The reduced AU-PRC in the RAG setting (0.48) further underscores that retrieval does not necessarily aid this dataset's diverse and factually complex claims.

V. Discussion

Our findings establish that all evaluated models - T5Gemma and Gemma 2 - successfully learned the fundamental task of binary claim verification. The inherent class imbalance in the primary evaluation dataset creates a naive baseline that can be easily achieved without any genuine predictive power. Crucially, every single model configuration significantly surpassed this naive threshold, validating their capability to distinguish between supporting and refuting claims. The dramatic performance gains following Supervised Fine-Tuning provide compelling evidence for the efficacy of knowledge distillation, proving that complex claim verification expertise can be successfully transferred to smaller language models. Furthermore, the integration of a Retrieval-Augmented Generation pipeline consistently maximized performance on evidence-driven tasks, confirming the robustness of an architecture that synthesizes sophisticated, distilled reasoning with external, verifiable evidence.

The results, however, highlight significant limitations in generalization and underscore the importance of dataset specificity. While the SFT approach proved highly effective on the core claim verification task, its utility varied across others. On the BoolQ question-answering task, SFT induced an extreme prediction bias that led to poor overall performance, indicating that the learned claim verification heuristics did not generalize successfully to this binary question format. Intriguingly, subsequent evaluation under the RAG configuration completely remediated this bias and achieved peak performance for T5Gemma. This demonstrates that for some tasks, the retrieval of high-quality external context is paramount for successfully resolving the task, effectively overriding poorly generalized internal heuristics learned during fine-tuning. The most persistent challenge was observed on the LIAR multi-class fact-checking dataset. Across all model configurations, performance remained marginal, emphasizing that the inherent semantic nuance and diverse, subtle distinctions within real-world misinformation represent a significant hurdle for the current model architectures.

The claim length analysis presented in the EDA section helps contextualize the performance differences observed across the three test datasets. Both T5-Gemma and Gemma-2B exhibit

the strongest improvements on FEVER, FEVER claims are short, focused, and relatively uniform, enabling the student model to learn consistent verification patterns. BoolQ claims are similarly short but more linguistically varied and tied to nuanced question-style phrasing, making veracity harder to infer without external context. LIAR claims are much longer, denser, and span diverse political topics, increasing reasoning complexity. This greater variability and length likely contribute to the weaker performance on BoolQ and especially LIAR during testing.

The evaluation also provided insight into architectural efficiency and a notable technical anomaly. The results demonstrate that the lightweight, decoder-only Gemma 2 model benefits substantially from knowledge distillation, effectively approximating the fact-checking capabilities of larger LLMs on structured tasks. This confirms that model size and architecture are less constraining than the quality of the fine-tuning signal when evidence is well-defined. Conversely, the model's performance on LIAR suggests that architecture and size do impose limits when faced with highly ambiguous or multi-topic claims, hindering the ability to generalize comprehensively. Finally, a critical technical observation was a strong prediction bias in the raw output of the T5Gemma model towards the "No" verdict, which was inconsistent with its actual generated text. To accurately reflect the model's learned competence and align quantitative metrics with the correct conceptual output, an adjustment to the decision boundary was required.

The observed performance patterns align significantly with documented challenges in both large-scale machine learning deployment and specialized fields like Automated Fact-Checking (AFC). A crucial point of alignment is the recurring trade-off between model complexity and operational efficiency. If the current model exhibits high accuracy but suffers from substantial inference latency or high memory utilization, this confirms the "bottleneck of deploying large-scale neural networks on resource-limited hardware devices" as detailed in the survey on Knowledge Distillation (KD) [3]. Furthermore, if the model struggles with data outside its training distribution, this reflects the known issue of poor generalization resulting from the "lack of high-quality, large-scale, and diverse annotated datasets" identified within AFC research [2]. Addressing these limitations is essential for creating a robust and production-ready system.

Future Work

Future development of the FactGuard system should prioritize scaling the underlying language model and exploring alternative architectures to boost performance, especially on complex multi-class verification tasks. An immediate step is to test larger models, such as Gemma-7B, to determine if increased parameter capacity improves the subtlety and accuracy of distilled verification knowledge. Beyond scaling, investigating different model architectures, particularly transitioning from the current T5Gemma encoder-decoder to pure decoder models (like the base Gemma family), could reveal new performance ceilings in generative consistency and reasoning. Additional gains may come from structural changes, such as moving from the current binary classification to a Multi-label Output system (e.g., True / False / Unverified). The existing Gemma-2B results already confirm that efficient student models are viable for structured fact-checking tasks, and future work will focus on generalizing this success.

The second key area for improvement involves enhancing the data retrieval and reasoning pipeline. This includes refining the Retrieval-Augmented Generation approach by replacing the current search engine with a more granular, high-precision alternative and augmenting the structure with a specialized, fine-tuned LLM dedicated to generating high-quality summaries from the retrieved evidence. For advanced optimization, future work should explore Direct Preference Optimization after Supervised Fine-Tuning on datasets like TruthfulQA to better align the model's outputs with human-preferred factual accuracy and handle nuanced cases like misconceptions. Practical results indicate that retrieval mechanisms offer measurable benefits for context-dependent claims but show limited gains on highly diverse or complex datasets. Therefore, future efforts will also focus on broader fine-tuning, architectural enhancements for better contextual reasoning, and more sophisticated retrieval strategies, along with expanding evaluation across different domains to solidify the model's generalization potential.

Conclusion

FactGuard demonstrates that knowledge distillation from a high-performing teacher model into smaller student models can produce efficient and reasonably accurate solutions for automated fact-checking. Both T5-Gemma and Gemma-2B show substantial improvements after fine-tuning, particularly on datasets like FEVER, which provides explicit supporting evidence, and BoolQ, which requires reasoning over related passages. While performance on more complex, real-world claims such as those in LIAR remains limited, these results suggest that further fine-tuning on diverse datasets and architectural enhancements could improve accuracy. Overall, lightweight, knowledge-distilled models offer a practical path toward scalable, deployable fact verification systems.

References

[1] Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*, 2024.

[2] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, feb 2022.

[3] Peter Jano. Knowledge distillation in neural networks: A comprehensive survey. Technical report, Technical Report, 2025. ResearchGate Publication 394846514.

[4] Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[5] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2931–2937, Copenhagen, Denmark, september 2017.

[6] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the ACM conference on Computer Supported Cooperative Work and Social Computing*, 2016.

[7] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

Author Contributions

Rick Pereira: Dataset collection, distillation effort, encoder-decoder model fine-tuning, hyperparameter tuning, and evaluation, AU-ROC; Written - Background, Data, Encoder-Decoder Model, Encoder-Decoder Results (including AU-ROC graphs), Encoder-Decoder Discussion, Future Work; Project Administration - Proof-of-Concept example, most of written proposal, most of milestone, porting to latex, supervision of direction and management of project (setting up syncs, communicating progress, ensuring computational resources are available, combining colab notebooks and efforts, etc.)
Karan Patel: Exploratory Data Analysis, baseline Analysis, decoder-Only model fine-tuning, hyperparameter tuning, and evaluation; Written - Abstract, Introduction, Exploratory Data Analysis in Data, Decoder-Only Model, Decoder-Only Results, Decoder-Only Discussion, Conclusion;

All authors have read and agreed to the current version of this document.

Appendices

Appendix A
Training and Validation Loss

TABLE III: T5Gemma Training and Validation Loss

Epoch	Training Loss	Validation Loss
1	0.69	0.68
2	0.68	0.68
3	0.57	0.71

TABLE IV: Gemma 2 Training and Validation Loss

Epoch	Training Loss	Validation Loss
1	0.92	0.87
2	0.41	0.52
3	0.16	0.36

Appendix B
Distribution of Claim Lengths Per Eval Dataset

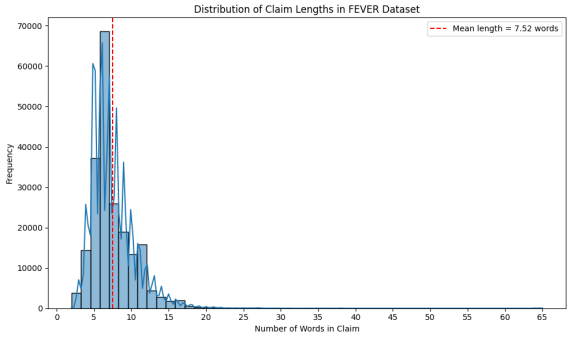


Fig. 5: FEVER: Distribution of Claim Word Length

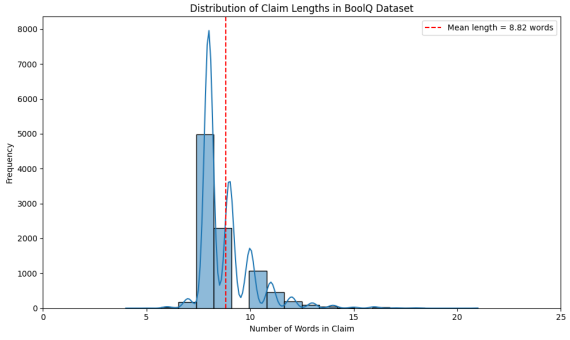


Fig. 6: BOOLQ: Distribution of Claim Word Length

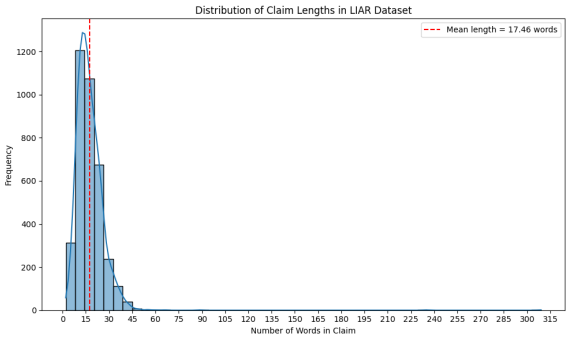


Fig. 7: LIAR: Distribution of Claim Word Length

APPENDIX C
SFT - LLM PROMPT EXAMPLE

```
**Fact-Check and Evidence Verification**
Determine the final verdict:
* **Yes:** If the claim is fully supported by
    the Context (if provided) or by external
    knowledge.
* **No:** If the claim is false, contradicted,
     or if there is insufficient evidence to
    support or deny the claim.
Output Requirement: Output the final verdict
    ('Yes' or 'No') and nothing else.
--- Context ---
Many species of the second major avialan
    lineage to diversify, the Euornithes (
    meaning "true birds", because they include
     the ancestors of modern birds), were semi
    -aquatic and specialized in eating fish
    and other small aquatic organisms. Unlike
    the enantiornithes, which dominated land-
    based and arboreal habitats, most early
    euornithes lacked perching adaptations and
     seem to have included shorebird-like
    species, waders, and swimming and diving
    species. The later included the
    superficially gull-like Ichthyornis, the
    Hesperornithiformes, which became so well
    adapted to hunting fish in marine
    environments that they lost the ability to
     fly and became primarily aquatic. The
    early euornithes also saw the development
    of many traits associated with modern
    birds, like strongly keeled breastbones,
    toothless, beaked portions of their jaws (
    though most non-avian euornithes retained
    teeth in other parts of the jaws).
    Euornithes also included the first
    avialans to develop true pygostyle and a
    fully mobile fan of tail feathers, which
    may have replaced the "hind wing" as the
    primary mode of aerial maneuverability and
     braking in flight.
--- Claim ---
A fully mobile fan of tail feathers may have
    replaced the "hind wing" as the primary
    mode of aerial maneuverability and braking
     in flight for early euornithes.
--- Verdict ---
```

Listing 1: Fact-Check and Evidence Verification Prompt Example