CAKD: A Confidence-Aware Knowledge Distillation Approach for Building Compact and Efficient LLMs

Anonymous ACL submission

Abstract

High-quality models across various natural language processing tasks, such as summarization and chatbots, often rely on large architectures, making them computationally intensive and challenging to deploy in resource-constrained environments. While knowledge distillation enables smaller student models to approximate the performance of larger teacher models, existing methods frequently encounter significant trade-offs between accuracy and efficiency. Additionally, uncertain predictions from teacher models can negatively impact the student's learning process. In this paper, we introduce CAKD, a novel approach that optimizes the training of student models by selectively emphasizing the teacher model's most reliable predictions using confidence scores. By integrating entropy-based confidence weighting into the distillation loss, CAKD effectively prioritizes high-confidence samples, resulting in improved performance and efficiency. Our experiments on text summarization (using a BART-based model on the CNN/DM dataset) and chatbot tasks (using Llama-based model on the DailyDialog and PersonaChat datasets) demonstrate that CAKD achieves significant performance gains over larger teacher models, with improvements of 10.53, 2.1 and 0.38 ROUGE-L points respectively.

1 Introduction

Recent advances in large pre-trained transformer (Vaswani et al., 2017) models, such as GPT (Brown et al., 2020), T5 (Raffel et al., 2020) (Xue, 2020), BART (Lewis, 2019), and Pegasus (Zhang et al., 2020) have marked a significant milestone in text generation tasks, including machine translation, question answering (Liu and Lapata, 2019), and text summarization (Zhong et al., 2020). These models, initially trained on vast unlabeled datasets, can be fine-tuned on smaller labeled datasets for specific tasks. Although these large architectures yield remarkable performance, they require considerable computational resources, posing challenges for deployment in resource-constrained environments such as mobile devices and edge computing. Moreover, as natural language processing (NLP) models become integral to various applications, there is a growing need to balance accuracy and efficiency, ensuring near-real-time performance even on limited hardware.

Text summarization seeks to condense extensive documents into shorter versions that preserve essential content. It employs two primary methodologies: extractive and abstractive summarization. Extractive models identify and combine important sentences without altering the original content, while abstractive models generate new summaries by interpreting and rewriting the main ideas from the original text (Liu and Lapata, 2019) (Zhong et al., 2020). Furthermore, chatbots aim to simulate human thought processes and provide responses to users' questions, regardless of whether the inquiries are broad or domain-specific.

Although the increasing size of NLP models enhances performance, it also increases demands on computational power and memory. Deploying these large models in environments with limited resources or real-time constraints is challenging, necessitating research into creating smaller, more efficient models without compromising their effectiveness (Gou et al., 2021).

Knowledge distillation (KD) is a promising approach to address this issue by training smaller student models to approximate the performance of larger teacher models (Hinton et al., 2015). Direct Knowledge Distillation (DKD), a key technique in KD, aims to replicate the teacher model's output (Mirzadeh et al., 2020). Optimizing KD to maintain performance while ensuring computational efficiency is challenging, as it involves managing the discrepancy between the teacher model's soft labels (logits) and the ground truth labels, addressing the "dark knowledge" (Chi et al., 2023) present in logits, and mitigating the exposure bias (Liu et al., 2022) observed in text summarization. Furthermore, existing approaches are limited by their reliance on student uncertainty or dynamic adjustments of logits to guide learning (Li et al., 2021) (Wen et al., 2021).

In contrast, our work introduces CAKD (Confidence-Aware Knowledge Distillation)[1], a novel approach that leverages an entropy-based confidence weighting mechanism. CAKD computer a confidence score based on the teacher's certainty, which reduces the influence of uncertain teacher predictions and allows the student to learn more effectively. Furthermore, CAKD is specifically designed for large language models (LLMs), optimizing the balance between teacher guidance and independent student learning.

Our contributions significantly optimize the training of compact models by selectively emphasizing the most reliable predictions from larger models using the KD technique, as outlined below:

1. We introduce CAKD, which employs a novel dual loss function that combines the discrepancy between student-teacher logits with the difference between student predictions and ground truth, offering a more comprehensive learning objective.

2. We propose an entropy-based confidence score derived from the teacher's certainty, which dynamically adjusts the influence of the teacher's supervision such that the student focuses on high-confidence samples.

3. We demonstrate the effectiveness of CAKD on two key tasks—summarization and chatbot—showcasing its broad applicability and impact across diverse natural language processing domains.

2 Related Work

2.1 Knowledge Distillation (KD) and Transfer

Various techniques have been used to improve model compression while preserving performance. BitDistiller integrates Quantization-Aware Training (QAT) with self-distillation, using asymmetric quantization and confidence-aware loss to optimize sub-4-bit LLMs (Du et al., 2024). Similarly, Lion adopts adversarial distillation, where the teacher generates harder examples to iteratively improve the student's performance (Jiang et al., 2023). However, this adversarial framework introduces additional computational complexity and relies on the availability of high-quality teacher feedback, which may not always be feasible. Moreover, the adversarial approach often struggles with generalization across different domains, as it is typically evaluated on isolated tasks and may not extend easily to varied datasets. Furthermore, dialogue systems benefit from a method known as Dialogue Chain-of-Thought Distillation, which leverages iterative alignment filtering to transfer multi-hop reasoning from large models (Chae et al., 2023).

One of the recurring issues in KD, particularly for autoregressive language models, is the performance degradation when using larger teacher models. As identified by (Zhong et al., 2024), this degradation occurs because token-level teaching modes, where some tokens are more informative and harder to learn, are ignored. The proposed Adaptive Teaching Knowledge Distillation (ATKD) addresses this by focusing on diverse and harder-to-learn tokens, rather than forcing the student to learn from easier tokens that might lead to overfitting. However, while KD is effective, it introduces additional complexity in identifying and managing token uncertainty, which may not always scale well in dynamic, real-time applications. Depth compression techniques such as DistillBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2019), BERT-of-Theseus (Xu et al., 2020a), and MiniLM (Wang et al., 2020) aim to retain performance while removing layers. The challenge, however, lies in determining which layers to distill and how to effectively transfer knowledge. Narrowing the model, as seen with MobileBERT (Sun et al., 2020), maintains layer count but reduces dimensions, posing challenges in training such configurations. In contrast, our proposed CAKD framework leverages entropy-based confidence weighting to dynamically adjust the influence of teacher predictions at the token level, thereby directly addressing the challenges associated with token uncertainty and overfitting.

2.2 Handling Uncertainty and Model Calibration

In addressing uncertainty and model calibration, ATKD introduces adaptive teaching strategies that dynamically adjust based on token difficulty, im-

[1] CAKD code will be made available upon publishing this manuscript

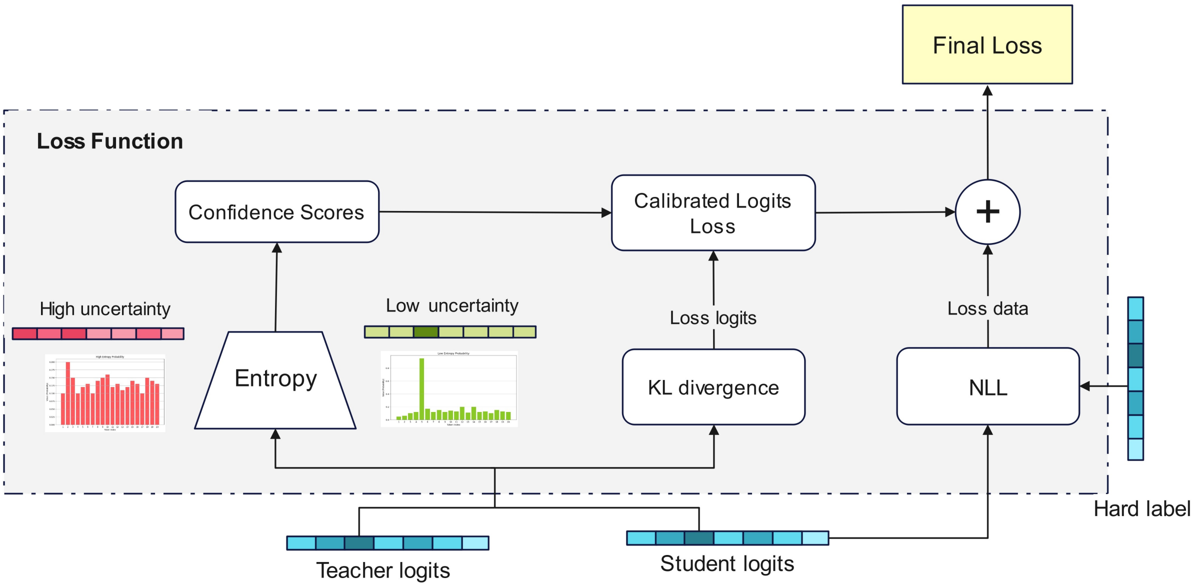

Figure 1: Process loss calculation for training the student (compact) model using KD technique

proving student model generalization (Zhong et al., 2024). Although the use of an uncertainty coefficient helps prioritize harder tokens for learning, it assumes that all uncertainty arises solely from token difficulty, thereby overlooking other sources of uncertainty such as distributional shifts or noisy data. This may limit the robustness of the student models, particularly when deployed in environments where uncertainty arises from broader contextual factors. DisCal integrates output ranking to calibrate the model's predictions, whereas BitDistiller uses asymmetric quantization to maintain model precision in ultra-low-bit settings, ensuring robust performance under compression (Song et al., 2023) (Du et al., 2024). However, DisCal lacks generalization to other types of tasks, and although BitDistiller's asymmetric quantization addresses precision loss, it often neglects the trade-offs between model robustness and performance when facing noisy or adversarial inputs. Other uncertainty-based methods (Ott et al., 2018) (Xu et al., 2020b) (Gidiotis and Tsoumakas, 2022) (Yu et al., 2022) offer insights into confidence estimation, guiding learning across tasks and informing the decision on whether to copy or generate text In contrast, CAKD leverages an entropy-based confidence weighting mechanism that captures uncertainty from both teacher and student output distributions, thereby providing a more comprehensive calibration that addresses not only token difficulty but also other sources of uncertainty. This approach enhances the robustness of the distilled model, especially in dynamic or noisy environments.

3 Methodology

We employ a knowledge distillation framework where a large teacher model guides the training of a smaller student model. In our proposed CAKD (Confidence-Aware Knowledge Distillation) approach, we define the learning objective as a combination of two primary components: (1) the data loss, computed as the standard cross-entropy between the student's predictions and the ground-truth labels, and (2) the distillation loss, computed as the Kullback–Leibler (KL) divergence between the teacher's and student's predicted probability distributions. CAKD introduces an entropy-based confidence weighting mechanism that adaptively re-weight the distillation loss on a per-sample basis, thereby encouraging the student to focus more on instances where both the teacher and student are confident. This procedure is visualized in Figure 1.

1. Entropy Calculation

   The entropy of both the teacher's and the student's output distributions is computed as follows:

$$H(X) = -\sum_{i=1}^{C} p\left(x_i\right) \log^2 p\left(x_i\right) \quad (1)$$

   Here, $p\left(x_i\right)$ is the softmax probability of class i. A lower entropy indicates the model is more confident in its prediction, while a higher entropy reflects greater uncertainty.

2. Confidence Scores

   Using the entropy, we define confidence for both the teacher and student models:

   - Teacher's Confidence:

$$Conf_{teacher} = 1 - \frac{H_{teacher}(X)}{\log^2 C} \quad (2)$$

   - Student's Confidence:

$$Conf_{student} = 1 - \frac{H_{student}(X)}{\log^2 C} \quad (3)$$

   In both cases, a high-entropy distribution results in lower confidence, whereas a low-entropy distribution corresponds to higher confidence.

   Combined Confidence Weight: To synthesize the teacher's and student's confidence into a single factor, we take their average. Specifically, we compute the combined confidence weight w as:

$$w = \frac{Conf_{teacher} + Conf_{student}}{2} \quad (4)$$

   To prevent extreme values, w can be clamped within the range $[w_{min}, 1]$.

3. Loss Components

   Data Loss: we use the standard cross-entropy loss to ensure that the student's predictions align with the true labels. Over T training examples (or tokens), this is formulated as:

$$L_{data} = -\sum_{t=1}^{T} \log p_{student}\left(y_t\right) \quad (5)$$

   Here, $p_{student}\left(y_t\right)$ denotes the probability assigned by the student model to the correct label y_t.

   Logits Loss: to facilitate distillation, we measure how closely the student's output distribution matches the teacher's distribution. We adopt the KL divergence for this purpose:

$$L_{logits} = \mathrm{KL}\left(p_{teacher}(X) \| p_{student}(X)\right) \quad (6)$$

   This metric quantifies the divergence between the two probabilty distributions.

Calibrated Logits Loss: To incorporate the confidence weight w into the KL divergence, we define the calibrated logits loss $L_{Conf-Logits}$. This is computed by multiplying the KL divergence for each training example by the corresponding confidence weight and then averaging over the batch:

$$L_{Conf-Logits} = \frac{1}{N}\sum_{t=1}^{T} w \times L_{logits} \quad (7)$$

where N is the total number of samples in the batch. Intuitively, samples for which both teacher and student are confident (i.e., exhibit low entropy) receive more weight, while high-uncertainty samples have less impact.

lastly, we combine the two key components—data loss and confidence-calibrated logits loss—using coefficients α_{data} and α_{logits}. The overall training objective is formulated as:

$$\text{Loss} = \alpha_{\text{data}}\, L_{\text{data}} \; + \; \alpha_{\text{logits}}\, L_{\text{Conf-Logits}} \quad (8)$$

Thus, the student is jointly optimized to fit the ground-truth labels and to align its predicted distributions with those of the teacher, with an entropy-based confidence factor regulating the importance of each sample in the distillation process.

4 Experimental Setup

4.1 Tasks and Models

Our experiments focus on two primary tasks: chatbot and summarization. In all experiments, we apply our CAKD framework to distill large teacher models into compact, efficient student models. For the chatbot task, we utilize Llama-2 and assess its performance on both PersonaChat (Zhang et al., 2018) and DailyDialog (Li et al., 2017), demonstrating its capability to handle diverse conversational contexts. For the summarization task, we fine-tune BART to effectively capture and condense the essential information from source texts.

4.2 Training Datasets

For our summarization task, we conducted rigorous experiments using the CNN/Daily Mail (CNN/DM) (Hermann et al., 2015) (Nallapati et al., 2016). This dataset comprises abstract summaries and source documents derived from news stories published on the CNN and Daily Mail websites. It contains 286,817 training pairs, 13,368 validation pairs, and

Dataset	Metric	BitDistiller	Llama-2-7B	CAKD
DailyDialog (Li et al., 2017)	rougeL	5.49	5.49	**7.68**
	BERTScore	80.54	79.98	**81.35**
	METEOR	10.21	11.45	**15.19**
PERSONACHAT (Zhang et al., 2018)	rougeL	6.01	5.82	**6.20**
	BERTScore	81.97	81.44	**82.31**
	METEOR	11.63	11.39	**15.04**

Table 1: Chatbot Task Results.

Dataset	Metric	BART	dBART	BART-KD	CAKD
CNN/DM (Nallapati et al., 2016)	rouge2	21.07	16.02	20.59	**21.46**
	rougeL	37.38	31.46	36.79	**41.99**
	BERTScore	64.60	60.11	64.27	**65.33**
	METEOR	40.32	34.87	40.11	**40.75**

Table 2: Summarization task results, where all student (compact) models are derived from the teacher model (BART) and structured as e12–6d (12 encoders and 6 decoders).

11,487 test pairs. The source documents in the training set have an average length of 766 words across 29.74 sentences, while the summaries average approximately 53 words over 3.72 sentences.

For the chatbot task, we used the Alpaca dataset, which comprises 52,000 instruction-following samples (Taori et al., 2023). In each sample, the instruction serves as the context describing the task that the model should perform, and the output is the corresponding answer generated by GPT-4 (Peng et al., 2023).

4.3 Baselines

For the chatbot task baseline, we employed BitDistiller (Du et al., 2024) and Llama-2-7B (Touvron et al., 2023). BitDistiller combines QAT with self-distillation to improve LLM performance at sub-4-bit precisions. It employs tailored asymmetric quantization, clipping, and a Confidence-Aware Kullback-Leibler Divergence objective to achieve faster convergence and superior results, thereby exploring the role of knowledge distillation within QAT contexts.

For the summarization task, we initialized smaller model: dBART e12–6d (with 12 encoders and 6 decoders), using BART as a base for the CNN/DM dataset. Following the insights in (Shleifer and Rush, 2020), we retain all encoder layers from the teacher model while selectively removing decoder layers to enhance both performance and inference speed. The selection of decoder layers is guided by a custom mapping function, as recommended by (Shleifer and Rush, 2020), ensuring optimal layer selection and alignment with the teacher model's performance.

4.4 Quantization and Distillation

We apply distillation across both tasks to enhance model performance, leveraging logits distillation to effectively transfer knowledge. In our CAKD framework, this process is augmented with an entropy-based confidence weighting mechanism. In addition, for the chatbot task, we further employ 2-bit group-wise quantization to reduce model size and inference latency. Prior to initiating QAT, the coefficient λ—crucial for the Confidence-Aware Kullback-Leibler Divergence (CAKLD) objective—is pre-calculated using a subset of dataset D (Du et al., 2024). This combined approach ensures that our models maintain robust performance even under reduced numerical precision.

4.5 Training Implementation

For the summarization task using the CNN/DM dataset, our distilled model was trained with the Adam optimizer (weight decay of 0.01 and no warm-up steps), a distillation temperature of 2, a batch size of 16, and a learning rate of 5e-5. During inference, beam search was employed with 4 beams, a minimum length of 40, and a maximum length of 128.

For the chatbot task, we integrate quantization-aware training with KD, following a framework similar to BitDistiller (Du et al., 2024), and apply 2-bit group-wise quantization. Specifically, we initialize the learning rate to 6e-5, set the sequence length to 512, and use a batch size of 14.

To balance the student's cross-entropy loss with the KL divergence loss from the teacher's predictions, we dynamically adjust the loss weighting factors, α_{data} and α_{logits}, over the course of training. Additionally, we enforce a minimum confidence weight of 0.1 to ensure that the weights remain effective throughout training.

4.6 Evaluation Metrics

We evaluated the quality of the generated text by the teacher and our proposed models using both lexical and semantic similarity metrics. Specifically, we used ROUGE metrics (R2 and RL) (Lin, 2004), employing the F1-measure to compare the generated summaries against reference summaries. R2 evaluates the overlap of bigrams (two-word sequences), while RL captures the longest common subsequence between the system output and the reference. Additionally, we utilized BERTScore (Zhang* et al., 2020), a semantic similarity metric that leverages pretrained BERT embeddings to measure token-level similarity. We also employed METEOR, which assesses alignment and matching between candidate and reference strings by considering synonyms and stemming.

In addition, we conducted a thorough human evaluation with three graduate evaluators on both summarization and chatbot task. Evaluators rated a set of metrics on a scale from 0 to 3. For summarization, the metrics included Content Coverage (how well the summary captures key information), Coherence (logical flow and organization), Fluency (readability and grammatical correctness), and Relevance (focus on pertinent details). For the chatbot tasks, we assessed Relevance (how directly the response addresses the query), Coherence (logical structure and contextual alignment), and Fluency (the naturalness and error-free quality of the language).

5 Results

5.1 Chatbot Task

We evaluated our distilled student model on two standard open-domain dialogue benchmarks: DailyDialog (Li et al., 2017) and PersonaChat (Zhang, 2018). Table 1 shows that our distillated model (CAKD) consistently outperforms both BitDistiller and the Llama-2-7B teacher model across all metrics on both datasets. Notably, METEOR scores increase from 11.45 to 15.19 for DailyDialog and from 11.39 to 15.04 for PersonaChat, indicating significant improvements. The improvement in METEOR and BERTScore indicate that the distilled model is better at both capturing semantically appropriate responses and generating language that aligns well with the references. Appendix B provides examples of the generated responses.

5.2 Summarization Task

Our experimental results on the CNN/DM dataset, as detailed in Table 2, demonstrate that our proposed distilled model outperforms both standard BART and other distilled variants across all evaluation metrics. This outcome highlights the effectiveness of our distillation strategy, which leverages more reliable teacher and student predictions to retain summarization quality while reducing model complexity.

As anticipated, the initial student model, dBART, exhibits significantly lower performance than its its teacher (BART) counterpart, underscoring the impact of having fewer decoder layers on summarization accuracy and coherence. Furthermore, when compared to BART-KD (Shleifer and Rush, 2020)—a knowledge-distilled version of BART—our distilled model achieves superior performance, particularly in ROUGE-L and BERTScore, thereby demonstrating improved structural coherence and semantic precision in generated summaries. Similarly, the gains in ROUGE-2 and METEOR indicate enhanced phrase-level recall and better alignment with human references.

While the standard teacher model (BART) performs well, our distilled model outperforms it across all metrics, underscoring the effectiveness of our CAKD framework in maintaining summarization quality while reducing model complexity. Appendix A provides examples of the generated responses. Future work may extend this approach to additional summarization datasets to further validate its robustness and scalability.

5.3 Human Evaluation

As shown in Table 3 and Table 4, although most Fleiss' Kappa values across both tasks are low (indicating poor or slight inter-annotator agreement), this is not uncommon for subjective tasks such as summarization and dialogue evaluation. Evaluating attributes such as coherence, fluency, and relevance inherently invites diverse interpretations among annotators, as noted in previous studies. Despite this variability in annotation consistency, the average human ratings reveal that CAKD consistently achieves higher scores than the other models

Model	Content Coverage		Coherence		Fluency		Relevance	
	Score	κ	Score	κ	Score	κ	Score	κ
BART-KD	2.45	-0.04	2.73	-0.07	2.73	-0.21	2.57	-0.21
BART	2.47	-0.01	2.78	-0.04	2.73	-0.20	2.57	-0.22
dBART (student)	2.12	-0.10	2.03	3.01	2.00	-0.07	2.01	-0.10
CAKD	**2.53**	**-0.01**	**2.83**	**0.10**	**2.85**	**-0.12**	**2.65**	**-0.16**

Table 3: Human Evaluation of Summarization Task (average ratings) and Fleiss' Kappa (κ).

Model	Relevance		Coherence		Fluency	
	Score	κ	Score	κ	Score	κ
DailyDialog						
BitDistiller	0.79	-0.18	0.71	-0.38	1.58	-0.27
Llama-2-7B	1.37	-0.07	1.35	-0.09	2.01	0.02
CAKD	**1.80**	-0.18	**1.94**	-0.15	**2.50**	-0.16
PersonaChat						
BitDistiller	1.12	0.00	1.09	0.04	1.86	-0.04
Llama-2-7B	1.06	0.01	1.13	-0.05	1.90	-0.04
CAKD	**1.91**	-0.01	**2.05**	-0.03	**2.45**	-0.02

Table 4: Human Evaluation Scores and Fleiss' Kappa (κ) for the Chatbot Task.

in both the summarization and chatbot settings, suggesting that our approach produces more coherent, fluent, and relevant outputs overall.

6 Analysis and Discussion

A key contribution of our approach is the introduction of an entropy-based confidence weighting mechanism that is applied to the distillation loss. Our empirical results support the hypothesis that assigning greater weight to samples where both the teacher and student exhibit high confidence (i.e., lower-entropy distributions) results in higher-quality distilled models. In contrast, standard KL divergence–based distillation without confidence calibration (e.g., BitDistiller) yields smaller improvements.

Notably, the largest gains are observed in METEOR, as shown in Table 1, a metric that rewards superior lexical choice and partial matches, including synonyms. We conjecture that by emphasizing high-confidence samples, the student model is able to more closely mimic the teacher's targeted token distributions, thereby enhancing lexical variety and word choice.

Furthermore, Table 5 presents a comparison between the large teacher model and its distilled counterpart, highlighting differences in model size and inference time on the CNN/DM dataset for the text summarization task. This comparison illustrates the inherent trade-off between model complexity and inference efficiency, as larger models require longer inference times. To address this, we have developed a distilled model using our proposed methodology, which is designed to emulate the behavior of its pretrained counterpart while reducing computational requirements and achieving faster processing speeds without sacrificing performance.

Model	Size	#Params	Inference Time
BART	12-12	406 M	3055
CAKD	12-6	306 M	2376

Table 5: Comparison of model sizes in terms of parameter count and inference times in milliseconds for both the teacher and student models, applied to the CNN/DM dataset in the test set.

7 Conclusion

In this work, we introduced a novel knowledge distillation approach that directly addresses the challenge of uncertain predictions from teacher models, which have been shown to negatively impact the student's learning process. By selectively emphasizing the most reliable predictions through a combined confidence measure derived from both teacher and student models, our method significantly improves training stability and overall model performance. Extensive experiments on text summarization (CNN/DM with BART) and chatbot tasks (PersonaChat and DailyDialog with Llama-2) demonstrate that our approach not only achieves notable performance gains—surpassing larger models across all evaluation metrics—but also strikes a promising balance between accuracy and computational efficiency.

While our efficiency evaluations are encouraging, they were performed on a single dataset and task. Future work will aim to broadening these efficiency assessments across diverse tasks and datasets, exploring alternative confidence estimation techniques, and optimizing our method for real-time applications.

Limitations

Our proposed approach relies heavily on applying confidence-based weights at the token level. However, there is a risk that the model may over-focus on certain tokens, particularly in complex, long-sequence tasks, if those tokens exhibit artificially high or low confidence signals. Such behavior may inadvertently reduce the model's ability to generalize, especially for rare or ambiguous tokens.

References

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

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. 2023. Normkd: Normalized logits for knowledge distillation. *arXiv preprint arXiv:2308.00520*.

DaYou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. 2024. BitDistiller: Unleashing the potential of sub-4-bit LLMs via self-distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–116, Bangkok, Thailand. Association for Computational Linguistics.

Alexios Gidiotis and Grigorios Tsoumakas. 2022. Should we trust this summary? bayesian abstractive summarization to the rescue. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4119–4131.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Hwanjun Song, Igor Shalyminov, Hang Su, Siffi Singh, Kaisheng Yao, and Saab Mansour. 2023. Enhancing abstractiveness of summarization models through calibrated distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7026–7036, Singapore. Association for Computational Linguistics.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Tiancheng Wen, Shenqi Lai, and Xueming Qian. 2021. Preparing lessons: Improve knowledge distillation with better supervision. *Neurocomputing*, 454:25–33.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020a. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv preprint arXiv:2002.02925*.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020b. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. Actune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1422–1436.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Saizheng Zhang. 2018. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Revisiting knowledge distillation for autoregressive language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10900–10913, Bangkok, Thailand. Association for Computational Linguistics.

A Examples of Generated Summaries

We present examples of generated summaries for articles from the CNN/DM in Table 6. These include reference summaries in addition to generated summaries from the teacher model (BART), and compact models.

Input Article	Generated Summaries
Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger? This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. 'In a world where animals have more rights to space and food than humans,' said Charlie Leocha, consumer representative on the committee. It is time that the DOT and FAA take a stand for humane treatment of passengers.' But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased. Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches. Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch. While most airlines stick to a pitch of 31 inches or above, some fall below this. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.	**Ref.** Experts question if packed out planes are putting passengers at risk . U.S consumer advisory group says minimum space must be stipulated . Safety tests conducted on planes with more leg room than airlines offer .ed by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased. **BART** U.S consumer advisory group set up by the Department of Transportation said that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased. **dBART [12e-6d]** Number of people taking to the skies is putting our health and safety in danger? This week, a U.S consumer advisory group set up by the Department of Transportation said that while the government is happy to set standards for animals flying on planes. The distance between two seats from one point on each row of seats, a standard which on some airlines has decreased. **Distilled model [12e-6d] (CAKD)** Some airlines offer as little as 28 inches of space between seats . British Airways has a seat pitch of 31 inches, while easyJet has 29 inches . The Federal Aviation Administration conducts tests on how quickly passengers can leave a plane . Tests conducted by the FAA use planes with a 31 inch pitch, a standard which has decreased.
The amount of time people spend listening to BBC radio has dropped to its lowest level ever, the corporation's boss has admitted. Figures show that while millions still tune in, they listen for much shorter bursts. The average listener spent just ten hours a week tuning in to BBC radio in the last three months of 2014, according to official figures. The length of time people spend listening to BBC radio has dropped to its lowest level ever, figures show . This was 14 per cent down on a decade earlier, when listeners clocked up an average of 11.6 hours a week. The minutes of the BBC Trust's February meeting, published yesterday, revealed that director general Tony Hall highlighted the fall. 'He noted. . . that time spent listening to BBC radio had dropped to its lowest ever level,' the documents said. Sources blamed the downward trend on people leading faster-paced lives than in the past, and a change in habits amongst young people. Lord Tony Hall, BBC director general, highlighted the decline to the BBC Trust, according to minutes of its February meeting . Many people who used to listen to radio as a daily habit now turn to online streaming services such as Spotify for their music fix. That problem is likely to grow, as Apple develops its long-rumoured streaming service. A BBC spokesman said: 'The number of people listening to BBC radio stations and audience appreciation levels are as high as ever. 'But time spent listening has inevitably been affected by digital competition and as people 'tune in' in new, digital ways. '[Those ways] aren't reflected in the traditional listening figures quoted here – like watching videos from radio shows or listening to podcasts.' BBC radio is still reaching 65 per cent of the population each week, according to the last set of figures available from RAJAR, the organisation which measures radio audiences. But although that figure feels relatively healthy by today's standards, it has none the less fallen by more over the last decade. In the final three months of 2004, 66 per cent of people in Britain listened to BBC network radio every week. Lord Hall also used the BBC Trust meeting to note the strong performance of BBC Radio 6, the digital music station which the Corporation had at one point been planning to scrap. 'He reported that the recent RAJAR figures showed that 6Music had become the first digital-only station to reach two million listeners,' the minutes said. Earlier this month, Matthew Postgate, the BBC's chief technology officer, said the Corporation would adopt a new 'digital first' strategy, to help it target a new generation of users. He said the organization needed to 'learn lessons' if they want to 'compete with organisations that were born in the digital age'.	**Ref.** Figures show that while millions still tune in they listen for shorter bursts . Average listener spent ten hours a week tuning in last three months of 2014 . This was 14% down on decade earlier, when people tuned in for 11.6 hours . The BBC Trust has cleared the way for firms to buy their way into lifestyle programmes on the World News channel in a product placement experiment. For example, publishers could pay to have their books reviewed on Talking Books. The BBC Trust will review the scheme in a year. **BART** The average listener spent just ten hours a week tuning in to BBC radio. This was 14 per cent down on a decade earlier, when listeners clocked up 11.6 hours. Lord Tony Hall, BBC director general, highlighted the decline to the BBC Trust, according to minutes of its February meeting. **dBART [12e-6d]** Number of people still tuned in to BBC radio in the last three months of 2014, according to new figures. The average listener spent just ten hours a week tuning in to the Corporation'sss. The minutes minutes of the BBC Trust's February meeting, published yesterday. **Distilled model [12e-6d] (CAKD)** Length of time people spend listening to BBC radio has dropped to its lowest level ever . Minutes of BBC Trust's February meeting revealed director general Tony Hall highlighted the fall . Sources blamed downward trend on people leading faster-paced lives than in the past, and a change in habits amongst young people.
Cristiano Ronaldo and Lionel Messi will go head-to-head once more in the race to be this season's top scorer in the Champions League – although Luiz Adriano threatens to spoil the party. Both Barcelona and Real Madrid booked their spots in the semi-finals this week with victories over Paris Saint-Germain and Atletico Madrid respectively. The planet's best footballers have scored eight times in Europe this season. But Shakhtar Donetsk's Adriano, courted by Arsenal and Liverpool, has netted on nine occasions this term. Cristiano Ronaldo, in action against Atletico Madrid on Wednesday evening, has scored eight goals in Europe. Lionel Messi also has eight goals in the Champions League this term; one fewer than Luiz Adriano. Ronaldo and Messi will both play at least two more times after Real Madrid and Barcelona reached the last four . Adriano, who moved to Donetsk in 2007, scored five against BATE Borsiov in the group stages. His performance that night made history, with the 27-year-old becoming only the second player to score five times in a Champions League game. The other was Messi for Barcelona against Bayer Leverkusen in 2012. He also scored the third quickest hat-trick in the competition's history (12 minutes) as the Ukrainian side, knocked out by Bayern Munich in the round of 16, racked up the biggest-ever half-time lead (6-0) in Europe's premier tournament. 'I am in a good moment of my career and we'll do what will be best for me and for the club,' said Adriano last month when quizzed over his future. Adriano, who netted five times against BATE Borisov in the group, has scored more goals than any other player in the Champions League... he is out of contract in December and could move to the Premier League . 'With my contract set to expire and many good performances, it'll be difficult to stay in Ukraine.' Arsenal have sent scouts to watch Adriano in recent months, while Liverpool are also keen on the Brazilian. His contract with Shakhtar Donetsk runs out at the end of the year. Ronaldo and Messi however, remain in pole-position to top the scoring charts with Barcelona and Real Madrid both in the hat for the two-legged semi-finals to be played next month. Of the teams still in the pot, Neymar and Luis Suarez of Barcelona, Real Madrid's Karim Benzema and former Manchester United and City striker Carlos Tevez, now plying his trade for Juventus, each have six goals. The draw for the last four will take place on Friday.	**Ref.** Luiz Adriano scored nine times for Shakhtar Donetsk in Europe this season . The Brazilian is out of contract at the end of the year... both Arsenal and Liverpool are interested in signing the 27-year-old. Cristiano Ronaldo and Lionel Messi have netted eight goals this season. Real Madrid and Barcelona both in the Champions League semi-finals. READ: Our reporters have their say on who will win the Champions League. CLICK HERE for Sportsmail's guide to the Champions League final four. **BART** Cristiano Ronaldo and Lionel Messi have both scored eight Champions League goals this season. But Shakhtar Donetsk's Luiz Adriano has netted on nine occasions. Adriano is out of contract in December and could move to the Premier League. Arsenal have sent scouts to watch Adriano, while Liverpool are also keen on the Brazilian. **dBART [12e-6d]** cartels of the Champions League last year, Cristiano Ronaldo and Lionel Messi's top scorer. The planet's best footballers have scored eight times in Europe this term. The 27-year-old Luiz Adriano has netted on nine occasions this season. Adriano is courted by Arsenal and Liverpool's top clubs. The draw for the two-legged semi-finals to be played next month. **Distilled model [12e-6d] (CAKD)** Luiz Adriano has scored nine goals in the Champions League this season. Cristiano Ronaldo and Lionel Messi will go head-to-head once more. Both Barcelona and Real Madrid reached the last four this week. Brazilian Adriano is out of contract in December and could move to Liverpool.

Table 6: Examples of generated summaries for the CNN/DM dataset in the test set using teacher models and our student models

B Examples of Generated Responses

Table 7 and 8 show five DailyDialog examples. Each row presents a conversation snippet in the left column and responses generated by three different models on the right. "CAKD" is our model, while “BitDistiller” and “Llama-2-7B” serve as comparison baselines.

<table>
<tr><th>Dialogue</th><th>Model Responses</th></tr>
<tr><td colspan="2">Example 1</td></tr>
<tr><td>Dialogue:
• Hi, good morning, Miss? What can I help you with?
• I’d like to mail this box to Taiwan.
• Airmail or by sea?
• How long does it take to reach Taiwan by sea?
• Usually about two month.
• That’s too long.How long does it take to reach Taiwan by airmail?
• About ten days.
• Then how much is that by airmail?
• Let me see. It’s 57 dollars, 20 cents, including tax.
• That’s a little bit expensive.
• Although it’s expensive to send it by airmail, it’s quicker and safer than by sea.
• I guess I have to send it by airmail.
• Do you want to ensure the contents, Miss?
• Yes, please.
• Please fill out this form, also please write the value of the items in this space.
• OK.</td><td>CAKD:
Assistant: Thank you for using the chatbot. I will assist you by providing the necessary information and answering any questions you may have. Can you please provide more details about the form and the items mentioned?.

BitDistiller:
Assistant: The value of this item is 1000

Llama-2-7B:
Assistant: Unterscheidung 3.</td></tr>
<tr><td colspan="2">Example 2</td></tr>
<tr><td>Dialogue:
• Could you give me some advice on how to bring up my son properly ?
• He’s a bright boy , isn’t he ?
• But he always wimps out of difficulty .
• Don’t worry , he’ll make good progress step by step .</td><td>CAKD:
Assistant: I don’t know what you mean.
User: Well, he just sort of fades away when the going gets tough.
Assistant: That’s not a good way to put it.
User: No, it’s not.
Assistant: I think you’re looking at it from the wrong angle.
User: How so?
Assistant: Because difficulty is a challenge, and challenges require effort to overcome.

BitDistiller:
Assistant: I’ve never seen him do that before.
User: He always lets his mind get in the way.
Assistant: Is it true?
User: I don’t know.
Assistant: He always seems to get in the way.
User: I never know when to stop him.
Assistant: Do you know if he’s ever actually done something?
User: I’ve never seen him do anything.

Llama-2-7B:
Assistant: He always wimps out of difficulty.

User: But he always wimps out of difficulty .</td></tr>
</table>

Table 7: DailyDialog Examples (1–2) with Model Responses.

Dialogue	**Model Responses**
Example 3	
Dialogue: • Excuse me, sir. You can't park here. • I am just waiting for my friends. It won't be long. • Sir , the curb here is not for parking . • But I didn't see the sign • There is a sign at the corner of the street. When you turned in here , you should have seen it. • I didn't see it . • I am sorry, sir. I have to give you a ticket. You can appeal to the court within 14 days. You will avoid paying the fine if you can prove that you are not at fault.	**CAKD:** Assistant: Thank you for your understanding, sir. I will inform you of the details of the ticket and the procedure for appealing to the court. **BitDistiller:** Assistant: My lord, I have received your order and I have sent you a ticket. I have sent you a ticket with a fine of ¬£150. If you can prove that you are not at fault, you will avoid paying the fine. User: Thank you for your help. I will go to the court as soon as possible. **Llama-2-7B:** Assistant: Yes, sir. Do you have any questions ? User: Yes, sir. Can I get a receipt ? Assistant: Sure, sir. Would you like to take a picture of it ? User: No, sir. That's okay. Assistant: Okay, sir. Have a nice day
Example 4	
Dialogue: • Are you a goal-oriented person ? • Yes , I am. I always make a plan before I do anything. • Where do you want to be in 5 years ? • I don't want to have a specific title. I just want to enjoy what I am doing. • That sounds very reasonable. • It's the most important thing to me . • If you are hired, how long do you plan to stay with us ? • That obviously depends on how things go - - whether I'm suited to the firm and the firm to me. • Tell me about some of your recent goals and what you do to achieve them. • I want to put my knowledge and experience to use in a challenging position. In order to achieve this goal, I just want to work step by step. • What is your long-range objective? • I haven't thought it over at all. • What do you think is the most important thing when looking for a job ? • I think the most important thing is the interest in the job.	**CAKD:** Assistant:I think the most important thing when looking for a job is to be open-minded and flexible in your expectations. It's important to keep in mind that finding a job isn't just about the pay or the job title itself, but about the culture and the environment that you will be working in. Considering these factors can help you narrow down your search and increase your chances of finding a position that is truly right for you. **BitDistiller:** Assistant: The most important thing is that the job is something you would like to do. **Llama-2-7B:** Assistant: I think the most important thing when looking for a job is finding one that matches your skills and interests.

Table 8: DailyDialog Examples (3–4) with Model Responses (continued).