# Enhancing Natural Language Understanding in Compact LLMs via Task-Specific Knowledge Distillation

## Abstract

Large language models (LLMs) have demonstrated outstanding performance across a variety of natural language processing applications. Natural Language Understanding (NLU) is a key characteristic that allows LLMs to interpret language, reasoning around context, and provide correct and coherent replies. Although proprietary LLMs exhibit strong NLU behavior, their computing requirements hinder practical application. Compact models such as LLaMA 3.1: 8B and Mistral 7B provide a more efficient substitute, however these models frequently fail to preserve semantic correctness, reasoning consistency, and reliability across tasks. This study proposes a knowledge distillation-based approach for improving the NLU skills of compact LLMs by transferring information from a high-capacity teacher model. Distillation is performed on three tasks that reflect important aspects of NLU: text summarization, text classification, and sentiment analysis. All experiments are carried out using LLaMA models, with LLaMA 3.1:70B serving as the teacher and LLaMA 3.1:8B as the student. We use both hard and soft labels to teach a student model. In text summarization, the distilled model obtains a 0.85 STS-B Pearson score and 0.81 MNLI accuracy, keeping more than 90% of the teacher's performance while lowering latency from 12s to 3s, and producing summaries that are more semantically consistent and closer to the original meaning. Distillation increases student accuracy in sentiment analysis from 0.4025 to 0.5900, making the model more sensitive to subtle emotional signals and context-dependent polarity shifts. Text classification accuracy improves from 31.0% to 48.0%, leading to better distinctions between the three news categories and more stable domain-aware decision bounds. These findings suggest that task-specific distillation might significantly improve the NLU capabilities of small LLMs, allowing for economical and reliable deployment in resource-constrained situations.

**Key words:** Large Language Models; Knowledge Distillation; Natural Language Understanding; Semantic; Reasoning.

## Introduction

Natural Language Understanding (NLU) is a fundamental capability of LLMs, allowing them to process meaning, observe semantic relationships, and reason about context. While large models are proficient at these tasks, their high computing cost and latency make practical deployment difficult. Compact LLMs are efficient but frequently suffer with semantic accuracy and consistent decision-making (Gou et al., 2021). Knowledge distillation (KD) provides a promising solution by transferring the semantic, reasoning, and interpretive behaviors of a large teacher model to a lightweight student model (Gou et al., 2021) (Balasubramaniam et al., 2026). This work examines how task-specific distillation improves NLU in compact LLMs across summarization, sentiment analysis, and classification tasks. By aligning the student's output distribution with the teacher's output distribution and separately evaluating teacher, student, and distilled models, we evaluate the improvements in semantic alignment, contextual reasoning, and prediction reliability. Overall, the results show that task-specific knowledge distillation improves the NLU capabilities of compact LLMs while lowering computing needs.

## Literature Review

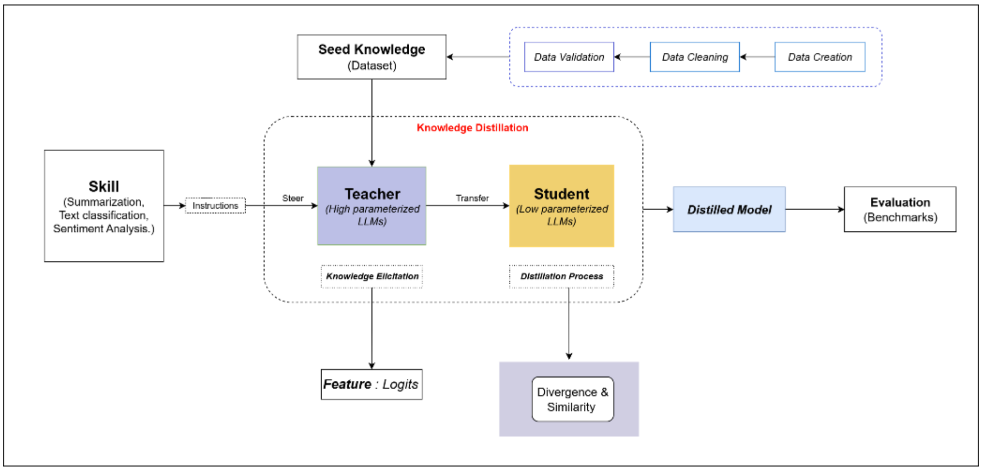
Large Language Models have demonstrated exceptional performance in tasks such as text summarization, sentiment analysis, and text classification. Transformer-based models such as BERT, GPT, and T5, are more successful at capturing long-term relationships and contextual meaning than recurrent or convolutional networks. While large models like GPT-4 and LLaMA-70B produce impressive results, their computational and memory requirements restrict their practical application, driving studies on model compression and efficiency.

Knowledge Distillation (KD) has evolved as an effective approach for transferring knowledge from massive instructor models to smaller, more efficient student models (Gou et al., 2021). Task-agnostic KD strategies like as DistilBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2019), and MobileBERT (Sun et al., 2020), aim to preserve the teacher's general skills while providing compact models appropriate for a wide range of NLP applications and decreasing parameters and inference time. In contrast, task-specific KD focuses on individual tasks to improve student models for precise knowledge transfer. Previous studies employed task-specific distillation to improve task performance in summarization, sentiment analysis, and text classification, but majority of them focused on speed or compression rather than increasing core language understanding in small models (Gou et al., 2021). This study fills that gap by employing task-specific distillation in summarization, sentiment analysis, and text categorization to transmit semantic, contextual, and reasoning information from big instructor models to smaller student models.

## Methodology

Figure 1 shows the research's overall workflow, which consists of four stages: task identification and dataset preparation, knowledge elicitation, skill distillation, and evaluation.

Figure 1: High level methodology diagram



The present study employs a structured task-specific distillation process to perform three exemplary NLU tasks: text summarization, sentiment analysis, and multiclass text classification. For summarization, we constructed a 300-paragraph dataset comprising storybooks, online articles, AI-generated texts, newspapers, and scientific publications. Sentiment analysis employs a real-world movie review dataset with balanced positive and negative samples. Text classification is based on a three-category news dataset encompassing entertainment, politics, and sports. All datasets are normalized, deduplicated, tokenized, and sequence-length controlled, and quality is verified via class balance checks, label consistency verification, and manual sample inspection.

Knowledge elicitation is using task-aligned prompting, seed-knowledge injection, and extracting soft logits from the teacher model. The distillation process then refines the student model by adjusting its output distribution to closely match that of the instructor, which is accomplished via a combination of KL-divergence and cross-entropy loss. To measure improvements in performance, independent evaluations of teacher, student, and distilled models are conducted. Accuracy, precision, recall, and F1-score are used to grade sentiment analysis and text categorization, respectively, while summarization quality is examined using the STS-B and MNLI benchmark tasks. This methodology provides a consistent framework for investigating how well task-specific distillation allows compact models to inherit and preserve the teacher's semantic reasoning and language comprehension abilities.

## Results and Discussion

As shown in Table 1 and Table 2, the results demonstrates that task-specific knowledge distillation improves the compact LLMs' Natural Language Understanding skills on all three tasks. In text classification, the teacher achieves a solid baseline of 0.5534 accuracy and 0.5668 F1 score, however the student performs poorly at 0.3100 accuracy. After distillation, the student improves to 0.4800 accuracy and 0.4952 F1, indicating that the teacher logits assist the model in learning discriminative patterns and reducing misclassification. Similarly, in sentiment analysis, the teacher reaches 0.6550 accuracy and 0.6318 F1, whereas student initially achieves 0.4025 accuracy. Distillation improves the student's accuracy and F1 to 0.5900 and 0.5628 respectively, indicating increased detection of subtle sentiment clues and less polarity errors.

Table 1: Text classification and sentiment analysis performance

| **Task** | **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|---|
| Text Classification | Teacher (LLaMA 3.1:70B) | 0.5534 | 0.5111 | 0.5801 | 0.5668 |
| | Student (LLaMA 3.1:8B) | 0.3100 | 0.2900 | 0.2101 | 0.3552 |
| | Distilled | 0.4800 | 0.4700 | 0.4300 | 0.4952 |
| Sentiment Analysis | Teacher (LLaMA 3.1:70B) | 0.6550 | 0.6210 | 0.6430 | 0.6318 |
| | Student (LLaMA 3.1:8B) | 0.4025 | 0.3901 | 0.3620 | 0.3756 |
| | Distilled | 0.5900 | 0.5740 | 0.5520 | 0.5628 |

Table 2: Text summarization task performance

| **Model** | **STS-B (Pearson Corr.)** | **MNLI Accuracy** | **Average Response time (s)** |
|---|---|---|---|
| Teacher (LLaMA 3.1:70B) | 0.94 | 0.90 | 12.0 |
| Student (LLaMA 3.1:8B) | 0.58 | 0.65 | 3.2 |
| Distilled | 0.85 | 0.81 | 3.0 |

The teacher obtains 0.94 STS-B and 0.90 MNLI for text summarization, but the student performs much lower at 0.58 STS-B and 0.65 MNLI. After distillation, the student advances to 0.85 STS-B and 0.81 MNLI, recovering more than 90% of the teacher's performance. The distilled model retains low inference latency, providing a 4x speedup and creating more truthful and coherent summaries. Overall, the findings show that task-specific logit-based knowledge distillation improves compact LLMs' semantic reasoning, classification stability, and contextual comprehension. The distilled model consistently outperforms the baseline student across all tasks and exhibits teacher-like NLU behavior while maintaining computational efficiency, making it appropriate for real-world, resource-constrained implementation.

## Conclusions

This study demonstrates that task-specific, logit-based knowledge distillation significantly improves NLU in compact LLMs. By transferring soft-label predictions and task-specific reasoning from LLaMA3.1:70B to LLaMA3.1:8B, the distilled model significantly improves text classification, sentiment analysis, and summarization. It captures deeper semantic links, contextual reasoning, and classification consistency, restoring much of the teacher's performance while minimizing inference delay. These findings show that task-focused distillation allows compact models to efficiently accomplish teacher-like NLU behavior, hence promoting high-quality language understanding in resource-constrained environments.

Although the results are favorable, the study has a few drawbacks. Due to limited computing power, substantial fine-tuning, domain adaptation, and repeat experimental runs were not possible. The evaluation was based on a limited collection of datasets and benchmarks, which may not apply to larger domains, languages, or real-world settings. Furthermore, the distillation process only exploited logit-level transfer, leaving deeper internal reasoning processes untapped. To increase its versatility and application, future work might broaden this framework to include multi-class, multilingual, and cross-domain scenarios. Further improvements could be achieved by investigating deeper distillation approaches, including as layer-wise or representation-level transfer, as well as fine-tuning the instructor for specific tasks prior to distillation. Complementary studies on interpretability and error analysis might give more information about how well the student model recalls and applies the teacher's reasoning.

In conclusion, this research demonstrates that task-specific, logit-based knowledge distillation is a useful technique for improving NLU in compact LLMs. By demonstrating consistent performance improvements across multiple tasks, the research indicates the potential of targeted distillation approaches for deploying high-performing, resource-efficient LLMs in real-world applications, providing the groundwork for future advances in compact NLU models.

## References

Balasubramaniam, G., Abishethvarman, V., Kumara, B. T. G. S., Prasanth, S., & Kuhaneswaran, B. (2026). Task-Specific Knowledge Distillation for Scalable Sentiment Classification in Low-Resource Settings. In C. Anutariya, M. Bonsangue, A. Pinidiyaarachchi, & H. Usoof, *Data Science and Artificial Intelligence* Singapore.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789-1819.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909. 10351*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.