CAAF: Confidence and Agreement-Aware Fusion for Multi-LLM Ensembles

Momojit Biswas
TCS Research, India
momojit.biswas@tcs.com

Preethu Rose Anish
TCS Research, India
preethu.rose@tcs.com

Abstract

Integrating multiple Large Language Models (LLMs) into ensembles improves robustness and generalization, yet existing methods often rely on coarse selection or static weighting, failing to integrate trustworthy information at a granular level. We introduce CAAF (Confidence and Agreement-Aware Fusion), a transformer-based encoder-decoder framework designed for dynamic, token-level fusion of diverse LLM outputs. CAAF utilizes a dual-source reliability signal combining LLM self-reported confidence with cross-model semantic agreement to guide the generation process. These signals are injected statically into encoder embeddings and dynamically as attention biases in the decoder, enabling the model to prioritize reliable content from heterogeneous sources. We evaluate CAAF on a comprehensive suite of twelve knowledge-intensive datasets, featuring an ensemble of heterogeneous LLMs. Across twelve knowledge-intensive datasets and averaged over four metrics (Rouge-L, Rouge-LSum, Meteor, and BERTScore-F_1), CAAF attains mean scores of 48.03 and 48.11 using models with only 160M and 330M parameters. In comparison, LLM-Blender achieves average scores of 44.01 (700M) and 45.14 (3B), while GPT-4 based fusion yields mean scores of 44.92.

1 Introduction

The advent of LLMs has significantly advanced machine reading comprehension and open-domain question answering. However, a fundamental challenge persists: for any given query, different LLMs each with distinct inductive biases and training regimes can produce widely varying outputs, ranging from accurate to erroneous (Lu et al., 2024). This variability is particularly pronounced in knowledge-intensive settings such as Retrieval-Augmented Generation (RAG), where a shared retrieved context is processed by multiple models yet yields divergent answers. This variability has motivated ensemble methods that aggregate outputs from multiple LLMs to achieve more accurate, robust, and reliable performance than any single model alone.

Recent research has explored diverse ensemble strategies, including multi-pipeline architectures (Chen et al., 2025b), confidence-based answer selection (Chen et al., 2025a), and sophisticated fusion mechanisms (Tekin et al., 2024; Jiang et al., 2023b), which collectively underscore the importance of model diversity, intelligent weighting, and efficient architectures. However, a significant limitation persists: most existing methods perform static fusion by either selecting a single best answer based on a coarse confidence score or applying fixed weighting schemes. They lack the granularity to integrate trustworthy information at the token level during generation, which often results in complementary knowledge being trapped within sub-optimal individual outputs.

To bridge this gap, we introduce CAAF, a confidence-aware encoder-decoder framework for dynamic answer fusion in multi-LLM RAG systems. CAAF synthesizes improved answers by selectively integrating reliable segments from multiple candidates, guided by a dual-source confidence signal that combines LLM's Self-Reported Confidence (via few-shot prompting) with Semantic agreement across responses. These confidence scores are incorporated into a transformer-based fusion model through two complementary mechanisms: (1) **Encoder**, where confidence is embedded into token representations for static up-weighting of reliable sources; and (2) **Decoder**, where confidence acts as dynamic attention biases to prioritize trustworthy content during generation. This dual injection strategy enables both static and context-aware confidence utilization, supporting fine-grained integration of complementary knowledge.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 456–471
October 17-21, 2026 ©2026 Association for Computational Linguistics

We evaluate CAAF on RAG-Bench (Friel et al., 2024), spanning twelve QA datasets across legal, biomedical, finance, customer support, and general knowledge. With only 160M and 330M parameters, CAAF fuses outputs from multiple LLMs, spanning Mistral-7B-Instruct-v0.2 (Jiang et al., 2023a), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a), and Llama-3-8B-Instruct (Grattafiori et al., 2024) and Gemma-2-2B-it (Team et al., 2024) outperforming finetuned baselines, surpassing LLM-Blender (Jiang et al., 2023b) (700M and 3B), and achieving competitive results against GPT-4 (Achiam et al., 2023), while maintaining a memory footprint under 1 GiB.

The key contributions of this work are: (1) We introduce CAAF, featuring a novel confidence estimation strategy that combines LLM self-confidence with semantic agreement to robustly weigh candidate answers from multiple LLMs; (2) We propose a transformer-based fusion architecture that injects confidence signals both statically into encoder embeddings and dynamically as a bias in the decoder's attention mechanism; (3) We demonstrate that our 160M and 330M parameter models outperform significantly larger ensemble baselines including LLM-Blender at 3B parameters and GPT-4 based fusion across twelve knowledge-intensive QA datasets; and (4) We provide extensive ablation studies and analysis, validating our design choices and offering insights into the model's behavior.

2 Related Work

Recent advances in ensemble methods for NLP and LLMs highlight the limitations of single-model approaches and the benefits of multi-model aggregation for performance, robustness, and generalization. Within the RAG paradigm, Chen et al. (2025b) addressed generalizability gaps in single-pipeline systems through a multi-RAG ensemble, showing via information-theoretic and mechanistic analyses that aggregating diverse pipelines and core modules. Chen et al. (2025a) introduced Mixture-Embedding RAG and Confident RAG, achieving up to 12.4% gains through confidence-based answer selection, underscoring the importance of model-level decision weighting. Christian (2025) extended RAG to long-text and domain-specific tasks through Optimized Ensembled Agency Construct (OEAC), integrating memory-augmented networks, hybrid retrieval, and meta-learning. Other ensemble approaches include Ensemble-GAN (Zhang et al., 2019a), RBAEM (Song et al., 2022) for retrieval-bias reduction, and Jin et al. (2024)'s geometry-aware LoRA adapter for zero-shot generalization. Beyond RAG, LLM ensembles continue to evolve. Agrawal et al. (2025) proposed EnsemW2S, a token-level boosting framework; Fang et al. (2024) introduced LLM-Ensemble for production-scale attribute extraction and LLM-Synergy for medical QA, achieving 96.36% accuracy; Xu et al. (2024) developed EVA for vocabulary alignment; and Tekin et al. (2024) proposed LLM-TOPLA, a pruning-based ensemble improving diversity and efficiency. Iterative and fusion-based systems such as ICE (Omar et al., 2025) and LLM-Blender (Jiang et al., 2023b) further show that critique-iteration, ranking, and generative fusion enhance quality and reduce hallucinations. Overall, prior work demonstrates that (1) model diversity improves robustness, (2) confidence-based and token-level aggregation enhances accuracy, (3) iterative and agentic ensembles aid adaptation and interpretability, and (4) bias-aware and parameter-space ensembles strengthen generalization with lower overhead. However, most methods rely on static weighting, lack dynamic attention, rarely unify RAG and LLM ensembles, and insufficiently address heterogeneous, knowledge-intensive tasks. Building on these gaps, our method integrates multi-level diversity, confidence-aware selection, and attention-guided ensemble weighting.

3 Proposed Method

3.1 Task Formulation

We formulate our problem as text fusion: given multiple candidate answers generated by different LLMs for the same query, the goal is to synthesize a superior final response by selectively integrating the most reliable information from all sources. This setting naturally arises in two scenarios: (1) within a RAG framework, where multiple LLMs operate over the same retrieved context to produce candidate answers; or (2) in a multi-LLM ensemble for a single question without retrieval. Formally, let q denote a query and $\{a_i \mid a_i = L_i(q)\}_{i=1}^{n}$ be the set of candidate answers produced by n LLMs $\{L_1, L_2, \ldots, L_n\}$. The objective is to learn a fusion function F that generates a final answer: $a^* = F(\{a_i\}_{i=1}^{n})$, where F integrates complementary knowledge across candidates while minimizing error propagation. To enable effective fusion, we first estimate the reliability of each candidate

answer. This involves scoring the outputs from multiple LLMs based on confidence and semantic agreement, which serve as key signals for our fusion model.

3.2 Confidence Score Estimation

Before fusion, each candidate answer a_i generated by an LLM L_i must be assigned a reliability score. Directly scoring these candidates is challenging due to the absence of ground-truth labels at this stage, which precludes direct comparison of answers. To address this, we propose a two-pronged strategy:

(1) **Self-Reported Confidence Score:** Each LLM is prompted to generate not only an answer but also a self-assessed confidence score, based on the query and retrieved context. This is achieved through few-shot prompting, which includes examples and instructions designed to elicit both the answer and its confidence estimate. The resulting output consists of two components: the answer a_i and the confidence score c_i. The prompt used for this purpose is detailed in Appendix B.

(2) **Voting-Based Semantic Score:** We hypothesize that answers exhibiting semantic overlap are more likely to be correct. Specifically, if m out of n answers share similar semantic content, they are considered more trustworthy. To quantify this overlap, each answer is encoded into an embedding vector e_i, using the all-MiniLM-L6-v2 (Wang et al., 2020), selected for its strong performance on semantic textual similarity benchmarks and computational efficiency. While we use this model in our experiments, the framework is agnostic to the choice of embedding model. For each embedding e_i, we compute a semantic agreement score as $v_i = \sum_{j \neq i} \cos(e_i, e_j)$. This score reflects the degree of semantic consistency between e_i and all other embeddings. A higher score indicates stronger alignment with peer outputs, suggesting greater reliability.

The moderate yet imperfect correlation between Self-Reported Confidence Score and answer quality (Appendix A) indicates that while confidence provides a useful signal, it is inherently noisy. In contrast, Voting-Based Semantic Score offers a complementary and more stable reliability estimate derived from cross-model consensus. Motivated by this observation, we combine both signals to form a unified confidence score, both scores are normalized to the range $[0, 1]$ using min-max scaling. We combine them via the arithmetic mean of the Self-Reported Confidence Score (c_i) and the Voting-Based Semantic Score (v_i), computed as: $s_i = \frac{c_i + v_i}{2}$. This choice is motivated by three considerations. First, as shown in Appendix A, Self-Reported Confidence scores exhibit similar correlation magnitudes with downstream evaluation metrics across all four models ($r = 0.28$–0.37), providing no empirical basis to systematically upweight one signal over the other. Second, equal weighting introduces no additional hyperparameters, keeping the framework simple and fully reproducible across domains. Third, uniform combination is a well-established strong baseline in ensemble aggregation, recommended precisely when prior evidence does not favor one source over another.

3.3 Confidence-Aware Encoder-Decoder

We designed a transformer architecture that explicitly integrates reliability signals into both the encoding and decoding stages, enabling the model to prioritize trustworthy information throughout the generation process.

(1) **Encoder:** The encoder ingests multiple candidate answers generated by different LLMs, each accompanied by a confidence score derived from the text ranking step (Section 3.2). Each answer a_i is first tokenized into a sequence of tokens and then converted into corresponding token embeddings $E_{i,j}$. To incorporate reliability, the confidence score s_i associated with answer a_i is converted into a trainable embedding vector $v_{a_i} = \text{Linear}(s_i)$ via a linear projection. This vector is then added to every token embedding of the corresponding answer, i.e., $E'_{i,j} = E_{i,j} + v_{a_i}$. By embedding the confidence score directly into the token representations, the encoder's self-attention mechanism is naturally guided to emphasize more reliable answers, while still preserving the semantic content of all inputs. To enable cross-candidate reasoning, these enriched sequences are concatenated into a single unified input sequence: $\mathbf{E} = [E'_{1,\cdot} \parallel E'_{2,\cdot} \parallel \cdots \parallel E'_{n,\cdot}]$. By processing the candidates as a continuous sequence, the encoder's self-attention mechanism is empowered to globally weigh information across all sources. This allows the model to emphasize reliable tokens from high-confidence answers while using them to contextualize or correct information from less certain candidates before the decoding phase begins.

(2) **Decoder:** During decoding, confidence is incorporated dynamically at each generation step. The decoder query vector Q_t interacts with the con-

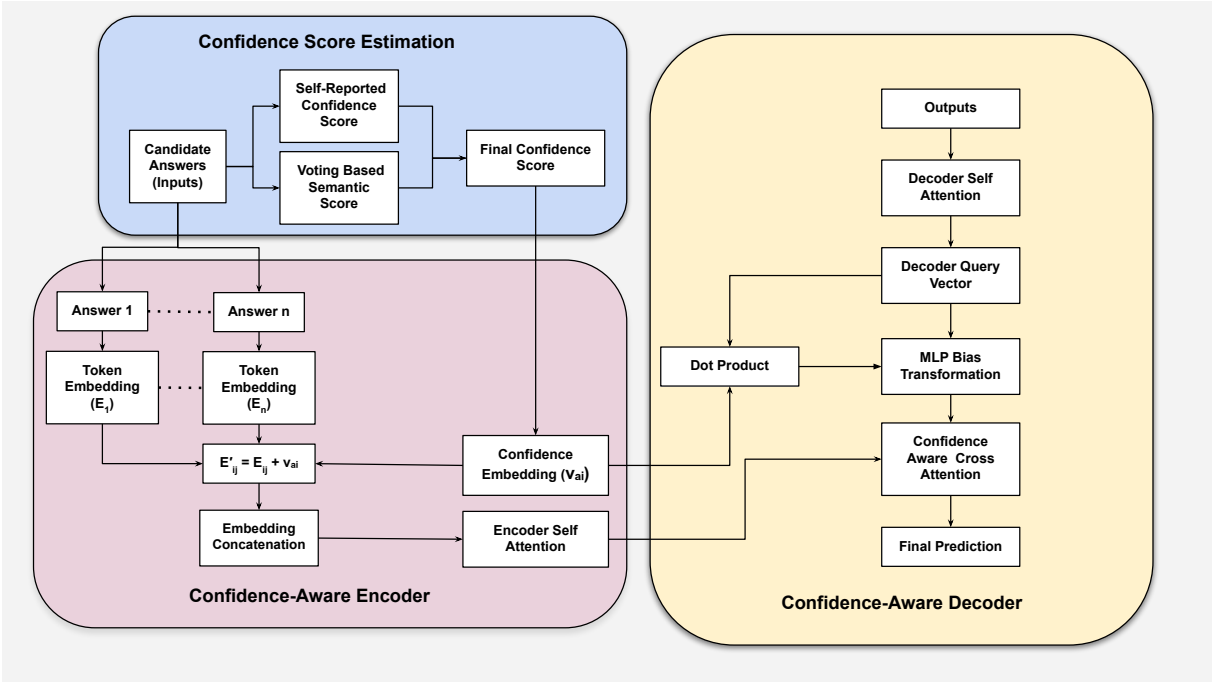


Figure 1: Overall Workflow of the Proposed Method.

fidence embedding via a dot product, $\beta_{t,i} = Q_t \cdot v_{a_i}$. This interaction is passed through a small multilayer perceptron (MLP) to produce a learnable bias term, $\text{Bias}_{t,i} = \text{MLP}(\beta_{t,i})$, which is then added to the standard attention logits, $\tilde{A}_{t,i} = \frac{Q_t \cdot K_i}{\sqrt{d}} + \text{Bias}_{t,i}$, where K denotes the encoder keys and d is the embedding dimension. A SoftMax over these modified logits yields the attention weights, which are then applied to the encoder values V_i to compute the context vector, $\text{Context}_t = \sum_i \text{Softmax}(\tilde{A}_{t,i})\, V_i$. This formulation allows the decoder to modulate its focus based not only on semantic relevance but also on the reliability of each input source, conditioned on the current generation context. As a result, answers with higher confidence receive proportionally greater attention, while less reliable or inconsistent responses are downweighted.

Together, the encoder and decoder form a unified mechanism for reliability-aware generation. The encoder embeds confidence into token-level representations to guide early attention, while the decoder dynamically adjusts its focus based on both semantic and reliability cues. This dual integration ensures that the model attends more strongly to answers that are both semantically relevant and supported by high confidence, thereby enhancing the overall robustness and trustworthiness of the generated output, the corresponding pseudocode is provided in Algorithm 1 and the complete methodology is illustrated in Figure 1.

4 Dataset and Experimentation Details

We detail the experimental framework used to evaluate CAAF, including datasets, model configurations, and competitive baselines.

(1) **RAG-Bench Dataset:** We evaluate on RAG-Bench (Friel et al., 2024), a comprehensive benchmark for retrieval-based QA spanning twelve datasets across five domains: **Biomedical QA** (PubMedQA, CovidQA-RAG), **Legal QA** (CUAD), **Customer-Support QA** (DelucionQA, EManual, TechQA), **Financial QA** (FinQA, TATQA), and **General-Knowledge QA** (HotpotQA, MS Marco, HAGRID, ExpertQA). For each dataset, 50 training instances are used to adapt the confidence-aware encoder–decoder. This minimal fine-tuning is a deliberate design choice: unlike LLM-Blender's (Jiang et al., 2023b) GenFuser, which requires large-scale pretraining on broad instruction-following corpora before deployment, CAAF is designed to be rapidly adaptable to new domains with minimal labeled data. The 50-sample regime tests whether confidence-aware fusion can generalize effectively under practical low-resource constraints, making it applicable in settings where large annotated corpora are unavailable.

(2) **Base Models:** Candidate answers are gen-

Algorithm 1 Confidence-Aware Fusion Framework.

Input: n LLMs $\{L_i\}_{i=1}^{n}$, query q, retrieved context C
Output: Fused answer a^*
Step 1: Self-Reported Confidence Score Estimation
for *each LLM* L_i **do**
Prompt L_i with (q, C) using few-shot examples to jointly produce answer a_i and self-reported confidence c_i
Step 2: Voting-Based Semantic Score Estimation
for *each answer* a_i **do**
Encode a_i into embedding e_i using MiniLM-L6-v2 Compute semantic agreement score: $v_i = \sum_{j \neq i} \cos(e_i, e_j)$
Step 3: Final Confidence Score
for *each answer* a_i **do**
Normalize c_i and v_i to $[0, 1]$ via min-max scaling Compute final reliability score: $s_i = \frac{c_i + v_i}{2}$
Step 4: Confidence-Aware Encoder
for *each answer* a_i **do**
Tokenize a_i into token embeddings $E_{i,j}$; Project s_i into confidence embedding: $v_{a_i} = \text{Linear}(s_i)$; Inject confidence into token representations: $E'_{i,j} = E_{i,j} + v_{a_i}$; Concatenate all enriched sequences: $\mathbf{E}' = [E'_{1,\cdot} \parallel E'_{2,\cdot} \parallel \cdots \parallel E'_{n,\cdot}]$ Pass $\mathbf{E}'$ through encoder self-attention to produce keys K and values V
Step 5: Confidence-Aware Decoder
for *each decoding step* t **do**
for *each candidate* i **do**
Compute confidence bias: $\beta_{t,i} = Q_t \cdot v_{a_i}$; $\text{Bias}_{t,i} = \text{MLP}(\beta_{t,i})$; Compute biased attention logit: $\tilde{A}_{t,i} = \frac{Q_t \cdot K_i}{\sqrt{d}} + \text{Bias}_{t,i}$
Compute context vector: $\text{Context}_t = \sum_i \text{Softmax}(\tilde{A}_{t,i})\, V_i$ Generate output token from Context_t
return a^* (sequence of generated tokens)

erated by four strong open-source LLMs deployed within the RAG pipeline: Mistral-7B-Instruct-v0.2, Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a), Llama-3-8B-Instruct (Grattafiori et al., 2024), and Gemma-2-2B-it-it (Team et al., 2024). These outputs serve as inputs to CAAF. To ensure full reproducibility, we utilize the official Hugging Face release repositories for all those models.

(3) **Backbone and Ensemble Configurations:** We adopt t5 (Raffel et al., 2020) as the backbone, utilizing two variants: t5-small (160M, **CAAF-s**) and t5-base (330M, **CAAF-b**), both initialized from publicly available pretrained weights. The CAAF framework is architecturally agnostic and remains compatible with other encoder-decoder models such as BART (Lewis et al., 2019) or Pegasus (Zhang et al., 2020). We evaluate two ensemble configurations: (a) **MML**: Mistral-7B-Instruct-v0.2, Mistral-7B-Instruct-v0.3, and Llama-3-8B-Instruct; and (b) **GML**: Gemma-2-2B-it, Mistral-7B-Instruct-v0.3, and Llama-3-8B-Instruct, introducing significant model-scale variance to test CAAF's robustness across diverse input pools.

(4) **Benchmark Choice:** We benchmark against two strong baselines: (a) **LLM-Blender** (Jiang et al., 2023b), a state-of-the-art ensemble framework that utilizes a two-stage pipeline consisting of a PairRanker (400M) and a GenFuser (evaluated at 700M and 3B). To ensure a fair and even comparison, we evaluate LLM-Blender using the same candidate answer pools as CAAF across all twelve RAG-Bench datasets. While LLM-Blender was originally pretrained on broad instruction-following data, we adopt its validated generative fusion weights to test CAAF's efficiency against established, large-scale learned fusers in knowledge-intensive domains; and (b) **GPT-4 Fusion**, using direct instruction-based prompting with candidate answers and Self-Reported Confidence scores (Appendix. B). These baselines enable comparison with both learned-ensemble and prompting-based fusion strategies.

(5) **Evaluation Metrics and Hyperparameters:** We report Rouge-L, Rouge-LSum (Lin, 2004), Meteor (Banerjee and Lavie, 2005), and BERTScore-F_1 (Zhang et al., 2019b), averaged across datasets. Training minimizes standard cross-entropy loss using AdamW with learning rate 1×10^{-5}, batch size 4, for up to 20 epochs with early stopping (patience = 3 based on validation loss). During candidate generation, all base LLMs and GPT-4 use temperature 0.1. The CAAF fusion model uses temperature 0.1 for decoding during evaluation. Experiments are run on a single NVIDIA TESLA V100 GPU (32 GiB).

5 Results

5.1 Memory Footprint Analysis

A key strength of our framework is its parameter efficiency and streamlined design. Unlike LLM-Blender, which requires a two-stage pipeline (PairRanker ∼400M plus GenFuser of 700M or 3B parameters), CAAF is a single-model solution, greatly reducing parameter and memory requirements. Memory usage is estimated as $\text{Memory (GiB)} = \frac{\text{Params} \times \text{Bytes per param}}{1024^3}$, with parameters occupying 4 bytes in FP32 or 2 bytes in FP16. The sentence encoder (MiniLM-L6-v2, ∼22M parameters) adds only 0.09 GiB in FP32 and 0.04 GiB in FP16. Table 1 reports memory footprints across configurations. GPT-4 is excluded

Table 1: Memory Footprint Comparison Across Models.

Model	Params	FP32 (GiB)	FP16 (GiB)	Additional Components (GiB)	Total Pipeline (FP32 / FP16)
CAAF-small	160M	0.60	0.30	Sentence Encoder: 0.09 / 0.04	0.69 / 0.34
CAAF-base	330M	1.23	0.62	Sentence Encoder: 0.09 / 0.04	1.32 / 0.66
gen-fuser-700M	700M	2.60	1.30	PairRanker (400M): 1.49 / 0.75	4.09 / 2.05
gen-fuser-3B	3,000M	11.18	5.59	PairRanker (400M): 1.49 / 0.75	12.67 / 6.34

Table 2: Comparison of CAAF against base LLMs and other baselines across twelve RAG-Bench datasets aggregated into five domains. The best model is marked in **bold**; the second best is underlined. All reported results are between [0,100].

Model Group	Model/Ensemble	Biomedical	Legal	Financial	Customer Support	General Knowledge	Overall Avg
Base Models	Mistral-7B-Instruct-v0.2	41.71	32.66	42.94	38.26	40.51	39.22
	Mistral-7B-Instruct-v0.3	47.71	33.01	44.75	41.55	41.90	41.78
	Llama-3-8B-Instruct	41.53	29.72	41.49	35.87	32.31	36.19
	Gemma-2-2B-it	41.06	25.94	33.52	39.59	36.85	35.39
Baselines	LLM-Blender-700M (MML)	50.74	33.04	48.37	42.39	45.49	44.01
	LLM-Blender-3B (MML)	51.42	34.00	50.94	42.72	46.62	45.14
	LLM-Blender-700M (GML)	51.26	27.20	44.91	39.08	46.11	41.71
	LLM-Blender-3B (GML)	53.02	30.09	46.88	41.34	47.17	43.70
	GPT-4 (MML)	47.79	36.83	51.67	42.99	45.33	44.92
	GPT-4 (GML)	47.96	34.97	50.84	44.47	44.58	44.56
Proposed	**CAAF-s (MML)**	50.79	38.54	**53.44**	45.28	45.71	46.75
	CAAF-b (MML)	51.56	**40.15**	53.24	45.28	45.53	47.15
	CAAF-s (GML)	**53.18**	39.46	52.01	48.23	**47.25**	48.03
	CAAF-b (GML)	52.68	39.83	51.95	**49.27**	46.82	**48.11**

due to its API-based nature and undisclosed size, though it is known to be far larger than GPT-3 (175B parameters) (Baktash and Dawodi, 2023), making local deployment impractical. Overall, CAAF achieves competitive performance with just 160M–330M parameters.

5.2 Experimental Results

5.2.1 Overall Performance Comparison

Table 2 presents the main performance comparison across twelve RAG-Bench datasets spanning five domains. The values reported in the table are averages over four evaluation metrics: Rouge-L, Rouge-LSum, Meteor, and BERTScore-F_1 computed across all datasets within each domain. We report aggregated results in the main paper to facilitate comparison across a large number of datasets and domains. Detailed results for each individual metric and dataset are provided in Appendix C.

Among individual base models, Mistral-7B-Instruct-v0.3 is strongest overall (41.78), followed by Mistral-7B-Instruct-v0.2 (39.22), Llama-3-8B-Instruct (36.19), and Gemma-2-2B-it (35.39), but all show a clear performance ceiling, motivating ensemble fusion. The fusion baselines improve on this ceiling, with LLM-Blender-3B (MML) reaching 45.14 and GPT-4 fusion (MML) reaching 44.92, confirming that aggregation helps over any single model.

Both CAAF variants consistently outperform all base models and fusion baselines. CAAF-b (GML) achieves the best overall score of 48.11, surpassing the strongest LLM-Blender configuration (3B, MML) by +2.97 points and GPT-4 fusion (MML) by +3.19 points. This trend holds consistently across all five domains: **Biomedical QA** rises from 47.71 (best base) to 53.18 (CAAF-s, GML), edging LLM-Blender-3B (GML) at 53.02 and well above GPT-4 (GML) at 47.96; **Legal QA** improves from 33.01 to 40.15 (CAAF-b, MML), ahead of GPT-4 (36.83) and LLM-Blender-3B (34.00); **Financial QA** improves from 44.75 to 53.44 (CAAF-s, MML), exceeding GPT-4 (51.67) and LLM-Blender-3B (50.94); **Customer-Support QA** rises from 41.55 to 49.27 (CAAF-b, GML), surpassing GPT-4 (44.47) and LLM-Blender-3B (42.72); and **General-Knowledge QA** improves from 41.90 to 47.25 (CAAF-s, GML), slightly above LLM-Blender-3B (GML) at 47.17 and GPT-4 (MML) at 45.33. Taken together, CAAF delivers consistent gains over both base models and fusion baselines across all evaluated domains.

Table 3: Performance shift between MML and GML ensembles, alongside individual base model scores. All reported results are between [0,100].

Dataset	MML Ensemble		GML Ensemble		Shift		Base Models			
	CAAF-s	CAAF-b	CAAF-s	CAAF-b	$\Delta(s)$	$\Delta(b)$	Mistral-7B Instruct-v0.2	Mistral-3 Instruct-v0.3	Llama-3 8B	Gemma-2 2B
CUAD	38.54	40.15	39.46	39.83	0.92	-0.32	32.66	33.01	29.72	25.94
PubMedQA	44.06	45.40	46.88	46.08	2.82	0.68	34.84	40.48	34.15	32.56
CovidQA-RAG	57.53	57.71	59.49	59.28	1.96	1.57	48.58	54.94	48.91	49.53
FinQA	48.38	47.86	45.29	45.48	-3.09	-2.38	36.75	38.07	35.13	22.21
TAT-QA	58.49	58.61	58.73	58.42	0.23	-0.19	49.14	51.43	47.85	44.82
DelucionQA	49.03	49.52	51.34	51.39	2.32	1.87	42.23	47.68	41.04	40.28
EManual	46.78	46.73	49.90	49.57	3.12	2.85	36.02	40.45	35.12	35.30
TechQA	40.02	39.60	43.46	46.83	3.44	7.23	36.53	36.52	31.47	43.19
HotpotQA	51.46	50.32	53.70	52.17	2.24	1.86	44.24	46.73	33.81	38.03
MS Marco	48.81	49.39	50.72	50.54	1.91	1.15	44.73	45.69	37.33	44.57
HAGRID	46.16	45.61	44.94	44.89	-1.22	-0.72	43.36	40.13	29.84	33.44
ExpertQA	36.42	36.79	39.62	39.67	3.20	2.89	29.72	35.05	28.26	31.36
Average	47.14	47.31	48.63	48.68	1.49	1.37	39.90	42.51	36.05	36.77

5.2.2 Impact of Ensemble Diversity

Table 3 reports CAAF's performance under two ensemble configurations: the MML ensemble and the GML ensemble, alongside individual base-model scores and the per-dataset performance shift Δ (GML score $-$ MML score). Rather than merely testing robustness to model heterogeneity, this comparison reveals how CAAF adapts to changes in ensemble composition by selectively leveraging stronger candidates when they are available. Across datasets, replacing Mistral-7B-Instruct-v0.2 with Gemma-2-2B-it in GML yields a positive average shift for both CAAF variants ($\Delta(s) = +1.49$, $\Delta(b) = +1.37$). In several cases, the direction of the performance shift aligns with the relative strength of the substituted model. For example, on FinQA, where Gemma-2-2B-it performs substantially worse than Llama-3-8B-Instruct (22.21 vs. 35.13) GML leads to the only consistent degradation ($\Delta(s) = -3.09$, $\Delta(b) = -2.38$). Conversely, on datasets where Gemma-2-2B-it is competitive or stronger, such as TechQA (43.19 vs. 31.47), switching to GML results in a pronounced gain, most notably for CAAF-b ($\Delta(b) = +7.23$). While the magnitude of Δ varies across tasks and does not strictly track base-model score differences, these trends indicate that CAAF's fusion mechanism is sensitive to relative candidate quality rather than fixed model identity, enabling it to exploit gains from stronger ensemble members while limiting degradation when weaker models are introduced.

5.2.3 Ablation Study

We isolate the impact of our architectural innovations and reliability signals through extensive ablation studies. First, we establish a learned-fusion performance floor using fine-tuned baselines: *t5-small* (160M parameters, **ft-t5-s**) and *t5-base* (330M parameters, **ft-t5-b**), both trained without confidence injection. Next, we evaluate an encoder-only variant (**CAAF-Enc**, *Ablation-1*) that combines a confidence-aware Encoder with a standard t5 decoder. To assess sensitivity to reliability signals, we conduct two additional studies: *semantic-only fusion* (*Ablation-2*), which uses only the Voting-Based Semantic Score denoted as **Semantic-only-s** for *t5-small* and **Semantic-only-b** for *t5-base* and *confidence-only fusion* (*Ablation-3*), which relies exclusively on the Self-Reported Confidence Score. Together, these ablations quantify the contributions of encoder-level confidence injection, decoder-level biasing, and dual-source reliability signals. All variants are evaluated on both MML and GML ensemble settings to ensure robustness to model variance. All experiments are conducted using the same training and evaluation protocol as the full CAAF models.

Table 4 reports ablations isolating the contributions of (i) the confidence-aware architecture, (ii) encoder-side injection (Ablation-1), (iii) decoder-side confidence-biased cross-attention, and (iv) the dual-source reliability signal (Ablation-2/3).

(1) Architecture vs. Backbone: The full CAAF models yield large gains over finetuned baselines trained on identical inputs without confidence

Table 4: Ablation study results across domains comparing fine-tuned baselines, architectural variants, and reliability-signal configurations under MML and GML ensembles. The best model is marked in **bold**; the second best is underlined. All reported results are between [0,100].

Model Group	Model/Ensemble	Biomedical	Legal	Financial	Customer Support	General Knowledge	Overall Avg
Finetuned Baseline	ft-t5-s (MML)	45.21	35.41	47.48	38.87	40.08	41.41
	ft-t5-b (MML)	45.29	34.59	49.70	38.68	41.25	41.90
	ft-t5-s (GML)	44.64	34.57	47.74	38.27	39.33	40.91
	ft-t5-b (GML)	46.38	34.87	48.26	40.26	41.83	42.32
Ablation-1	CAAF-Enc-s (MML)	47.60	35.96	48.19	42.61	42.45	43.36
	CAAF-Enc-b (MML)	48.61	36.91	48.87	42.10	42.65	43.83
	CAAF-Enc-s (GML)	47.29	38.26	48.73	46.56	44.47	45.06
	CAAF-Enc-b (GML)	47.42	37.23	49.09	45.61	44.29	44.73
Ablation-2	Semantic-only-s (MML)	50.57	37.67	51.54	45.59	45.08	46.09
	Semantic-only-b (MML)	50.24	38.91	51.53	45.41	44.87	46.19
	Semantic-only-s (GML)	48.63	36.53	48.11	45.46	44.75	44.70
	Semantic-only-b (GML)	48.98	36.94	48.08	46.55	44.29	44.97
Ablation-3	Confidence-only-s (MML)	50.02	38.04	51.64	44.94	44.96	45.92
	Confidence-only-b (MML)	50.70	38.62	51.61	45.18	45.07	46.23
	Confidence-only-s (GML)	49.24	36.54	48.07	45.67	44.66	44.84
	Confidence-only-b (GML)	49.60	37.31	47.99	46.70	44.27	45.17
Proposed	**CAAF-s (MML)**	50.79	38.54	**53.44**	45.28	45.71	46.75
	CAAF-b (MML)	51.56	**40.15**	53.24	45.28	45.53	47.15
	CAAF-s (GML)	**53.18**	39.46	52.01	48.23	**47.25**	48.03
	CAAF-b (GML)	52.68	39.83	51.95	**49.27**	46.82	**48.11**

injection. Under GML, CAAF-b improves from 42.32 (ft-t5-b) to 48.11 (+5.79), confirming that gains arise from reliability-aware fusion rather than the t5 backbone alone.

(2) Encoder Contribution (Ablation-1): Encoder-only variants consistently outperform fine-tuned baselines: CAAF-Enc-s (GML) reaches 45.06 vs. 40.91 for ft-t5-s (GML) (+4.15) confirming that static confidence injection already biases fusion toward more reliable candidates.

(3) Decoder Contribution: Despite encoder gains, CAAF-Enc remains notably below full CAAF. The largest gap appears under GML: 48.11 (CAAF-b) vs. 44.73 (CAAF-Enc-b), a +3.38 improvement, supporting the necessity of dynamic confidence-biased cross-attention at each generation step.

(4) Dual Confidence Signal (Ablation-2/3): Semantic-only and Confidence-only variants perform comparably in isolation (e.g., 46.19 vs. 46.23 under MML base), suggesting overlapping but non-identical cues. Both fall short of the full model (47.15 MML; 48.11 GML), confirming that combining cross-model agreement with self-reported confidence yields the most effective reliability signal.

Overall, the ablations establish that static and dynamic reliability conditioning are complementary, and the dual-signal design is necessary for peak performance.

6 Data and Code Availability

All datasets used in this work are publicly available[1]. The code is provided in a public repository[2].

7 Limitations

Our framework, while effective, has several limitations: (1) **Computational Overhead:** Like all ensemble-based approaches, generating multiple candidate answers introduces additional latency, even though the fusion model itself is lightweight and efficient. In this work, we essentially propose a text-fusion method that can be integrated into multi-LLM RAG frameworks to improve robustness and generalization. (2) **Embedding Quality Dependence:** The semantic agreement module relies on MiniLM-L6-v2 embeddings, making confidence estimation sensitive to embedding biases; future work could explore learnable or end-to-end alternatives. (3) **Evaluation Scope:** Experiments were limited to 12 RAG-Bench datasets; broader cross-domain and cross-lingual evaluations are needed to validate general applicability.

[1] https://huggingface.co/datasets/galileo-ai/ragbench/

[2] https://github.com/mb16biswas/Caaf/

8 Conclusion

We present CAAF, a novel confidence-aware encoder-decoder framework for dynamic answer fusion across multiple LLMs within a RAG ensemble. Our key contribution is a two-stage confidence estimation strategy that integrates LLM self-assessment with semantic agreement voting. This confidence signal is injected both statically into the encoder and dynamically into the decoder's attention mechanism via a transformer-based architecture. Extensive evaluations across twelve diverse QA datasets demonstrate that our approach establishes a new state-of-the-art in answer fusion. With only 160M and 330M parameters, CAAF achieves mean scores of 48.03 and 48.11, surpassing substantially larger ensemble methods such as LLM-Blender (700M and 3B parameters) and remaining competitive with GPT-4-based fusion, while maintaining a memory footprint under 1 GiB. This efficiency enables high-quality ensemble performance on commodity hardware. Ablation studies validate the complementary contributions of the confidence-aware encoder and decoder, and together the framework provides a unified, efficient solution for practical ensemble-based QA deployment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aakriti Agrawal, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, Bang An, Bayan Bruss, John Langford, and Furong Huang. 2025. Ensemw2s: Enhancing weak-to-strong generalization with large language model ensembles. *arXiv preprint arXiv:2505.21959*.

Jawid Ahmad Baktash and Mursal Dawodi. 2023. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Shiting Chen, Zijian Zhao, and Jinsong Chen. 2025a. Each to their own: Exploring the optimal embedding in rag. *arXiv preprint arXiv:2507.17442*.

Yifei Chen, Guanting Dong, Yutao Zhu, and Zhicheng Dou. 2025b. Revisiting rag ensemble: A theoretical and mechanistic analysis of multi-rag system collaboration. *arXiv preprint arXiv:2508.13828*.

Ibiloye Abiodun Christian. 2025. Long text language ensemble models: Extension of retrieval-augmented generation (rag).

Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2910–2914.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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. 2023a. Mistral 7b.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023b. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Pengfei Jin, Peng Shu, Sekeun Kim, Qing Xiao, Sifan Song, Cheng Chen, Tianming Liu, Xiang Li, and Quanzheng Li. 2024. Retrieval instead of fine-tuning: A retrieval-based parameter ensemble for zero-shot learning. *arXiv preprint arXiv:2410.09908*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*.

Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. Refining llms outputs with iterative consensus ensemble (ice). *Computers in Biology and Medicine*, 196:110731.

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. *Journal of machine learning research*, 21(140):1–67.

Yiping Song, Zheng Xie, Jianping Li, Luchen Liu, Ming Zhang, and Zhiliang Tian. 2022. Retrieval bias aware ensemble model for conditional sentence generation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6602–6606. IEEE.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size, 2024. *URL https://arxiv.org/abs/2408.00118*, 1(3).

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Llm-topla: Efficient llm ensemble by maximising diversity. *arXiv preprint arXiv:2410.03953*.

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.

Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024. Bridging the gap between different vocabularies for llm ensemble. *arXiv preprint arXiv:2404.09492*.

Jiayi Zhang, Chongyang Tao, Zhenjing Xu, Qiaojing Xie, Wei Chen, and Rui Yan. 2019a. Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 435–444.

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.

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

A Correlation Analysis of Self-Reported Confidence Scores

To assess the usefulness of Self-Reported Confidence as a reliability signal for answer fusion, we compute Pearson correlations between each LLM's reported confidence and downstream evaluation metrics. Table 5 shows consistently positive correlations across all models and metrics, with correlation coefficients ranging from 0.28 to 0.37. Mistral-7B-Instruct-v0.2 exhibits the strongest alignment with Rouge-based metrics, while Llama-3-8B-Instruct shows the highest correlation with BERTScore-F_1. Notably, correlation strength varies across metrics and models, indicating that self-confidence captures complementary aspects of answer quality rather than serving as a fully reliable proxy.

B Prompts Used

The prompts used for GPT-4 Fusion and for obtaining the self-reported confidence score are presented in Table 6 and Table 7, respectively.

C Detailed Experimental Results

This appendix provides the complete set of experimental results across all domains, including all evaluation metrics, baselines, and ablation variants. While the main paper reports averaged performance for clarity, we include full results here for transparency.

C.1 Legal QA

Table 8 reports detailed results on the Legal QA benchmark. Our proposed CAAF variants consistently outperform base models and fine-tuned baselines across all evaluation metrics. Both MML and GML variants show improvements over GPT-4 and LLM-Blender, while ablation results demonstrate performance degradation when removing either encoder or decoder components, confirming their contributions.

C.2 Biomedical QA

Table 9 presents results on CovidQA-RAG and PubMedQA. CAAF models achieve strong and consistent improvements across both datasets compared to base models and fine-tuned baselines. While LLM-Blender shows competitive performance on CovidQA, CAAF variants remain more stable across datasets. Ablation results further highlight the importance of confidence-aware components.

C.3 Customer QA

Detailed results for DelucionQA, EManual, and TechQA are shown in Table 10. CAAF models

Table 5: Pearson Correlation between Self-Reported Confidence and Evaluation Metrics.

Model	Correlation Rouge-L	Correlation Rouge-LSum	Correlation Meteor	Correlation BERTScore-F_1
Mistral-7B-Instruct-v0.2	0.369	0.337	0.298	0.334
Mistral-7B-Instruct-v0.3	0.334	0.328	0.284	0.313
Llama-3-8B-Instruct	0.320	0.325	0.336	0.368
Gemma-2-2B-it	0.310	0.312	0.363	0.311

You are a helpful assistant.

You will receive three answers to the same question, each with a confidence score (0–1). Your task is to produce a single, best possible answer by combining the most accurate and relevant parts of the inputs. Use confidence scores as guidance, but do not copy any answer verbatim. Resolve conflicts and ensure clarity.

Inputs
Answer 1 (Confidence Score: $C_1 = \{c1\}$): {a1}
Answer 2 (Confidence Score: $C_2 = \{c2\}$): {a2}
Answer 3 (Confidence Score: $C_3 = \{c3\}$): {a3}

Output
Provide only the improved final answer. Do not include reasoning or any extra text.

Table 6: Prompt template used for GPT-Fusion.

consistently outperform baseline and fusion-based approaches across all datasets. The results also show that removing confidence-aware components leads to noticeable drops in performance, reinforcing the importance of both encoder and decoder design.

C.4 Financial QA

Table 11 summarizes performance on FinQA and TATQA. CAAF models demonstrate strong improvements over base models and competitive methods such as GPT-4 and LLM-Blender. While some baselines perform well on individual datasets, CAAF maintains more consistent performance across both tasks. Ablation results confirm the contribution of confidence-aware mechanisms.

C.5 General Knowledge QA

Table 12 reports results on HotpotQA, MS MARCO, HAGRID, and ExpertQA. CAAF models achieve competitive or superior performance across most datasets. Although LLM-Blender performs strongly on certain tasks, CAAF shows better consistency overall. Ablation studies again indicate that both encoder and decoder components are necessary for optimal performance.

C.6 Overall Summary

Across all datasets, the full CAAF model consistently outperforms base models, fine-tuned baselines, and most fusion-based approaches. The ablation results demonstrate that both the confidence-aware encoder and decoder contribute significantly to performance gains. These results validate the robustness and generalizability of the proposed framework across diverse QA domains.

You are a helpful assistant. Your task is to answer a question based on a given context.

You must return two things:
1. A concise answer to the question.
2. A confidence score between 0.0 and 1.0 that reflects how confident you are in your answer being correct — whether the answer is present in the context or not.

Guidelines:
- If the answer is **clearly stated in the context**, answer it directly and return a **high confidence score (e.g., 0.9–1.0)**.
- If the answer is **not present**, but you are confident it is not present, respond accordingly and still return a **high confidence score** (e.g., 0.9–1.0).
- If you are **unsure** or the information is ambiguous, return an appropriate answer (e.g., "I don't know") and a **lower confidence score** (e.g., 0.0–0.6).

Always respond in the following format:

```
{
"answer": "<your generated answer>",
"score": <confidence_score>
}
```

Examples:

Example 1
Question: What animal is thought to have originated the 2003 SARS outbreak?
Context: The 2003 SARS outbreak was caused by a coronavirus known as SARS-CoV. It was believed to have originated in civet cats.
Output:

```
{
"answer": "civet cats",
"score": 0.95
}
```

Example 2
Question: Is Favipiravir mentioned as a treatment in the passage?
Context: The passage discusses treatments for COVID-19 including Remdesivir and dexamethasone. It does not mention Favipiravir.
Output:

```
{
"answer": "The answer is not present in the context.",
"score": 0.98
}
```

Example 3
Question: How many countries does the Nile flow through?
Context: The Nile River is over 6,600 kilometers long and flows through multiple countries including Egypt and Sudan.
Output:

```
{
"answer": "I don't know",
"score": 0.3
}
```

Now Answer This:
Question: {question}
Context: {context}

Table 7: Prompt used to obtain Self-Reported Confidence Score.

Table 8: Results on Legal QA. All reported results are between [0,100].

Model	**Rouge-L**	**Rouge-LSum**	**Meteor**	**BERTScore** F_1	**Avg**
Mistral-7B-Instruct-v0.2	19.68	21.09	11.46	78.40	32.66
Mistral-7B-Instruct-v0.3	21.09	22.12	11.76	77.07	33.01
Llama-3-8B-Instruct	17.76	19.99	9.03	72.10	29.72
Gemma-2-2B-it	12.11	12.67	7.30	71.66	25.94
CAAF-s (MML)	27.13	28.62	16.80	81.62	38.54
CAAF-b (MML)	28.33	30.25	20.52	81.51	40.15
CAAF-s (GML)	26.31	15.10	36.62	79.81	39.46
CAAF-b (GML)	26.39	16.57	36.90	79.47	39.83
LLM-Blender-700M (MML)	19.65	21.10	12.39	79.02	33.04
LLM-Blender-3B (MML)	20.51	22.71	14.21	78.58	34.00
LLM-Blender-700M (GML)	12.50	14.08	11.03	71.19	27.20
LLM-Blender-3B (GML)	15.14	17.60	14.94	72.67	30.09
GPT-4 (MML)	24.47	25.93	15.80	81.13	36.83
GPT-4 (GML)	22.75	23.65	14.23	79.27	34.97
ft-t5-small (MML)	24.01	26.05	13.75	77.84	35.41
ft-t5-base (MML)	22.90	24.82	12.40	78.22	34.59
ft-t5-small (GML)	23.00	22.89	13.17	79.21	34.57
ft-t5-base (GML)	23.01	23.94	13.21	79.32	34.87
CAAF-Enc-s (MML) (ab1)	24.32	25.88	14.13	79.51	35.96
CAAF-Enc-b (MML) (ab1)	25.02	26.50	15.71	80.42	36.91
CAAF-Enc-s (GML) (ab1)	25.87	14.80	32.42	79.94	38.26
CAAF-Enc-b (GML) (ab1)	24.70	14.21	31.31	78.72	37.23
Semantic-only-s (MML) (ab2)	24.99	14.01	33.85	77.82	37.67
Semantic-only-b (MML) (ab2)	26.97	17.19	33.96	77.52	38.91
Semantic-only-s (GML) (ab2)	23.37	12.12	33.68	76.95	36.53
Semantic-only-b (GML) (ab2)	23.49	13.73	34.00	76.53	36.94
Confidence-only-s (MML) (ab3)	25.77	13.94	34.69	77.76	38.04
Confidence-only-b (MML) (ab3)	27.56	15.84	33.42	77.65	38.62
Confidence-only-s (GML) (ab3)	23.41	12.12	33.70	76.94	36.54
Confidence-only-b (GML) (ab3)	23.94	14.01	34.39	76.91	37.31

Table 9: Results on Biomedical QA. All reported results are between [0,100].

Model	CovidQA-RAG					PubMedQA				
	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg
Mistral-7B-Instruct-v0.2	38.39	38.45	33.18	84.30	48.58	21.94	21.92	15.55	79.94	34.84
Mistral-7B-Instruct-v0.3	45.99	46.03	41.28	86.45	54.94	28.41	28.47	20.21	84.85	40.48
Llama-3-8B-Instruct	38.72	38.77	33.51	84.63	48.91	20.90	20.96	14.15	80.60	34.15
Gemma-2-2B-it	37.56	30.98	46.15	83.52	49.55	15.69	10.59	29.17	74.78	32.56
CAAF-s (MML)	48.23	48.34	45.94	87.62	57.53	32.45	32.62	25.24	85.91	44.06
CAAF-b (MML)	47.94	47.99	46.97	87.96	57.71	33.14	33.30	29.47	85.70	45.40
CAAF-s (GML)	48.06	45.21	57.20	87.49	59.49	32.53	25.11	43.97	85.90	46.88
CAAF-b (GML)	47.15	45.63	56.86	87.50	59.28	31.36	24.53	43.10	85.35	46.08
LLM-Blender-700M (MML)	51.52	51.55	50.42	87.20	60.18	29.07	29.25	22.01	84.89	41.31
LLM-Blender-3B (MML)	51.02	51.10	49.71	87.17	59.75	30.86	31.07	25.24	85.22	43.10
LLM-Blender-700M (GML)	52.91	52.94	52.24	88.94	61.76	28.50	28.64	21.20	84.72	40.77
LLM-Blender-3B (GML)	54.44	54.46	52.96	89.60	62.87	31.11	31.34	25.25	85.03	43.18
GPT-4 (MML)	43.17	43.20	43.20	86.88	54.11	28.75	28.90	23.10	85.13	41.47
GPT-4 (GML)	44.83	44.88	44.45	87.34	55.37	28.11	28.22	21.26	84.63	40.55
ft-t5-small (MML)	40.97	41.09	40.28	84.26	51.65	26.69	26.75	18.82	82.78	38.76
ft-t5-base (MML)	41.31	41.37	40.36	84.44	51.87	26.78	26.79	18.37	82.88	38.70
ft-t5-small (GML)	41.58	41.75	39.53	85.42	52.07	24.91	24.91	15.46	83.59	37.22
ft-t5-base (GML)	43.27	43.31	41.07	85.83	53.37	27.13	27.17	19.05	84.19	39.38
CAAF-Enc-s (MML) (ab1)	45.02	45.08	42.20	85.07	54.34	29.10	29.20	21.38	83.78	40.87
CAAF-Enc-b (MML) (ab1)	45.75	45.77	43.61	85.61	55.18	30.35	30.52	23.29	83.97	42.03
CAAF-Enc-s (GML) (ab1)	37.62	30.84	47.15	82.52	49.53	30.30	23.48	42.06	84.33	45.04
CAAF-Enc-b (GML) (ab1)	37.46	30.38	47.02	81.52	49.10	30.00	26.20	42.35	84.44	45.75
Semantic-only-s (MML) (ab2)	44.12	41.76	52.37	85.47	55.93	31.08	27.18	39.45	83.13	45.21
Semantic-only-b (MML) (ab2)	43.81	42.80	52.57	85.80	56.25	29.68	26.01	38.37	82.84	44.22
Semantic-only-s (GML) (ab2)	43.92	41.35	52.15	85.36	55.70	28.00	19.74	35.92	82.59	41.56
Semantic-only-b (GML) (ab2)	43.10	41.50	51.80	85.38	55.45	28.63	21.79	36.81	82.81	42.51
Confidence-only-s (MML) (ab3)	44.44	41.89	51.79	85.22	55.84	29.64	26.00	38.35	82.82	44.20
Confidence-only-b (MML) (ab3)	44.52	42.01	52.65	85.58	56.19	31.02	27.25	39.44	83.12	45.21
Confidence-only-s (GML) (ab3)	44.06	41.21	52.20	85.39	55.72	28.97	22.25	37.07	82.77	42.76
Confidence-only-b (GML) (ab3)	43.21	41.66	51.90	85.40	55.55	30.24	23.01	38.04	83.31	43.65

Table 10: Results on Customer-QA. All reported results are between [0,100].

Method	DelucionQA					EManual					TechQA				
	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg
Mistral-7B-Instruct-v0.2	30.60	31.95	23.96	82.42	42.23	23.46	24.42	15.54	80.67	36.02	22.52	24.50	17.90	81.21	36.53
Mistral-7B-Instruct-v0.3	37.13	38.46	30.09	85.04	47.68	28.35	30.62	20.55	82.29	40.45	25.01	25.07	17.60	78.41	36.52
Llama-3-8B-Instruct	29.82	30.59	21.53	82.21	41.04	22.35	23.71	14.57	79.83	35.12	18.32	19.38	11.65	76.52	31.47
Gemma-2-2B-it	25.48	19.56	37.12	78.96	40.28	20.07	13.72	31.98	75.43	35.30	27.34	23.34	39.92	82.17	43.19
CAAF-s (MML)	38.23	40.04	31.75	86.08	49.03	34.27	38.48	29.72	84.64	46.78	28.21	29.33	20.50	82.05	40.02
CAAF-b (MML)	38.74	40.49	32.45	86.38	49.52	34.44	38.06	29.72	84.68	46.73	27.07	29.22	20.18	81.93	39.60
CAAF-s (GML)	40.02	31.01	48.75	85.59	51.34	38.72	29.39	46.83	84.67	49.90	27.47	24.48	40.20	81.69	43.46
CAAF-b (GML)	39.82	31.30	48.75	85.70	51.39	38.19	28.79	46.58	84.73	49.57	32.36	28.64	43.81	82.51	46.83
LLM-Blender-700M (MML)	34.08	35.96	30.66	84.89	46.40	31.51	35.13	27.48	82.37	44.12	23.14	24.37	19.14	79.91	36.64
LLM-Blender-3B (MML)	36.78	39.40	34.36	84.44	48.74	30.04	33.40	23.15	82.23	42.20	22.23	24.59	21.42	80.56	37.20
LLM-Blender-700M (GML)	32.11	34.16	29.67	82.13	44.52	28.63	31.59	23.46	81.17	41.22	16.98	17.62	15.65	75.81	31.51
LLM-Blender-3B (GML)	37.55	40.10	35.86	84.54	49.51	31.45	34.52	26.91	81.97	43.72	15.24	16.24	15.81	75.82	30.78
GPT-4 (MML)	36.93	39.40	32.35	85.74	48.60	30.10	32.69	25.30	83.40	42.87	22.52	24.66	21.87	80.91	37.49
GPT-4 (GML)	37.09	39.07	31.66	85.47	48.32	31.74	34.31	25.61	83.80	43.86	26.55	26.55	31.47	80.32	41.23
ft-t5-small (MML)	32.44	34.09	25.93	81.50	43.49	25.97	29.07	19.98	80.37	38.85	21.03	22.55	15.51	77.99	34.27
ft-t5-base (MML)	32.24	33.42	25.39	81.21	43.06	25.49	27.68	17.82	79.98	37.75	21.99	23.93	15.25	79.73	35.23
ft-t5-small (GML)	27.50	29.20	20.68	81.91	39.82	26.30	27.24	20.45	80.49	38.62	23.29	24.11	19.02	79.01	36.36
ft-t5-base (GML)	32.81	33.94	26.28	83.64	44.17	27.30	29.04	20.52	81.49	39.59	24.87	25.09	19.03	79.06	37.01
CAAF-Enc-s (MML) (ab1)	36.92	38.37	29.47	84.31	47.27	31.79	35.79	26.40	82.31	44.07	23.86	25.14	18.34	78.68	36.51
CAAF-Enc-b (MML) (ab1)	37.32	38.84	30.02	84.58	47.69	32.34	36.16	26.66	82.33	44.37	20.93	22.51	15.48	78.09	34.25
CAAF-Enc-s (GML) (ab1)	37.24	28.88	46.15	83.02	48.82	35.97	26.12	44.00	82.15	47.06	29.09	23.29	42.56	80.26	43.80
CAAF-Enc-b (GML) (ab1)	37.50	28.72	46.29	83.04	48.89	35.03	26.28	43.72	82.02	46.76	25.64	20.12	39.64	79.34	41.19
Semantic-only-s (MML) (ab2)	36.77	28.47	45.73	82.73	48.43	36.14	27.45	44.41	82.19	47.55	26.19	17.38	39.39	80.18	40.78
Semantic-only-b (MML) (ab2)	37.10	29.06	46.12	82.99	48.82	35.63	27.32	44.30	82.23	47.37	25.71	16.02	38.57	79.89	40.05
Semantic-only-s (GML) (ab2)	37.20	28.18	45.88	82.69	48.49	35.23	26.58	43.99	81.35	46.79	24.10	21.12	39.42	79.82	41.11
Semantic-only-b (GML) (ab2)	36.91	28.39	45.84	82.80	48.49	35.76	26.44	43.17	81.30	46.67	29.03	25.31	43.04	80.63	44.50
Confidence-only-s (MML) (ab3)	36.64	28.36	45.62	82.67	48.32	33.95	25.07	43.21	81.13	45.84	25.99	17.17	39.25	80.17	40.65
Confidence-only-b (MML) (ab3)	37.10	29.04	46.12	82.98	48.81	34.61	26.26	43.27	81.22	46.34	25.88	16.84	38.83	80.05	40.40
Confidence-only-s (GML) (ab3)	37.12	28.11	45.85	82.69	48.44	36.27	26.93	44.38	82.21	47.45	24.10	21.12	39.42	79.82	41.11
Confidence-only-b (GML) (ab3)	36.91	28.38	45.84	82.81	48.49	35.78	26.32	44.13	82.27	47.12	29.03	25.31	43.04	80.63	44.50

Table 11: Results on Financial QA. All reported results are between [0,100].

Model	FinQA					TATQA				
	Rouge-L	**Rouge-LSum**	**Meteor**	**BERTScore** F_1	**Avg**	**Rouge-L**	**Rouge-LSum**	**Meteor**	**BERTScore** F_1	**Avg**
Mistral-7B-Instruct-v0.2	25.47	25.85	20.49	75.18	36.75	40.46	40.60	33.57	81.91	49.14
Mistral-7B-Instruct-v0.3	27.32	27.67	21.15	76.16	38.07	43.25	43.53	35.76	83.17	51.43
Llama-3-8B-Instruct	23.08	23.13	19.67	74.62	35.13	39.75	38.82	32.49	80.35	47.85
Gemma-2-2B-it	5.74	4.61	18.92	59.57	22.21	33.23	28.80	42.51	74.76	44.82
CAAF-s (MML)	39.14	39.66	32.55	82.19	48.38	50.15	50.46	46.77	86.59	58.49
CAAF-b (MML)	38.03	39.02	31.96	82.43	47.86	50.18	50.81	46.43	87.03	58.61
CAAF-s (GML)	33.27	26.08	42.74	79.09	45.29	48.34	44.50	56.58	85.48	58.73
CAAF-b (GML)	33.32	26.55	42.85	79.22	45.48	47.70	44.25	56.19	85.55	58.42
LLM-Blender-700M (MML)	24.67	25.05	22.25	75.88	36.96	52.19	52.57	49.34	85.04	59.78
LLM-Blender-3B (MML)	28.79	29.44	26.90	78.73	40.96	53.12	53.64	50.91	85.95	60.91
LLM-Blender-700M (GML)	18.98	19.37	16.85	71.63	31.71	50.02	50.27	47.23	84.95	58.12
LLM-Blender-3B (GML)	21.69	22.27	20.03	73.82	34.45	50.56	51.06	49.61	85.99	59.30
GPT-4 (MML)	35.94	36.77	30.41	81.80	46.23	48.68	48.79	44.12	86.83	57.11
GPT-4 (GML)	35.92	36.58	30.24	80.80	45.89	47.58	48.06	42.13	85.37	55.78
ft-t5-small (MML)	33.36	34.11	28.97	78.14	43.64	41.36	41.89	40.46	81.53	51.31
ft-t5-base (MML)	35.70	36.26	28.96	79.50	45.10	45.82	46.11	41.92	83.31	54.29
ft-t5-small (GML)	31.25	31.90	27.01	79.24	42.35	43.12	43.49	42.61	83.28	53.13
ft-t5-base (GML)	30.95	31.60	24.52	78.43	41.37	46.03	46.45	43.26	84.80	55.14
CAAF-Enc-s (MML) (ab1)	35.97	36.59	28.85	79.63	45.26	42.53	42.90	39.40	79.63	51.12
CAAF-Enc-b (MML) (ab1)	35.89	36.46	29.02	79.54	45.23	44.05	44.44	40.80	80.79	52.52
CAAF-Enc-s (GML) (ab1)	32.79	22.89	40.02	78.61	43.58	43.20	39.70	51.66	80.94	53.88
CAAF-Enc-b (GML) (ab1)	31.31	23.74	40.93	78.99	43.74	43.60	40.46	52.21	81.46	54.43
Semantic-only-s (MML) (ab2)	34.19	27.16	43.03	79.40	45.94	47.07	43.29	55.06	83.12	57.14
Semantic-only-b (MML) (ab2)	33.94	26.85	42.82	79.74	45.84	47.31	42.89	55.11	83.55	57.21
Semantic-only-s (GML) (ab2)	28.30	21.07	37.76	76.43	40.89	44.96	41.11	53.19	82.05	55.33
Semantic-only-b (GML) (ab2)	28.25	21.52	37.82	76.55	41.03	44.61	40.82	52.94	82.10	55.12
Confidence-only-s (MML) (ab3)	34.49	27.37	43.25	79.46	46.14	47.02	43.37	55.05	83.10	57.13
Confidence-only-b (MML) (ab3)	34.24	26.97	43.00	79.80	46.00	47.31	42.91	55.10	83.52	57.21
Confidence-only-s (GML) (ab3)	28.54	21.21	37.94	76.56	41.06	44.73	40.78	52.95	81.89	55.09
Confidence-only-b (GML) (ab3)	28.35	21.55	37.87	76.54	41.07	44.20	40.72	52.68	82.04	54.91

Table 12: Results on General Knowledge QA. All reported results are between [0,100].

Model	HotpotQA					MS Marco				
	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg
Mistral-7B-Instruct-v0.2	35.47	35.47	26.49	79.53	44.24	33.66	33.77	27.66	83.84	44.73
Mistral-7B-Instruct-v0.3	38.21	38.21	29.59	80.91	46.73	35.11	35.31	28.02	84.34	45.69
Llama-3-8B-Instruct	23.35	23.38	14.33	74.16	33.81	25.95	26.02	17.24	80.12	37.33
Gemma-2-2B-it	24.82	18.12	35.38	73.78	38.03	30.68	23.78	41.77	82.04	44.57
CAAF-s (MML)	41.58	41.58	40.58	82.12	51.46	37.78	37.93	33.72	85.82	48.81
CAAF-b (MML)	39.99	39.99	39.68	81.61	50.32	38.38	38.58	34.58	86.02	49.39
CAAF-s (GML)	41.43	40.33	51.25	81.81	53.70	36.90	32.52	47.93	85.52	50.72
CAAF-b (GML)	39.13	39.09	49.56	80.90	52.17	36.58	32.29	47.71	85.56	50.54
LLM-Blender-700M (MML)	44.19	44.19	42.87	83.79	53.76	31.81	31.95	32.33	83.13	44.81
LLM-Blender-3B (MML)	43.81	43.81	43.84	83.67	53.78	33.14	33.26	35.66	83.69	46.44
LLM-Blender-700M (GML)	50.26	50.26	51.52	79.19	57.81	31.69	31.82	32.81	83.10	44.85
LLM-Blender-3B (GML)	49.01	49.01	51.25	78.96	57.06	33.17	33.32	36.19	83.67	46.59
GPT-4 (MML)	42.18	42.18	36.66	82.62	50.91	37.62	37.80	36.10	86.22	49.44
GPT-4 (GML)	42.18	42.18	36.66	82.62	50.91	37.65	37.88	34.18	86.18	48.97
ft-t5-small (MML)	31.78	31.78	30.36	76.58	42.62	33.89	33.97	31.50	82.94	45.58
ft-t5-base (MML)	35.86	33.07	34.72	78.20	45.46	33.82	33.93	28.61	83.06	44.86
ft-t5-small (GML)	32.86	32.86	28.77	78.30	43.19	34.76	34.90	32.25	84.77	46.67
ft-t5-base (GML)	36.94	36.94	36.46	79.35	47.42	34.28	34.35	29.19	84.53	45.59
CAAF-Enc-s (MML) (ab1)	38.17	38.17	36.73	78.92	48.00	35.98	36.13	31.53	83.98	46.90
CAAF-Enc-b (MML) (ab1)	38.09	38.09	36.63	79.00	47.95	36.04	36.22	31.89	84.02	47.04
CAAF-Enc-s (GML) (ab1)	38.86	37.13	48.39	78.70	50.77	34.94	30.05	45.77	83.32	48.52
CAAF-Enc-b (GML) (ab1)	38.05	36.29	47.71	78.44	50.12	34.81	30.47	45.83	83.37	48.62
Semantic-only-s (MML) (ab2)	38.56	37.57	48.45	79.12	50.93	35.73	31.52	46.63	83.67	49.39
Semantic-only-b (MML) (ab2)	36.78	36.37	47.12	78.54	49.70	36.50	32.55	47.31	83.88	50.06
Semantic-only-s (GML) (ab2)	38.58	37.45	48.37	78.89	50.82	34.81	30.44	45.84	83.40	48.62
Semantic-only-b (GML) (ab2)	36.14	36.06	46.55	77.86	49.15	34.49	30.21	45.62	83.44	48.44
Confidence-only-s (MML) (ab3)	38.53	37.48	48.41	79.09	50.88	35.85	31.61	46.71	83.70	49.47
Confidence-only-b (MML) (ab3)	36.97	36.62	47.29	78.61	49.87	36.53	32.53	47.33	83.90	50.07
Confidence-only-s (GML) (ab3)	38.31	37.21	48.16	78.80	50.62	34.79	30.39	45.81	83.40	48.60
Confidence-only-b (GML) (ab3)	36.21	36.21	46.64	77.95	49.25	34.53	30.31	45.66	83.45	48.49

Model	HAGRID					ExpertQA				
	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg	Rouge-L	Rouge-LSum	Meteor	BERTScore F_1	Avg
Mistral-7B-Instruct-v0.2	33.68	33.68	25.50	80.57	43.36	15.02	16.02	10.59	77.25	29.72
Mistral-7B-Instruct-v0.3	30.04	30.04	21.32	79.10	40.13	21.06	22.50	15.70	80.93	35.05
Llama-3-8B-Instruct	18.45	18.47	10.84	71.61	29.84	13.98	14.39	8.27	76.41	28.26
Gemma-2-2B-it	19.13	11.61	30.58	72.44	33.44	14.07	8.29	27.78	75.32	31.36
CAAF-s (MML)	36.38	36.38	30.02	81.85	46.16	22.33	24.02	16.93	82.39	36.42
CAAF-b (MML)	35.64	35.64	29.40	81.77	45.61	22.66	24.21	17.85	82.43	36.79
CAAF-s (GML)	31.39	26.31	42.23	79.83	44.94	24.03	16.68	36.23	81.53	39.62
CAAF-b (GML)	31.43	26.19	42.20	79.75	44.89	23.98	16.80	36.27	81.65	39.67
LLM-Blender-700M (MML)	38.35	38.35	33.02	83.66	48.35	21.26	22.54	16.42	79.95	35.04
LLM-Blender-3B (MML)	40.06	40.06	36.26	84.65	50.26	22.15	23.46	18.07	80.38	36.02
LLM-Blender-700M (GML)	36.17	36.17	31.71	83.02	46.77	21.24	22.60	16.36	79.84	35.01
LLM-Blender-3B (GML)	38.31	38.31	35.25	84.49	49.09	21.98	23.51	18.02	80.33	35.96
GPT-4 (MML)	36.04	36.04	29.14	81.88	45.78	19.89	21.96	17.66	81.25	35.19
GPT-4 (GML)	32.21	32.21	25.07	79.63	42.28	21.53	23.63	17.52	81.93	36.15
ft-t5-small (MML)	30.23	30.23	24.43	77.81	40.67	17.05	18.86	12.84	77.03	31.45
ft-t5-base (MML)	33.09	33.09	27.72	79.57	43.37	17.17	18.76	11.43	77.87	31.31
ft-t5-small (GML)	26.57	26.57	20.54	78.02	37.93	13.80	15.80	10.14	78.41	29.54
ft-t5-base (GML)	30.11	30.11	25.02	77.56	40.70	19.69	20.71	13.30	80.71	33.60
CAAF-Enc-s (MML) (ab1)	32.47	32.47	26.18	78.19	42.33	18.47	20.08	13.10	78.72	32.59
CAAF-Enc-b (MML) (ab1)	32.23	32.23	26.06	78.02	42.13	19.55	21.08	14.13	79.10	33.47
CAAF-Enc-s (GML) (ab1)	28.94	24.29	40.05	78.02	42.82	17.43	14.36	32.12	79.22	35.78
CAAF-Enc-b (GML) (ab1)	28.58	23.59	39.58	77.56	42.33	17.64	14.91	32.43	79.41	36.10
Semantic-only-s (MML) (ab2)	32.49	26.14	42.27	78.96	44.96	17.64	14.93	29.25	78.28	35.03
Semantic-only-b (MML) (ab2)	31.78	25.48	42.72	78.85	44.71	17.68	14.88	29.55	77.88	35.00
Semantic-only-s (GML) (ab2)	29.49	24.41	40.33	77.93	43.04	18.47	15.48	32.89	79.28	36.53
Semantic-only-b (GML) (ab2)	29.58	24.18	40.30	77.85	42.98	18.40	15.67	32.94	79.39	36.60
Confidence-only-s (MML) (ab3)	32.47	26.13	42.25	77.94	44.70	17.71	14.96	28.26	78.24	34.79
Confidence-only-b (MML) (ab3)	32.74	26.51	42.72	77.87	44.96	17.79	15.84	29.58	78.27	35.37
Confidence-only-s (GML) (ab3)	28.55	24.46	40.38	77.94	42.83	18.50	15.56	32.93	79.28	36.57
Confidence-only-b (GML) (ab3)	28.52	24.28	40.30	77.86	42.74	18.38	15.65	32.93	79.40	36.59