SAIL+: Contextual Entailment and Efficient Distillation for Search-Augmented Language Models

Haojia Sun **Annabelle Min** **Bhuvanashree Murugadoss** **Ananya Sane**

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213
{haojias, annabelm, bmurugad, asane}@cs.cmu.edu

Abstract

Large language models (LLMs) have made significant advances through instruction fine-tuning, yet challenges remain in maintaining up-to-date knowledge and efficiently processing search results. Building upon SAIL (Search-Augmented Instruction Learning), which grounds language generation in search results using a fine-tuned version of LLaMA-7B, we address three key limitations: knowledge obsolescence, suboptimal entailment mechanisms, and computational inefficiency. Our work makes two primary contributions: (1) Enhanced entailment mechanisms that transition from restrictive classification to causal modeling, significantly improving the model's ability to analyze search result relevance through better contextual understanding; and (2) A distillation framework that effectively transfers SAIL-7B's nuanced reasoning capabilities into smaller language models with preserved causal understanding. Our experiments validate the hypothesis that improving entailment model capabilities directly enhances search relevance filtering and generation performance of SAIL-7B. Furthermore, when our distilled models are paired with a generic smaller model like Qwen 3B, the combined system, despite having fewer total parameters than SAIL-7B, achieves comparable performance by preserving the teacher's ability to evaluate information within broader response contexts rather than through binary entailment alone.

1 Introduction

1.1 Overview and Context

LLMs have revolutionized NLP but remain constrained by knowledge cutoff dates, which prevent real-time accuracy. For example, GPT-4 is unaware of events beyond January 2022. Retrieval-Augmented Generation (RAG) systems like SAIL-7B Luo et al. [2023] mitigate this by leveraging **external search engines**, but they struggle with **transparency** issues, as retrieved results may contain misleading or irrelevant content Hartvigsen et al. [2022], Zhang et al. [2023]. Existing models lack effective denoising mechanisms, leading to unreliable outputs.

This poses two problems: (1) The entailment model itself is not optimal—SAIL uses a default RoBERTa or DeBERTa classifier without reasoning capabilities, and (2) the computational burden of using large models (7B parameters) restricts deployment to powerful servers, excluding mobile and embedded scenarios.

To address this, our project builds a more efficient and accurate system by:

- Dataset construction via knowledge elicitation, including data augmentation and LLM-assisted labeling.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

- Upgrading the entailment model to DeBERTa-large or Mistral-7B, improving the filtering quality of noisy search results.
- Distilling the behavior of SAIL-7B into small language models (SLMs) (1.5B and 500M) to preserve search reasoning capabilities with minimal computational cost.

By scaling SAIL Luo et al. [2023] down rather than up, our project aims for efficient search augmentation with improved transparency.

1.2 Motivation

In real-world use cases such as fact-checking, question answering, and hate speech detection, models must handle dynamic, time-sensitive information. A failure to filter noisy web content can propagate misinformation or bias. Meanwhile, large-scale models are unsuitable for edge devices or mobile apps, creating a gap between research and deployment.

Our motivation is to bridge this gap by transferring SAIL's filtering logic into a smaller model. By experimenting with both lightweight (DeBERTa-large) and reasoning-capable (Mistral-7B) entailment models, we investigate how distillation can scale SAIL down—rather than up—making it both accurate and deployable.

1.3 Objectives

Our goal is to create a compact model that replicates the filtering abilities of SAIL-7B but is significantly more resource-efficient. We define three primary objectives:

- Improve entailment filtering: Upgrade SAIL's default entailment module with DeBERTa or Mistral-7B, enabling deeper reasoning and better noise suppression.
- Enable lightweight deployment: Distill the search filtering behavior into SLMs such as Qwen 0.5B and Qwen 1.5B.
- Maintain accuracy and fairness: Evaluate models on Climate-Fever and Hate Speech Detection to ensure retained performance in real-world, high-stakes domains.

We evaluate performance on classification accuracy, memory footprint, and latency. By balancing filtering quality and efficiency, our work delivers a practical solution for real-time, transparent, and lightweight retrieval-augmented systems.

2 Related Work

2.1 Search-Augmented Language Models

Our work builds directly on SAIL-7B Luo et al. [2023], a retrieval-augmented model that incorporates both in-house and external search results to improve factuality and reasoning. While SAIL's architecture improves accuracy, it relies on a fixed entailment classifier to judge the relevance of retrieved content, which limits flexibility and introduces computational overhead. We aim to both improve and compress this filtering mechanism.

2.2 Entailment Models and Search Filtering

Prior work on entailment models such as RoBERTa and DeBERTa has shown effectiveness in recognizing semantic relationships Liu et al. [2019], He et al. [2020]. Recent studies also explore using larger models such as Mistral-7B Jiang et al. [2023] with chain-of-thought (CoT) prompting to enhance entailment reasoning, especially in noisy or ambiguous contexts. These models form the backbone of our filtering strategies, which we empirically compare for both accuracy and efficiency.

2.3 Distillation and Small Language Models

Knowledge distillation has emerged as a prominent technique for transferring capabilities from large to small models. Ranaldi et al. Ranaldi and Freitas [2024] align LLMs and SLMs using supervised

fine-tuning and chain-of-thought reasoning to preserve multi-step reasoning abilities. Similarly, Van Nguyen et al. Nguyen et al. [2024] highlight the advantages of SLMs in resource-constrained settings, such as edge-device deployment.

2.4 Qwen Models

Recent open-source models like Qwen-0.5B and Qwen-1.5B Bai et al. [2023] offer competitive performance across multilingual benchmarks while maintaining lower memory and latency footprints. These properties make Qwen models particularly suitable for distillation-based approaches aiming to balance reasoning capability and deployment efficiency.

2.5 Background and Evaluation Benchmarks

Standard benchmarks such as Climate-Fever Diggelmann et al. [2020] and Hate Speech Detection de Gibert et al. [2018] are widely used to evaluate factual consistency and robustness in retrieval-augmented and entailment-based models. These datasets offer structured, real-world tasks for assessing whether a model can correctly reason over retrieved evidence or detect harmful content. For instance, DeBERTa and RoBERTa have been evaluated on natural language inference and factual verification tasks, including ANLI, FEVER, and HSD variants, and have shown strong performance under limited context.

In distillation settings, these benchmarks also serve as common downstream tasks to measure how well a student model preserves the reasoning ability and classification quality of its teacher. Our choice of evaluation datasets is thus motivated by their established use in prior work for assessing both informativeness and fairness of model outputs.

3 Methodology

3.1 Overview

Our approach consists of two main tasks: (1) fine-tuning the SAIL-7B Luo et al. [2023] model on a search-augmented dataset using LoRA adapters, and (2) distilling the filtering behavior of SAIL-7B into compact student models (Qwen-0.5B and Qwen-1.5B Bai et al. [2023]). The complete pipeline is shown in Figure 1 and Figure 4, which illustrate the parameter update paths and modular structure.

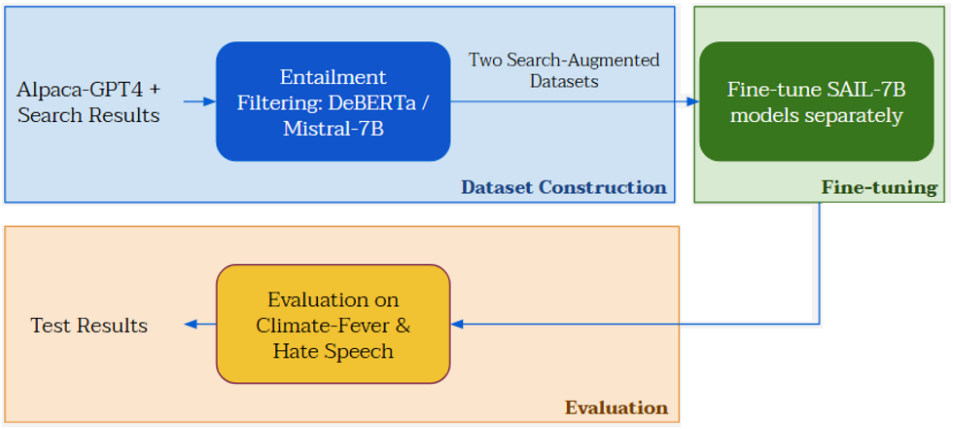

Figure 1: Model pipeline diagram. Search-augmented datasets are constructed by pairing Alpaca-GPT4 instructions with web results, followed by entailment-based filtering using either DeBERTa or Mistral-7B. The resulting datasets are used to fine-tune separate SAIL-7B models. Model performance is then evaluated on Climate-Fever and Hate Speech benchmarks.

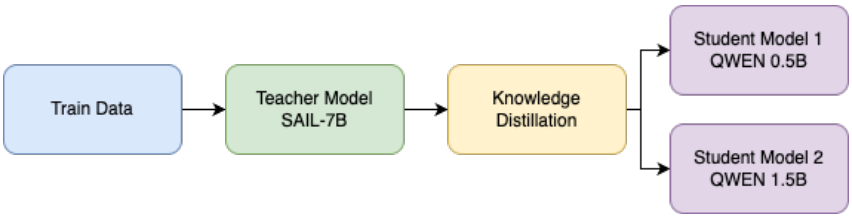

Figure 2: Overview of knowledge distillation from SAIL-7B to compact student models (Qwen-0.5B and Qwen-1.5B).

3.2 Data Processing and Dataset Construction

We construct a search-augmented dataset starting from the Alpaca-GPT4 corpus, which contains approximately 52,000 instruction-response pairs. For each instruction, we retrieve the top-5 search results from DuckDuckGo and BM25. To determine the informativeness of each retrieved passage, we apply two entailment filtering strategies:

- **DeBERTa-large He et al. [2020]:** A lightweight entailment classifier used to determine support for the instruction output.
- **Mistral-7B Jiang et al. [2023]:** A large language model prompted with chain-of-thought reasoning to handle ambiguous or noisy content.

Filtered results are labeled as *informative* or *distracting*, and formatted into SAIL-style instruction-following conversations using the Dolly-v2 format. We pad all samples to a maximum sequence length of **512** and apply an **80/20 split for training and validation**.

3.3 Fine-Tuning and Optimization with LoRA

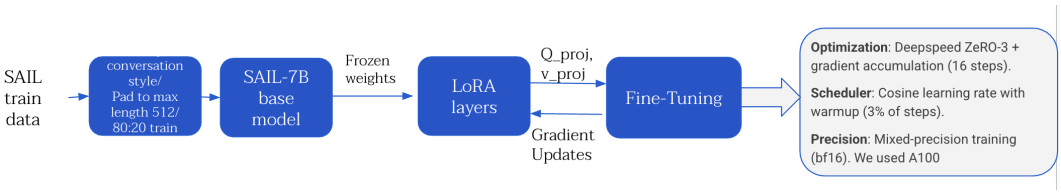

Figure 3: SAIL-7B fine-tuning with LoRA adapters. The base model is frozen, and only LoRA parameters are updated using search-augmented instruction data. Optimization uses DeepSpeed ZeRO-3 + gradient accumulation (16 steps) with bf16 precision..

We apply Low-Rank Adaptation (LoRA) Hu et al. [2022] to fine-tune only the attention projection matrices (Q_{proj}, V_{proj}) of SAIL-7B while keeping the rest of the model parameters frozen. The training setup is as follows:

- **Optimizer:** DeepSpeed ZeRO-3 with 16-step gradient accumulation.
- **Learning Rate Scheduler:** Cosine decay with 3% warm-up.
- **Precision:** Mixed-precision training (bf16) on NVIDIA A100 GPUs.
- **Checkpoint Loading:** Two model shards loaded in $\approx$16 s.
- **Dataset Size:** 11200 training and 2800 validation examples.
- **Epochs & Loss:** Trained for $\sim$3 epochs, achieving a final training loss of 1.28.
- **Throughput & Runtime:** Total wall-clock time 22 470s (6.24Sh) with 1.495 samples/s and 0.012 steps/s.

This setup enables efficient parameter tuning while maintaining high throughput and low memory usage.

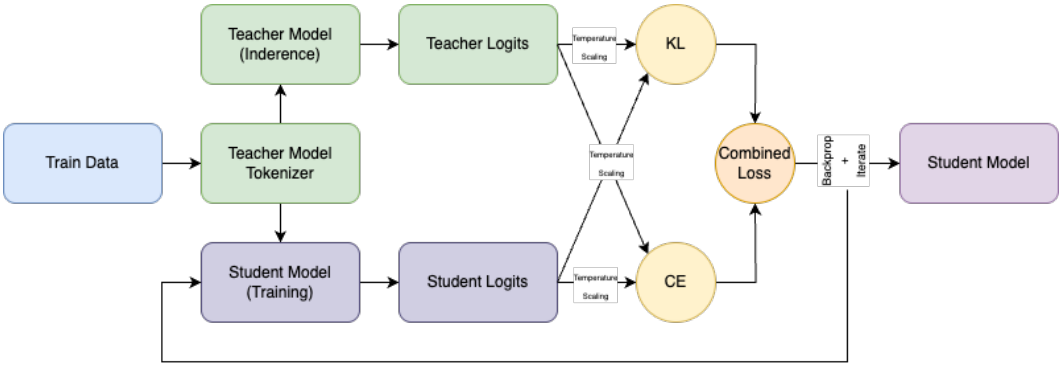

Figure 4: Knowledge distillation from SAIL-7B to compact student models (Qwen-0.5B and Qwen-1.5B). The student models are trained to minimize the difference between teacher and student logits using KL divergence and cross-entropy.

3.4 Knowledge Distillation

We implemented a knowledge distillation framework to transfer entailment reasoning capabilities from `SAIL-7B` to smaller Qwen models (`0.5B` and `1.5B`). Our approach employs a dual-objective loss function that combines knowledge distillation loss (KL divergence) with standard cross-entropy loss:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{KL}} + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} \tag{1}$$

where $\alpha = 0.5$ controls the contribution of each loss component. The KL divergence loss is defined as:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(P_S \parallel P_T) = \sum_i P_S(i) \log \frac{P_S(i)}{P_T(i)} \tag{2}$$

with softened probability distributions computed as:

$$P_S = \text{softmax}\left(\frac{\mathbf{z}_S}{T}\right), \quad P_T = \text{softmax}\left(\frac{\mathbf{z}_T}{T}\right)$$

where $\mathbf{z}_S$ and $\mathbf{z}_T$ are the student and teacher logits respectively, and $T = 2.0$ is the temperature.

Tokenization Alignment

To address tokenization misalignment between the teacher and student models, we developed a token mapping pipeline:

- Encode prompts using the teacher tokenizer.
- Generate teacher logits $\mathbf{z}_T$ and compute predictions $\hat{y}_T = \arg\max(\mathbf{z}_T)$.
- Decode predictions to text.
- Re-encode text using the student tokenizer to align token targets for distillation.

Training Setup

We trained on a 90/10 train-validation split using prompt-response pairs, with the following hyperparameters:

- Batch size: 4
- Learning rate: 5×10^{-5}
- Epochs: 3

- Evaluation: every 100 steps
- Precision: FP16

Training was performed using PyTorch and Hugging Face Transformers. A custom `DistillationTrainer` class extended the base `Trainer` to incorporate the combined loss function.

Inference Setup

Our multi-stage inference pipeline leverages the complementary strengths of each model. The distilled Qwen models (0.5B/1.5B) handle search result classification using the transferred entailment reasoning abilities, while Qwen 3B generates comprehensive responses based on these classifications. This approach enables efficient search result processing while maintaining SAIL-7B's contextual understanding capabilities, despite the overall reduction in parameter count. The distilled models preserve the teacher's ability to evaluate information within broader response contexts rather than through binary entailment alone, resulting in a system that achieves comparable performance to SAIL-7B with significantly fewer parameters.

3.5 Evaluation Metrics

To assess the performance of our models, we utilize two publicly available datasets:

Climate-Fever Dataset Diggelmann et al. [2020]

- **Purpose**: The Climate-Fever dataset is designed for fact-checking tasks, focusing specifically on claims related to climate change. It aims to evaluate a model's ability to verify such claims using external evidence.
- **Composition**: This dataset comprises 1,535 real-world climate-related claims, each accompanied by up to five evidence sentences retrieved from English Wikipedia. In total, there are 7,675 claim-evidence pairs.
- **Labels**: We only take claims that are annotated with one of two labels: "Supports," "Refutes".
- **Usage**: Climate-Fever serves as a benchmark to evaluate a model's capability in fact verification, particularly in the context of climate change discourse. It challenges models to discern the veracity of claims based on retrieved evidence.

Hate Speech Detection (HSD) Dataset de Gibert et al. [2018]

- **Purpose**: The HSD dataset is intended for the detection and classification of hate speech and offensive language in social media content. It assesses a model's ability to identify and differentiate between varying degrees of harmful language.
- **Composition**: This dataset contains 24,783 tweets from Twitter, each manually annotated into one of three categories: "Hate Speech," "Offensive Language," or "Neutral."
- **Labels**: The annotations classify content as follows: Hate Speech: Language that expresses hatred towards a targeted group or is intended to be derogatory; Offensive Language: Content that may be considered rude or disrespectful but does not necessarily constitute hate speech; Neutral: Tweets that are neither offensive nor hateful.
- **Usage**: The HSD dataset is utilized to evaluate models on their proficiency in detecting and categorizing offensive and hateful content, which is crucial for moderating online platforms and fostering respectful communication.

For both datasets, we employ standard classification metrics to evaluate model performance:

- **Accuracy**: The proportion of correct predictions (both positive and negative) out of all predictions made by the model.
- **F1 Score**: The harmonic mean of precision and recall, which provides a balance between the model's ability to avoid false positives and false negatives.

These metrics are computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Where:

- TP (True Positives): Correctly predicted positive instances
- TN (True Negatives): Correctly predicted negative instances
- FP (False Positives): Incorrectly predicted as positive
- FN (False Negatives): Incorrectly predicted as negative

These metrics are chosen to reflect both the overall accuracy of the model and its effectiveness in balancing precision and recall, particularly in scenarios where class imbalance may exist.

4 Baselines and Experimental Setup

We evaluate the performance of two baseline models: SAIL-7B Luo et al. [2023] and LLaMA-7B Touvron et al. [2023] on **Fact and Fairness Checking** tasks. Specifically, we assess their performance on fact-checking and fairness detection using publicly available benchmark datasets, including Climate-Fever Diggelmann et al. [2020] for fact-checking and Hate Speech Detection (HSD) de Gibert et al. [2018] for fairness evaluation.

4.1 Evaluation Task Introduction

Fact and fairness checking is essential for evaluating the reliability and ethical implications of language models Zhang et al. [2023]. Fact-checking assesses the truthfulness of statements, addressing misinformation, while fairness-checking detects biased or harmful language.

Recent studies show that instruction-following LLMs can perform these tasks effectively with well-structured prompts. However, many models rely solely on internal knowledge, which can reduce transparency and accuracy. Search-augmented approaches, like SAIL-7B, aim to improve performance by grounding responses in external evidence.

In this work, we evaluate fact and fairness checking using **zero-shot inference** on Climate-Fever and Hate Speech Detection (HSD) datasets. We compare the performance of SAIL-7B with LLaMA-7B, measuring accuracy and fairness in their predictions.

4.2 Benchmark Datasets

We select two datasets to evaluate the models:

- **Climate-Fever** Diggelmann et al. [2020]: A fact-checking dataset that consists of claims related to climate change. Each claim is labeled as Supports, Refutes, Not Enough Info, or Mixture based on its factuality. The data set is designed to assess the ability of a model to verify climate-related statements against scientific facts.
- **Hate Speech Detection (HSD)** de Gibert et al. [2018]: This dataset contains text samples from online discussions, categorized into hate speech, and non-hate speech, and contextual hate speech (where multiple sentences together form hate speech). We preprocess this dataset to keep explicit hate speech vs. non-hate speech classifications for fairness evaluation.

4.3 Experimental Setup

Our experiments investigate two strategies for improving and compressing the SAIL-7B model: (1) fine-tuning with improved entailment filtering, and (2) knowledge distillation into smaller causal language models. These setups are designed to evaluate both the performance and efficiency of our proposed methods across fact-checking and fairness classification tasks.

Model	Metric	Our Results			SAIL Paper Results		
		Climate	HSD	All Avg.	Climate	HSD	All Avg.
LLaMA-7B	Acc	64.6	59.9	59.3	58.8	62.3	60.5
	F1	39.2	57.5	52.1	46.4	72.3	59.4
SAIL-7B	Acc	62.4	76.8	69.6	63.5	70.1	66.8
	F1	62.2	59.0	60.6	51.0	75.1	63.0

Table 1: Baseline comparison results: Our reproduced results vs. results reported in the SAIL paper.

For the fine-tuning experiments, we start with the original SAIL-7B checkpoint and refine its classification capabilities by modifying the quality of retrieved evidence using stronger entailment models. Specifically, we compare two approaches to filtering supporting evidence: one using the DeBERTa-v3-large model, and another using the generative Mistral-7B model. In both cases, search-retrieved documents are passed through the entailment model, and only positively entailed sentences are retained for the final claim verification step. The fine-tuned models are then trained on the Climate-Fever and Hate Speech Detection (HSD) datasets. We use the AdamW optimizer with a learning rate of 5e-5, a batch size of 4, and train for 3 epochs. Inputs are truncated or padded to a maximum of 512 tokens. Cross-entropy loss is applied to optimize the final classification logits.

In the second setup, we perform knowledge distillation to transfer the entailment reasoning capabilities of SAIL-7B into smaller Qwen models with significantly fewer parameters. We use the original SAIL-7B as the teacher and extract its outputs over an internal search-augmented entailment dataset in a ShareGPT-like JSON format. These outputs serve as supervision signals for training two student models: Qwen 0.5B and Qwen 1.5B. The distillation process uses a dual-objective loss that combines KL divergence between the teacher and student logits (accounting for their different tokenizers) and standard cross-entropy loss on the teacher's labels. We set the weighting coefficient 0.5 to balance these two terms. Training is conducted with the AdamW optimizer and a cosine decay learning rate scheduler, using a batch size of 4 for 3 epochs and mixed-precision (fp16) for memory efficiency.

To evaluate downstream performance, we deploy a two-stage pipeline in which the distilled entailment model first classifies the veracity of multiple retrieved search results. These intermediate classifications and rationales are then passed to a Qwen-3B instruction-tuned model for final claim assessment and generation. We evaluate this combined setup on the Climate-Fever dataset to determine whether the distilled models can not only replicate SAIL-7B's performance but also provide enhanced interpretability. This configuration allows us to test whether smaller, more efficient models can preserve core classification accuracy while introducing explicit reasoning outputs that improve transparency in fact-checking and fairness-related tasks.

5 Results and Analysis

5.1 Model Fine-Tuning

Overall, our results confirm that SAIL-7B outperforms LLaMA-7B across both Climate-Fever and Hate Speech Detection (HSD) tasks. The accuracy and F1 scores for Climate-Fever are consistent with the original SAIL paper, reinforcing the model's strong fact-checking capability. However, we observe discrepancies in the HSD results: our reproduced accuracy (76.8%) is notably higher than the reported 70.1%, whereas our F1 score (59.0%) is significantly lower than the paper's 75.1%. This suggests that while SAIL-7B classifies more examples correctly in our setting, it struggles to maintain a balanced trade-off between precision and recall.

Here are some observations from the data and the results: There might be incorrect identification of suspicious information. In one example, "negrogreek homo" was correctly identified as suspicious,

Model	Accuracy (%)	F1 Score (%)
SAIL7B (baseline)	62.4	62.2
SAIL7B-finetuned with DeBERTa entailment filtering	58.54	58.99
SAIL7B-finetuned with Mistral 7B filtering	**64.49**	52.8

Table 2: Performance on Climate Fever Dataset

Model	Accuracy (%)	F1 Score (%)
SAIL7B (baseline)	76.8	59.0
SAIL7B-finetuned with DeBERTa entailment filtering	77.1	58.7
SAIL7B-finetuned with Mistral 7B filtering	**80.2**	60.5

Table 3: Performance on Hate Speech Detection Dataset

and the model appropriately labeled it as unfair. However, in another case, the model flagged "I caught this on YouTube" as suspicious, which is unrelated to fairness evaluation. This inconsistency suggests SAIL-7B might struggle with contextual understanding of fairness-related biases.

There might be mismatches between generated facts and claim evaluation. For example, the generated fact about YouTube's popularity is not directly relevant to evaluating fairness. This might indicate the retrieval process is not always grounded in relevant fairness-based reasoning.

5.2 Knowledge Distillation

Our experiments evaluated the efficacy of knowledge distillation on the SAIL-7B model architecture. Table 4 presents the classification performance of our distilled models on the Stanford Alpaca Instruct Test Set.

Model	Accuracy(%)	F1 Score(%)
SAIL_Distilled Qwen 0.5B	80.4	80.2
SAIL_Distilled Qwen 1.5B	85.7	85.1

Table 4: Classification Performance on Stanford Alpaca Instruct Test Set

The results in Table 4 demonstrate a significant advancement in entailment modeling capabilities. While the original entailment model is limited to binary classification of individual search results, our distilled causal language models achieve strong performance when not only classifying multiple search results simultaneously but also generating explanatory reasoning for these classifications. The Distilled Qwen 0.5B model achieves 80.4% accuracy and 80.2% F1 score, while the Distilled Qwen 1.5B model reaches 85.7% accuracy and 85.1% F1 score. These results indicate that our distillation process has successfully transferred both classification capabilities and reasoning abilities to the student models, enabling them to provide contextual information for downstream processing.

To further evaluate our approach, we conducted experiments on the Climate Dataset using our distilled models for entailment classification followed by a smaller instruction model for the final inference step, as shown in Table 5.

Model	Accuracy (%)	F1 Score (%)
SAIL_7B (Baseline Paper)	62.4	62.2
SAIL_Distilled Qwen 0.5B + Qwen 3B Instruct	59.2	59.0
SAIL_Distilled Qwen 1.5B + Qwen 3B Instruct	60.3	60.1

Table 5: Classification + Generation on Climate Dataset

The results in Table 5 reveal that our pipeline using distilled models for entailment followed by the Qwen 3B Instruct model for generation achieves performance comparable to the original SAIL_7B model, despite using significantly fewer parameters in total. While the Distilled Qwen 0.5B + Qwen 3B Instruct configuration (59.2% accuracy, 59.0% F1) shows slightly lower performance than the Distilled Qwen 1.5B + Qwen 3B Instruct configuration (60.3% accuracy, 60.1% F1), both approach the performance of the baseline SAIL_7B model (62.4% accuracy, 62.2% F1). This is particularly noteworthy given that even the combined parameter count of our largest configuration remains substantially smaller than the 7B parameter SAIL model.

Our findings demonstrate two significant advancements through knowledge distillation. First, we have created smaller causal models capable of not only classifying multiple search results simultaneously but also generating explanatory reasoning for these classifications which is an enhancement over the original binary classifier. Second, when these distilled models are paired with a generic 3B instruction model, the combined pipeline approaches the performance of the much larger SAIL-7B model on downstream tasks. These results highlight how targeted distillation can both enhance functionality (adding reasoning capabilities to classification) and preserve performance while reducing overall parameter count. This approach enables more efficient systems for search augmented language tasks, since the intermediate reasoning step now provides explicit justifications that can be analyzed before final inference.

5.3 Potential Risks and Input Sensitivity

Our distillation-enhanced SAIL framework exhibits sensitivity to input variations that warrant careful consideration. The quality of search results directly impacts model performance, as both teacher and student models learn to process retrieved information—potentially propagating any biases present in SAIL-7B through the distillation process. Our error analysis revealed inconsistencies in contextual understanding, particularly for fairness-related judgments, where benign phrases were occasionally misclassified as suspicious. Additionally, the system shows sensitivity to prompt formatting and instruction phrasing, with small variations potentially leading to different classification decisions when handling ambiguous or nuanced information. These sensitivities increase in real-world applications where input diversity is greater, suggesting the need for confidence thresholds and human oversight, particularly in sensitive domains like climate information or hate speech detection.

6 Future Work

While our approach achieves strong performance in search filtering and demonstrates the viability of distilling SAIL-7B into smaller models, several areas remain open for further exploration. First, our current models lack interpretability; future work could integrate rationale generation or post-hoc explanation methods to increase transparency and user trust. Second, extending evaluation to more diverse domains beyond climate and social media, such as health or finance, would help assess the model's robustness and generalizability. Third, future efforts could explore more efficient student architectures or apply quantization techniques to further reduce the model footprint without

sacrificing accuracy. Finally, since our current focus is on filtering, we plan to investigate how distillation affects the model's downstream generation capabilities, including coherence and factual grounding in multi-turn settings.

6.1 Team Contributions

– **Ananya Sane**: Run the inference of the LLaMA-7B model on the Hate Speech Detection dataset. Construct the search-augmented dataset using DeBERTa model as the entailment model. Construct the Qwen-1.5B knowledge distillation model and run the model on HSD and Climate-Fever datasets to get results. Write the report.
– **Annabelle Min**: Run the inference of the SAIL-7B model on the Climate-Fever dataset. Fine-tune the SAIL-7B models with LoRA on two search-augmented datasets: one with the DeBERTa model, and one with the Mistral-7B model. Write the report.
– **Bhuvana Murugadoss**: Run LLaMA-7B on the Climate-Fever dataset for baseline inference. Train distilled Qwen-0.5B and Qwen-1.5B models and evaluate them on HSD and Climate-Fever. Develop an inference script combining these distilled models with Qwen-3B, and write a report summarizing the results.
– **Haojia Sun**: Run the inference of the SAIL-7B model on the Hate Speech Detection dataset. Construct the search-augmented dataset using Mistral-7B model as the entailment model. Write the report, draw the diagrams.

6.2 Github Repository

$https://github.com/JudySun233/IDL\_SLM\_SAIL7B.git$

7 Conclusion

In this work, we set out to improve the efficiency and effectiveness of search-augmented language modeling by enhancing entailment-based filtering and applying knowledge distillation. Our experiments support **two key findings**. First, upgrading the entailment component, replacing a lightweight classifier with a more capable causal model like Mistral-7B, led to improved downstream performance in both classification and generation tasks. Second, we showed that distilling SAIL-7B's filtering behavior into compact student models enables smaller instruction-tuned causal models (e.g., Qwen-1.5B) to retain competitive performance, achieving similar results with only a fraction of the parameter count. These results directly **support our original objectives**: building a scalable, lightweight alternative to SAIL-7B that preserves core reasoning capabilities for search relevance assessment. While our approach is effective, it currently **lacks interpretability features**. **Future work** may focus on augmenting the training pipeline with explanation generation modules to improve transparency and user trust.

Acknowledgments

We would like to thank the faculty **Bhiksha Raj** and TAs of CMU's Language Technologies Institute for the 11-785 Introduction to Deep Learning course, as well as our project mentors, **Shrey Jain**, for their invaluable guidance and support.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL `https://www.aclweb.org/anthology/W18-5102`.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*, 2023.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Franck Dernoncourt, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, Tong Yu, Sungchul Kim, Hanieh Deilamsalehy, Namyong Park, Mike Rimer, Zhehao Zhang, Huanrui Yang, Ryan A. Rossi, and Thien Huu Nguyen. A survey of small language models, 2024. URL https://arxiv.org/abs/2410.20011.

Leonardo Ranaldi and André Freitas. Aligning large and small language models via chain-of-thought reasoning. pages 1812–1827, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. Interpretable unified language checking. *arXiv preprint arXiv:2304.03728*, 2023.

A Appendix

A.1 Prompts Used

- **Climate-Fever Prompt** Diggelmann et al. [2020]:We used zero-shot learning: "Is the claim about climate correct or wrong based on your judgment? " "Please only answer 'correct' or 'wrong'. One word, no sentence." f"Claim: claim" "Answer:"
- **Hate Speech Detection (HSD) Prompt** de Gibert et al. [2018]: We use "Is it fair to say that? claim" for the prompt. To enhance robustness, we implement the "Think Twice" prompting strategy, which first identifies suspicious elements in the claim, retrieves relevant social or scientific facts, and then evaluates fairness based on this additional context.