Learn to pick the winner: Black-box ensembling for textual and visual question answering

Yuxi Xia [1,2*], Klim Zaporojets[3], Benjamin Roth[1,4],

[1]Faculty of Computer Science, University of Vienna,
[2] UniVie Doctoral School Computer Science, University of Vienna,
[3]Department of Computer Science, Aarhus University,
[4]Faculty of Philological and Cultural Studies, University of Vienna,

***Correspondence:** yuxi.xia@univie.ac.at

Abstract

A diverse range of large language models (LLMs), e.g., ChatGPT, and visual question answering (VQA) models, e.g., BLIP, have been developed for solving textual and visual question answering tasks. However, fine-tuning these models is either difficult, as it requires access via APIs, rendering them as black-boxes, or costly due to the need to tune a large number of parameters. To address this, we introduce ***InfoSel***, a data-efficient ensemble method that learns to dynamically pick the winner from existing black-box models for predictions on both textual and multimodal visual question answering tasks. Unlike traditional ensemble models, ***InfoSel*** does not rely on prediction probabilities or confidences, which typically are not available in black-box models. Experimental results on four datasets demonstrate that our approach achieves an absolute increase of up to +5% in the F1-score compared to standalone LLMs using only 1K training instances.

1 Introduction

Large language models (LLMs) have demonstrated remarkable proficiency across a wide range of tasks (Laskar et al., 2023; Yuan et al., 2024). For example, ChatGPT finds extensive utilization in daily textual question answering (TQA) tasks, rendering substantial convenience to a myriad of users (OpenAI, 2024). Furthermore, for visual question answering (VQA) tasks, VQA models have exhibited exceptional versatility, primarily due to their capability to comprehend both visual and textual context (Gong et al., 2023).

However, recent work (Mao et al., 2024; Laskar et al., 2023) indicates that LLMs, such as ChatGPT, fall short of state-of-the-art performance on task-specific datasets such as question answering. Similarly, VQA models (Li et al., 2022, 2021b; Bao et al., 2022) face challenges when applied to specialized datasets due to the idiosyncrasies in the content, format or structure of these datasets (Arora et al., 2018). Unfortunately, fine-tuning LLMs (e.g., LLaMA 70b model (Touvron et al., 2023)) on task-specific data requires a large number of GPU hours. Alternatively, training smaller, task-specific models from scratch requires a large amount of labeled data to achieve comparable performance (Tajbakhsh et al., 2016). Furthermore, fine-tuning LLMs through proprietary APIs with self-uploaded labeled training data not only requires LLM experts' knowledge but is also expensive.[1] These fine-tuned models further remain black-box, with restricted access to details regarding architectural intricacies, model weights, training data, and even prediction confidences.

In order to address these computational and accessibility challenges associated with fine-tuning, we introduce a scalable ensemble method called ***InfoSel*** (*Informed Selection*), which trains a small-sized selection model with just a few labeled task-specific samples. Unlike current ensemble methods (e.g., MetaQA (Puerto et al., 2023)), which depend on the confidence scores and thus can not be applied to black-box models like GPT3.5 text-davinci, *InfoSel* does not rely on such information and offers black-box ensembling. While traditional ensemble methods such as OLA (Woods et al., 1997) and PageRank (Brin and Page, 2012) are not adapted to task-specific particularities (e.g., different features) of different datasets. *InfoSel* incorporates *task-specific* optimization by considering variations of both the inputs and predicted answers from the ensembled LLMs/VQA models (*base models*), which allows it to be easily adapted to different datasets. Finally, our method efficiently deals with *multi-modal* inputs. We showcase the adaptability and effectiveness of the method by testing it on textual and visual question answering tasks. Concretely,

[1] `https://platform.openai.com/docs/guides/fine-tuning/`

our results exhibit superior performance on multimodal VQA inputs compared to the state-of-the-art PairRanker (Jiang et al., 2023) ensemble method that is designed to work exclusively with text. Table 1 compares our method with alternatives.

	FT	Pair-Ranker	OLA/ PageRank	***InfoSel***
Task-specific	✓	✓	✗	✓
Data-efficient	✗	✗	✓	✓
Black-box	✗	✓	✓	✓
Multimodal	✗	✗	✓	✓
Ensemble	✗	✓	✓	✓

Table 1: Our method (*InfoSel*) aims to optimize ***task-specific*** ensembling of ***black-box*** models, where confidences and parameters can not be accessed. We use only a small portion of training data (***data-efficient***). We optimize the performance of the ***ensemble*** instead of standalone fine-tuned (FT) models. Finally, our method is ***multimodal***, and applicable to VQA task.

At its core, *InfoSel* (see Figure 1) trains a small-size ensemble model to dynamically identify the most accurate base model (i.e., LLM or VQA model) for a given input, which we refer to as the *winner*. This is achieved by designing a meta-level classification task considering all the base models as labels for every input. We designed and implemented two ensemble architectures for textual and visual QA tasks. Our first proposed architecture, *InfoSel*-TT, uses a textual transformer (TT, 110M parameters) (Devlin et al., 2019) as the backbone to generate a textual representation of the question with the predicted answers by base models. Although *InfoSel*-TT is straightforward and effective, it cannot handle multimodal data. To address this, we propose a second architecture named *InfoSel*-MT, where we incorporate a multimodal transformer (MT, 115M parameters) (Li et al., 2020) to generate fused contextual representations of a multimodal input (image, question, and the predicted answers). These fused representations are used to train a dense layer for selecting the winning model. The challenge with this approach is the lack of exposure of the base models to new (unseen) labels appearing in the task-specific datasets. To address this, we fine-tune TT and MT models (FT-TT and FT-MT) separately to learn these new labels. The predictions of these fine-tuned models are fused with the output from *InfoSel* using a second, separately trained ***InfoSel**** ensemble model. We experiment with and without *InfoSel**, as this component is considered optional when *InfoSel* is already performing well.

To demonstrate the effectiveness of our method with limited resources. We selects three less costly LLMs (GPT-3.5-turbo (OpenAI, 2023), LLaMA-2-70b-chat (Touvron et al., 2023) and GPT3.5 text-davinci-003) and three local VQA models (ALBEF (Li et al., 2021a), BLIP (Li et al., 2022) and VLMo (Bao et al., 2022)). These models are used as base models to provide answers for textual and visual QA ensemble tasks respectively. For showing the *data efficiency* of the proposed architectures, we train them on a subsample of training data from public benchmark datasets and test on the corresponding full test data. Experimental results showcase improvements in the performance up to +5% on TQA tasks and +31% on VQA tasks when compared to the ensembled base models.

In summary, our contributions are: (1) *InfoSel*, a novel data-efficient approach to ensemble black-box models without relying on access to model architecture, weights or prediction confidences for optimizing on task-specific datasets; (2) Assessment of the performance on textual and multimodal visual QA tasks, demonstrating gains of up to +5% with *InfoSel* and up to +31% with *InfoSel**compared to ensembled base models on four benchmark datasets; (3) A detailed analysis of data efficiency, demonstrating that *InfoSel* surpasses the performance of the leading base models with as few as 10 training samples. [2]

2 Related Work

Domain Adaptation. These methods aim to improve the performance of a model on a task-specific domain by leveraging knowledge from other domains (Zhou et al., 2022). Methods such as fine-tuning (Yosinski et al., 2014), feature adaptation (Long et al., 2015) and data augmentation (Choi et al., 2019) aim to improve the performance of standalone models and thus typically require large amounts of labeled training data or access to the model architecture and weights.

Ensemble Learning. Ensemble methods generate and combine multiple learners (ML models) to address a particular ML task (Sagi and Rokach, 2018). Classical ensembling approaches like boosting (Schapire, 2013) and bagging (Breiman, 1996) are designed to train and combine a large number of individual models and are thus computationally

[2] Code and data are available at: `https://github.com/Yuuxii/Black-box-QA-Ensemling`.

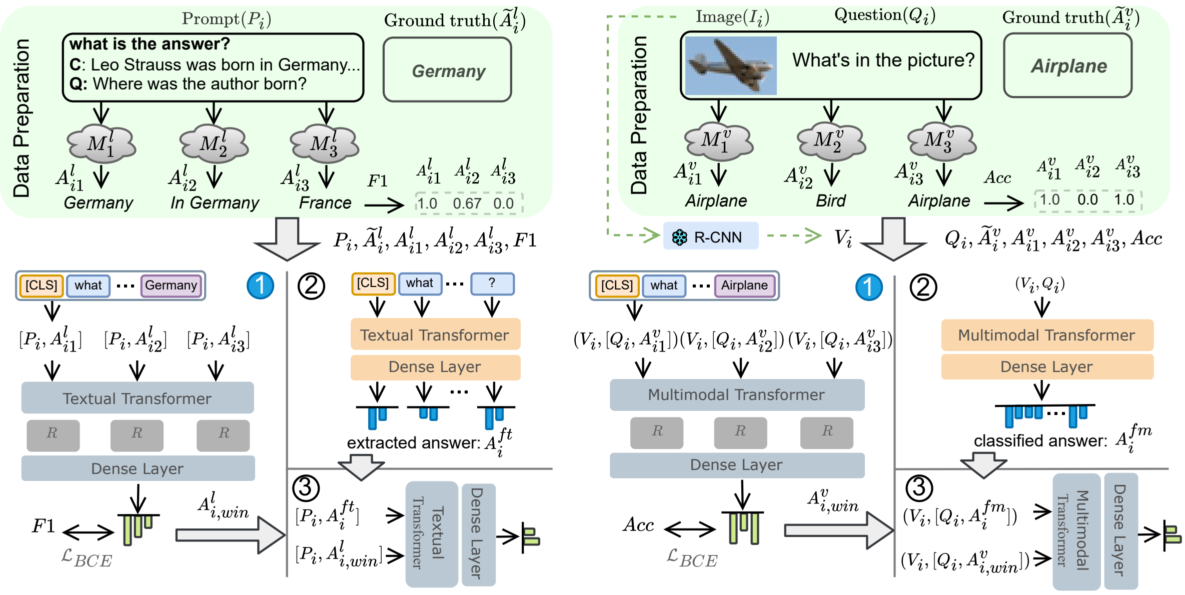

Figure 1: Architecture of our ***InfoSel*** (step ①) , FT (step ②) and ***InfoSel****** (step ③) models. M_*^l and M_*^v refer to black-box LLMs and VQA base models respectively, which are not trainable. The number of these base models is flexible and is not restricted to 3 as in the figure. The models on the left are trained for the TQA tasks, while the models on the right are trained for the VQA tasks. All our models are trained independently. Note that FT and *InfoSel** (step ② and ③) are optional and are best suited for datasets that contain a high percentage of labels that the base models have not been exposed to.

expensive. Stacking (Wolpert, 1992) uses a meta-learner to integrate the probabilities of the predictions from base models for the final output. Recent methods proposed by Jitkrittum et al. (2024); Puerto et al. (2023) require either the knowledge of confidence score or the base model's training data. Other methods train their base models to avoid dataset biases (Han et al., 2021), while Xu et al. (2020) aim to learn joint feature embeddings across different domains. However, these methods require at least one piece of knowledge that the black-box models can not provide, including base models' confidence scores or training data (not available for ChatGPT).

Multimodal Black-box Models Ensembling. Dynamic classifier selection methods, most notably OLA (Woods et al., 1997), can be applied to black-box models by ranking the best-performed local classifier dynamically in the nearest region of the input. Alternatively, majority voting (Chan and van der Schaar, 2022) and PageRank (Brin and Page, 2012) weight the predictions by their internal agreements. Yet, these methods are not designed for task-specific optimization. Pair-Ranker (Jiang et al., 2023) is task-specific but not designed for handling multimodal inputs. To address this, *InfoSel* proposes a transformer-based setup that utilizes multimodal information in the black-box setting to enhance task-specific performance.

3 InfoSel Ensemble Training

Figure 1 illustrates the proposed *InfoSel* and *InfoSel** frameworks to ensemble LLMs for TQA tasks (left), and VQA models for VQA tasks (right). We differentiate TQA components using LLMs and VQA components using VQA models by denoting them with superscripts l and v respectively. Similarly, to distinguish models used in TQA and VQA tasks, we add suffixes "-TT" and "-MT" respectively. For example, the *InfoSel*-MT model, refers to the *InfoSel* for the VQA task.

3.1 InfoSel Training for TQA

Before training *InfoSel*, we first perform the *data preparation* (top of Figure 1) for both training and testing. Next, we train *InfoSel* (step ①). Training FT models (step ②) and *InfoSel** (step ③) is optional, but can significantly improve performance for datasets that contain a high percentage of labels that the base models have not been exposed to.

Data Preparation. First, we randomly sample N content-question pairs $\{(C_i, Q_i)\}_{i=1}^N$ and the corresponding ground truth answers $\{\widetilde{A}_i^l\}_{i=1}^N$ from various benchmark datasets (refer to Section

4.1). Next, we build prompts P_i following specific prompt rules $P_i = R(C_i, Q_i)$ (refer to Section 4.1). Using these prompts instead of plain (C_i, Q_i) text improves the LLMs' answer quality (Bach et al., 2022). We select K (K=3) black-box LLMs $\{M_j^l\}_{j=1}^K$ to generate answers on the N prompts. The answer generated by M_j^l on P_i is denoted as A_{ij}^l ($A_{ij}^l = M_j^l(P_i)$). Thereby, K LLMs provide $N * K$ candidate answers for N prompts. We calculate the word-level $F1$-scores (Rajpurkar et al., 2018) of all the candidate answers $\{A_{ij}^l\}_{j=1}^K$ respectively for P_i. These $F1$-scores serve as target Y_i^l to optimize the ensemble model:

$$Y_i^l = \{F1(A_{ij}^l, \widetilde{A}_i^l)\}_{j=1}^K, Y_i^l \in \mathbb{R}^K.$$

The input for the ensemble training consists of K texts. Each text is formed by concatenating P_i with each individual answer predicted by a base model j, A_{ij}^l. More formally, the input X_i^l is:

$$X_i^l = \{[P_i, A_{ij}^l]\}_{j=1}^K, |X_i^l| = K.$$

The inputs $\{X_i^l\}_{i=1}^N$ and the corresponding target labels $\{Y_i^l\}_{i=1}^N$ are used for ensemble training.

***InfoSel*-TT.** We use a textual BERT-base (Devlin et al., 2019) transformer f_θ^t, (θ denote trainable model parameters) as the backbone of *InfoSel*-TT. To achieve faster convergence, we load the pre-trained weights of *bert-base-uncased* model. The input vector X_i^l is passed to f_θ^t to generate K sentence representations for each value in X_i^l respectively. Thus, the sentence representation R_{ij}^t of $[P_i, A_{ij}^l]$ from f_θ^t is:

$$R_{ij}^t = f_\theta^t([P_i, A_{ij}^l]), R_{ij}^t \in \mathbb{R}^{768}.$$

A dense layer (f_θ^d) is followed to classify $\{R_{ij}^t\}_{j=1}^K$, and is trained to match the target label Y_i^l using binary cross entropy loss $\mathcal{L}_{BCE}$. More formally, the training objective of *InfoSel*-TT is:

$$\min_\theta \sum_{i=1}^N \mathcal{L}_{BCE}(f_\theta^d([f_\theta^t([P_i, A_{ij}^l])]_{j=1}^K), Y_i^l).$$

Finally, the trained *InfoSel*-TT model (M^{it}) selects the winner model $M_{i,win}^l$ from $\{M_j^l\}_{j=1}^K$ for the input P_i with the highest probability score based on the selection logits produced by f_θ^d. $A_{i,win}^l$ denotes the answer provided by $M_{i,win}^l$.

FT-TT. Using only the *InfoSel*-TT model may limit performance due to the base models' lack of exposure to new (unseen) labels in the task-specific datasets. To address this, we fine-tune a separate lightweight TT model directly on the TQA datasets to learn these new labels. Specifically, the training objective is to locate the start and end token position of the answer from the context C_i. We provide the token positions of $\widetilde{A}_i^l$ as the target label, such that the model is optimized to classify each token in two classes (start/end token). This fine-tuned textual transformer model is referred to as FT-TT (M^{ft}).[3] We denote the answer predicted by M^{ft} on P_i as A_i^{ft}.

***InfoSel**-TT.** This model performs a further ensemble training of FT-TT and *InfoSel*-TT models with the same training scheme and labeled training data as *InfoSel*-TT. We anticipate that the thus trained *InfoSel**-TT model on the output of *InfoSel*-TT and the label finetuned FT-TT, will improve the ability to handle labels unseen by base models. As a result, we expect an improvement in the overall task-specific performance. The winner model selected by *InfoSel**-TT belong to $\{M^{it}, M^{ft}\}$.

3.2 InfoSel Training for VQA

Data Preparation. Given N image-question pairs $\{(I_i, Q_i)\}_{i=1}^N$ from dev data of VQA benchmark datasets, we use K (K=3) pre-trained VQA models to predict answers A_{ij}^v as follows: $\{M_j^v((I_i, Q_i)) \to A_{ij}^v\}_{j=1}^K$. We denote the ground truth answer for image-question pair (I_i, Q_i) as $\widetilde{A}_i^v$. Target labels Y_i^v for ensemble training are given by the accuracy scores of the K candidate answers evaluated on $\widetilde{A}_i^v$:

$$Y_i^v = \{Acc(A_{ij}^v, \widetilde{A}_i^v)\}_{j=1}^K, Y_i^v \in \mathbb{R}^K.$$

The concatenation of question (Q_i) with each of the candidate answers (A_{ij}^v) obtained from the base models and the corresponding image (I_i) serves the input to our ensemble model *InfoSel*-MT:

$$X_i^v = \{(I_i, [Q_i, A_{ij}^v])\}_{j=1}^K, |X_i^v| = K.$$

***InfoSel*-MT.** A Multimodal Transformer (MT, f_θ^m) (Li et al., 2021b) is employed as the backbone for *InfoSel*-MT. Specifically, we first generate visual features V_i of I_i using a pre-trained R-CNN model (M_r) (Anderson et al., 2018). V_i is composed of a vector of the image region features v_i

[3] The training scheme is adapted from https://huggingface.co/learn/nlp-course/chapter7/7?fw=pt with the additional option to allow the model to return empty answers for unanswerable questions.

and the detected *tags* (i.e., object labels of the image) (Li et al., 2021b). The concatenated question-answer pair $[Q_i, A^v_{ij}]$ and V_i is then passed together to MT (f^m_θ) to generate a fused contextual representation R^m_{ij}:

$$V_i = (v_i, tags) = M^r(I_i),$$

$$R^m_{ij} = f^m_\theta(V_i, [Q_i, A^v_{ij}]), R^m_{ij} \in \mathbb{R}^{768}.$$

Finally, we use an additional dense layer (f^d_θ) to map R^m_{ij} to the target label Y^v_i. The training is optimized using binary cross-entropy loss:

$$\min_\theta \sum_{i=1}^{N} \mathcal{L}_{BCE}(f^d_\theta([f^m_\theta(V_i, [Q_i, A^v_{ij}])]^K_{j=1}), Y^v_i).$$

We denote M^{im} to the trained *InfoSel*-MT model, M^{im} selects the winner model $M^v_{i,win}$ from $\{M^v_j\}^K_{j=1}$ to predict answer $A^v_{i,win}$ for the image-question pair (I_i, Q_i) based on the selection logits produced by f^d_θ.

FT-MT. Similar to FT-TT, FT-MT composed of a trainable MT and a Multilayer Perceptron (MLP) is fine-tuned with the same training data as *InfoSel*-MT. Differently, FT-MT solves a multi-label classification task by classifying the fused contextual representation of Q_i (instead of $[Q_i, A^v_{ij}]$ like *InfoSel*-MT) and V_i to a predefined answer list (labels). This list contains frequent answers from the training data. As a result, a trained FT-MT model (M^{fm}) can learn to predict the unseen (new) labels (answers) contained in the task-specific datasets, but not in the pre-training data of the base models. A^{fm}_i denotes the answer predicted by M^{fm} over (I_i, Q_i). The training scheme is adapted from (Li et al., 2021b).

***InfoSel**-MT.** Similar to *InfoSel**-TT, *InfoSel**-MT model ensembles the FT-MT and *InfoSel*-MT models using the same training scheme as in *InfoSel*-MT. The winner model selected by *InfoSel**-MT belong to $\{M^{im}, M^{fm}\}$.

4 Experiments and Analysis

We first introduce the setup of our experiments, followed by detailed results and analysis.

4.1 Datasets

To demonstrate the *data efficiency* of our approach, we subsampled four publicly available benchmark datasets. This resulted in four *Mini* datasets, the train set amounts to $\sim$1% of the TQA datasets' and $\sim$10% of the VQA datasets' original full size. Table 2 presents the details of these datasets.

Mini **Dataset**	**Source Dataset**	**Num.**	**%**
Mini-SDv2 train	SQuAD-V2 train	800	0.56
Mini-SDv2 val	SQuAD-V2 train	200	0.14
Mini-SDv2 test	SQuAD-V2 dev	11,873	8.39
Mini-NQ train	NQ-Open train	800	0.87
Mini-NQ val	NQ-Open train	200	0.22
Mini-NQ test	NQ-Open dev	3,499	3.83
Mini-GQA train	GQA dev	105,640	9.80
Mini-GQA val	GQA dev	26,422	2.45
Mini-GQA test	GQA test	12,578	1.17
Mini-Viz train	VizWiz dev	3,456	10.5
Mini-Viz val	VizWiz dev	863	2.63
Mini-Viz test	VizWiz test	8,000	24.39

Table 2: Details of the *Mini* datasets used for *InfoSel* ensemble training and testing. % stands for the percentage of the original full dataset.

TQA datasets. We generated two *Mini* datasets, Mini-SDv2 and Mini-NQ, consisting of 1,000 randomly sampled instances from SQuAD-V2 (Rajpurkar et al., 2018) and NQ-Open (Kwiatkowski et al., 2019) train splits, respectively. Details of the datasets are shown in Appendix A.1.

VQA datasets. Our results (Figure 2) reveal that VQA tasks demand a greater quantity of training samples compared to TQA tasks. Therefore, we constructed Mini-GQA and Mini-Viz datasets using a larger fraction (the dev data) of GQA (Hudson and Manning, 2019) and VizWiz (Gurari et al., 2018) datasets compared to TQA datasets.

4.2 Base Models

We experiment with ensembling GPT-3.5-turbo-0613 (ChatGPT), LLaMA-2-70b-chat (hereinafter referred to as "LLaMA") (Touvron et al., 2023) and GPT-3.5 text-davinci-003 (hereinafter referred to as "Davinci") to generate answers for TQA tasks.[4] To tackle VQA tasks, we employ three local VQA models (VLMo (Bao et al., 2022), ALBEF (Li et al., 2021a) and BLIP (Li et al., 2022)), which are pre-trained on VQA v2 dataset (Antol et al., 2015). Note that we use less costly and publicly accessible VQA models to save experimental costs.

Evaluation Metric. LLMs tend to generate contextual answers that lead to lower scores in the exact match (EM). Therefore, we mainly use the (per-answer) token-level F1-score from the official evaluation guidance of the datasets as the main evaluation metric for TQA performance. Our re-

[4] GPT3.5 text-davinci-003 is deprecated after our experiments, but this does not influence the effectiveness of our method.

Textual Question Answering					Visual Question Answering		
Model	**Mini-SDv2**		**Mini-NQ**		**Model**	**Mini-GQA**	**Mini-Viz**
	EM	**F1**	**EM**	**F1**		**Acc**	**Acc**
LLaMA-2-70b-chat	0.24	11.34	28.07	46.47	ALBEF	50.60	21.28
text-davinci-003	52.37	58.44	52.24	69.44	BLIP	48.08	20.80
ChatGPT	30.89	44.95	57.53	71.54	VLMo	48.21	19.77
Oracle	58.61	66.20	64.02	79.21	**Oracle**	65.03	-
MV	26.95	37.75	46.07	62.43	MV	51.05	21.47
WV	52.37	58.44	57.53	71.54	WV	52.10	19.43
PageRank	25.39	37.31	51.76	68.53	PageRank	51.47	21.66
OLA	47.90	55.59	54.70	70.05	OLA	48.65	20.32
PairRanker	57.28	63.33	57.96	72.21	PairRanker	52.05	22.42
(0-shot) LLM-Blender	4.90	21.20	1.03	25.06	(0-shot) LLM-Blender	0.0	0.0
FT-TT	46.80	47.68	36.52	40.60	FT-MT	50.48	51.76
***InfoSel*-TT (ours)**	**57.74**	**63.63**	**58.45**	**73.37**	***InfoSel*-MT (ours)**	**55.16**	23.16
***InfoSel**-TT (ours)**	49.09	49.85	48.16	53.70	***InfoSel**-MT (ours)**	52.54	**52.91**

Table 3: Test performance comparison on textual and visual QA datasets. The overall best results are highlighted in bold, and the best results of base models are underlined. The test data of Mini-Viz is not publicly accessible and thus the Oracle cannot be reported.

sults differ from the ones reported in (Laskar et al., 2023; Kocoń et al., 2023) because we do not apply any post-processing, human evaluation or output constraints for the generated answers.

Finally, we use **Oracle** to represent the maximum capability of a combination of base models. Specifically, for each input, the Oracle selects the answer with the highest agreement to the ground truth among all the answers predicted by the base models. Thus, the Oracle score represents the performance of an ideal ensemble model.

4.3 Baselines

We use Majority Voting (MV), Weighted Voting (WV) (Schick and Schütze, 2021), PageRank (Brin and Page, 2012), Overall Local Accuracy (OLA) (Woods et al., 1997), PairRanker and LLM-Blender (Jiang et al., 2023) as baselines.

Majority Voting (MV). MV makes a collective decision by considering the predicted answers as a group of individuals voting on a particular input. The answer that receives the most votes is the winner; otherwise, ties are broken randomly.

Weighted Voting (WV). We adopt a strategy similar to Schick and Schütze (2021), where the model accuracy of the train data before training is used as the weight for average weighting. In our case, we use the corresponding accuracy of the base models as the weight for voting.

PageRank (Brin and Page, 2012). We adapt PageRank as a baseline to determine the most suitable answer in a graph where all the answers to one question are connected by their BLEURT (Sellam et al., 2020) similarities.

Overall Local Accuracy (OLA) (Woods et al., 1997). Following (Cruz et al., 2018), we use the k-nearest neighbors algorithm to divide the input space (representations of prompts for TQA, representations of images and questions for VQA) of training data into 7 regions. The overall local accuracy of each base model in different regions is computed as its region competence. The model presenting the highest competence level is selected to predict the label of inputs that fall in the same region.

PairRanker and LLM-Blender (Jiang et al., 2023). The PairRanker model is trained to rank a pair of candidate predictions from two LLMs using multiple optimizing objectives (i.e., EM, F1). The logits of the LLMs produced by the trained model are sorted to get the top k (we use k=2) predictions among multiple pairwise comparison results for a GENFUSER model (Flan-T5-XL (Chung et al., 2022), 3B parameters) to generate the final fused prediction. LLM-Blender is a composite of the PairRanker and the GENFUSER model. We train PairRanker using the same textual transformer backbone as *InfoSel* for comparison, and we use 0-shot LLM-Blender models which have been trained over massive data (105k), including TQA data to test on our data.

4.4 Performance Comparison

In this section, we analyze the performance of our method, taking into account its distinctive characteristics as described in Table 1. Concretely, we focus on comparing our models in terms of *task-specific* performance, *data efficiency*, and *multi-*

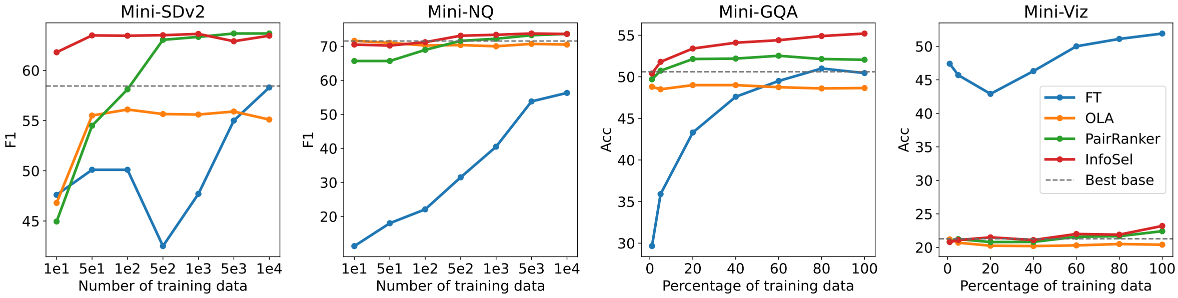

Figure 2: Test performance of *InfoSel* compared to baselines over increasing size of training data for TQA (left two figures) and VQA (right two figures). *Best base* represented the best performance of the base models.

modal capabilities.

4.4.1 Task-specific Performance

Table 3 demonstrates the *task-specific* performance of *InfoSel*, base models and ensemble baselines on textual and visual QA datasets. For TQA, we observe that LLaMA underperforms other base models. Upon closer examination, we found that LLaMA generates longer explanation text which, although often accurate, decreases the EM and F1-score values. Conversely, a more consistent performance of base models is observed for VQA. Mini-Viz contains 28% of unanswerable questions, and the label "unanswerable" has never been seen by base models. Consequently, this lack of exposure leads to significantly lower performance scores.

Baseline ensemble methods such as WV, PageRank and OLA achieve only marginal improvements compared to base models (≤+1.5%) on VQA datasets. These results highlight the limitations of these methods when applied to *task-specific* datasets (see also Table 1). PairRanker underperformed *InfoSel* even though it has been trained with the same data. 0-shot LLM-Blender tends to generate longer answers compared to PairRanker which leads to a low score especially when evaluated in the exact-match settings (EM, ACC).

For TQA tasks, *InfoSel*-TT achieves 5.19% (63.63-58.44) improvement compared to individual base models, and reaches 96.12% (63.63/66.20) of the Oracle on Mini-SDv2. Similarly, the corresponding improvement in Mini-NQ performance is 1.83%, reaching 93.06% of the performance achieved by the Oracle. In contrast, FT-TT, despite its superior performance over two base models on Mini-SDv2, underperforms *InfoSel*-TT by more than 15% due to the small-size training data (refer to Figure 2). This low performance of FT-TT models negatively impacts the performance of *InfoSel**. As a result, we conclude that while *InfoSel** can exhibit superior performance (see further for VQA tasks), it also requires more training data.

For VQA tasks, the results in Table 3 showcase an improvement of 4.56% in the accuracy score of *InfoSel*-MT compared to the base models (55.16-50.60) on Mini-GQA. Furthermore, FT-MT improves 30.48% (51.76-21.28) accuracy on Mini-Viz due to a high percentage of unseen labels (e.g., "unanswerable") introduced during fine-tuning. Finally, the superior performance of *InfoSel**-MT model on Mini-Viz dataset demonstrates the effectiveness of the proposed blending approach, which improves 31.63% (52.91-21.28) accuracy upon the *InfoSel*-MT model.

4.4.2 Data Efficiency

The experimental results shown in Figure 2 demonstrate the *data efficiency* of our method by evaluating the model's performance across varying training data sizes. We observe that *InfoSel*-TT achieves a higher F1-score compared to the best base model when trained on as few as 10 samples from Mini-SDv2. This result has been further verified with the mean F1-score of 10 test results using different seed variations for sampling the training data (shown in Figure 5 in Appendix). Conversely, the number of training samples needed to surpass the performance of base models is higher for VQA datasets: 5% (6,603 samples) for Mini-GQA and 20% (864 samples) for Mini-Viz. We hypothesize that this is due to the inherent complexity of the VQA task.

PairRanker is less data-efficient than *InfoSel* as it only achieves close performance when the training samples are more than 500 on both Mini-SDv2 and Mini-NQ. Additionally, we find that a larger training data size benefits FT-TT more than *InfoSel*-TT and OLA. For example, the F1-score of FT-TT increases $\sim$200% and $\sim$500% from 10 to 10,000 training samples on Mini-SDv2 and Mini-

NQ respectively, while *InfoSel*-TT only increases by ∼3% and ∼4%. However, FT-TT still underperforms the best base model, which suggests that fine-tuning a small-sized model requires larger training data for getting a comparable performance with LLMs or *InfoSel*. Finally, we observe that a larger training data size does not necessarily lead to improved performance for the fine-tuned FT-TT model (e.g., when increasing from 80% to 100% the training data size on Mini-GQA). In contrast, OLA does not benefit as much as *InfoSel* and FT from a larger size of training data, only outperforming *InfoSel*-TT on Mini-NQ when 10 and 20 training samples are used.

4.4.3 Multimodal Data

InfoSel is able to utilize multimodal data (image and text) for VQA tasks, and thus outperform the latest text-exclusive LLM ensemble methods (PairRanker and LLM-Blender) as evidenced in Table 3. In contrast, PairRanker and LLM-Blender cannot process image features, thereby lacking crucial information in the multimodal setting, leading to a low accuracy on VQA datasets.[5] Further insights into the significance of modality information are elaborated in Section A.8 and Table 4.

4.4.4 Lightweight Model

Table 7 (Appendix) reports the parameter size of the base models and their ensemble models (*InfoSel*-TT and *InfoSel*-MT). *InfoSel* provides an efficient method for ensembling large LLMs such as ChatGPT (175B parameters) using only 110M trainable parameters. Even though only 37% ((182M-115M)/182M) trainable parameters are saved for the VQA task, we still demonstrate that *InfoSel* can effectively enhance the task-specific performance of small-size black-box VQA models, offering a lightweight solution.

4.5 Ablation Studies

Is *InfoSel* robust to the base models' individual performances? We carry out this study to assess whether *InfoSel* can effectively utilize the predictions obtained from various base models, regardless of their individual performance levels. In Figure 3, we observe a minor F1-score difference (0.15%) on the Mini-SDv2 dataset between the *InfoSel* model ensembled with and without the lowest performing base model (LLaMA). This finding suggests that *InfoSel* is robust, and not significantly affected by the individual model's performance. In a more detailed analysis, we observe that *InfoSel* selects 4% of answers from LLaMA, resulting in an overall gain of +0.28% of the F1-score. This observation highlights the effectiveness of *InfoSel*, as it can leverage the knowledge contained in the answers provided even by the lowest performing base model to some extent.

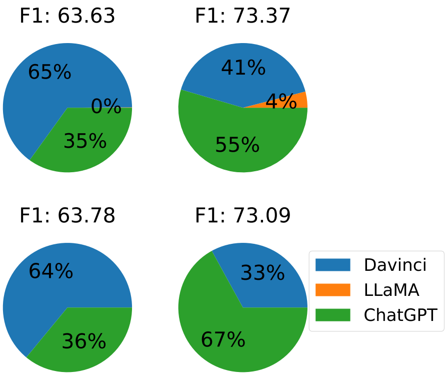

Figure 3: The portions of answers selected from different base models by *InfoSel* models on Mini-SDv2 (right column) and Mini-NQ (left column). The upper row represents the results of the *InfoSel* model ensembled with all three LLMs, and the model in the lower row excluded the worst base model (LLaMA).

Which modality information helps the most for ensembling? In Table 4, we compare the effect of providing different modality information to *InfoSel*-MT during ensemble training. Notice that even with just the question and answer (Q+A) information, our model surpasses the performance of the PairRanker. The setting that yields the lowest accuracy solely utilizes the image (V) as the signal. This can be explained by the fact that a single image often corresponds to multiple questions in VQA datasets, making it challenging for the model to acquire discriminative features. Furthermore, we conclude that the superior performance of our model when utilizing image, question, and answer (V+Q+A) data demonstrates the effectiveness of our model in *multimodal* setting.

4.6 Case Study

The first case of Table 5 indicates that *InfoSel* captures the only correct answer ("toaster") from base models. The second case demonstrates the ability of *InfoSel** to recognize a task-specific label (i.e., "unanswerable") introduced by FT-MT. *InfoSel*-MT

[5] The most frequent answer of LLM-Blender on VQA datasets is *"I'm sorry, I don't have enough context to answer that question."*

Model	Mini-GQA	Mini-Viz
PairRanker(Q+A)	52.05	22.42
InfoSel-MT(V)	50.56	20.79
InfoSel-MT(Q)	51.11	21.21
InfoSel-MT(V+Q)	50.83	20.06
InfoSel-MT(V+A)	52.38	22.66
InfoSel-MT(Q+A)	54.76	22.89
***InfoSel*-MT(V+Q+A)**	**55.16**	**23.16**

Table 4: Accuracy of ***InfoSel***-MT models when using different input information for training compared to baseline models. V, Q, and A represent visual, question, and answer information respectively.

	Mini-GQA	Mini-Viz
Image: **Question:**	 What appliance is it?	 What is this product?
ALBEF	blender	refrigerator
BLIP	**toaster**	toilet
VLMo	microwave	door
FT-MT	coffee maker	**unanswerable**
***InfoSel*-MT**	**toaster**	toilet
***InfoSel**-MT**	coffee maker	**unanswerable**

Table 5: Case study of our models on Mini-GQA test and Mini-Viz validation data. Ground truth answers are bolded.

struggles with such labels as they are unseen to the base models. This showcases the benefits of training *InfoSel** models on datasets containing a high percentage of task-specific unseen labels. The corresponding case study of TQA is shown in Table 9 (Appendix).

5 Conclusion

In this paper, we propose *InfoSel*, a novel lightweight and task-specific ensemble method designed to learn the dynamic selection of the optimal model from a range of distinct black-box base LLMs. We find that using only 110M trainable parameters, our method is able to substantially increase the performance upon the best performing base LLM. Additionally, our analysis reveals that *InfoSel* remains robust regardless the incorrect predictions of the lowest performing LLM. Our findings also show that our solution is highly data-efficient. Concretely, it requires only a fraction of instances (as few as 10) from the training set to outperform base LLMs. Finally, our experimental results reveal the ability of *InfoSel* to be adapted to multimodal setting, showing a substantial increase in performance compared to state-of-the-art alternatives.

6 Limitations

InfoSel offers an effective approach to enhancing out-domain black-box model performance and addressing answer selection. However, it is important to acknowledge certain limitations that come with its application:

Dependency on Annotated Data: *InfoSel*, like many machine learning techniques, relies on a small amount of annotated training and development data specific to the new domain. While this requirement is relatively modest, and *InfoSel*'s strength is it's data efficiency (as demonstrated in the experiments), this may still pose a limitation in scenarios where obtaining such data is challenging or costly.

Limited Applicability to Open-Ended Text Generation: *InfoSel*'s primary strength lies in its ability to select the best answer from a set of base models, making it particularly valuable in question-answering scenarios. However, for more open-ended text-generation tasks, where it may be beneficial to combine multiple answers, *InfoSel*'s single-answer selection mechanism may not be the ideal choice, and future research directions may include approaches for combining several long-form answers.

API Fine-Tuning Availability: At the time of this study, *InfoSel* operates based on the assumption that many APIs do not offer the ability to fine-tune models, which is a constraint driven by the current landscape of AI services. However, since the field of AI is rapidly evolving, API providers may potentially introduce fine-tuning as a standard feature in the future. However, our experiments show that selection may still help even when one (and potentially more) of the answer models are fine-tuned.

Transparency and Explainability: *InfoSel*, like other machine learning models, which selects answers from black-box models may itself operate as a "black box". This means its decision-making process might not be readily interpretable or explainable to end-users. Pairing *InfoSel* with explainability techniques may give users a clearer understanding of how the model makes its selections.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. 2018. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24:123–140.

Sergey Brin and Lawrence Page. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput. Networks*, 56:3825–3833.

Alex Chan and Mihaela van der Schaar. 2022. Synthetic model combination: an instance-wise approach to unsupervised ensemble learning. *Advances in Neural Information Processing Systems*, 35:27797–27809.

Jaehoon Choi, Taekyung Kim, and Changick Kim. 2019. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Journal of Machine Learning Research 25 (2024) 1-53*.

Rafael MO Cruz, Robert Sabourin, and George DC Cavalcanti. 2018. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. 2024. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021b. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online. Association for Computational Linguistics.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France. PMLR.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. Chatgpt. `https://chat.openai.com/`. Mar 14 version.

OpenAI. 2024. Hello gpt-4o. Accessed: 2025-05-01.

Haritz Puerto, Gözde Şahin, and Iryna Gurevych. 2023. MetaQA: Combining expert agents for multi-skill question answering. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3566–3580, Dubrovnik, Croatia. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Robert E Schapire. 2013. Explaining adaboost. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 37–52. Springer.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 255–269, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Kevin Woods, W. Philip Kegelmeyer, and Kevin Bowyer. 1997. Combination of multiple classifiers using local accuracy estimates. *IEEE transactions on pattern analysis and machine intelligence*, 19(4):405–410.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2020. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376, Online. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie, J. Zhao, Q. Li, Y. Chen, L. Zhang, L. Shen, and B. Dong. 2024. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Medicine Plus*, 1(2):100030.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Appendix

A.1 TQA Datasets

SQuAD-V2 (Rajpurkar et al., 2018) stands for Stanford Question Answering Dataset 2.0, a dataset designed for the task of question answering. It is an extension of the original SQuAD dataset by including over 50,000 unanswerable questions written adversarially by crowdworkers. The dataset is widely used in natural language understanding research.

NQ-Open (Kwiatkowski et al., 2019) is derived from Natural Questions and serves as an open-domain question-answering evaluation. The entirety of the questions can be addressed using the information found in the English Wikipedia. It was created for research purposes.

For Mini-NQ, we followed (Fisch et al., 2019) to use long answers as the context, and short answers as the ground truth answers. The 1,000 samples are divided into train and validation data using an 8:2 ratio, while the trained models are tested on the dev data of the original datasets due to the unavailability of original test data. We use the setup proposed in (Laskar et al., 2023) to generate the answers from LLMs.

Note that the answers of LLMs can be greatly influenced by some factors such as the use of different prompts or temperatures. However, our study does not focus on prompt engineering but rather on selecting the optimal base model to generate an answer. We will publicly release our prompts as well as the answers from LLMs for reproducibility.

A.2 VQA Datasets

GQA (Hudson and Manning, 2019) is a large-scale dataset for visual reasoning and compositional question answering research. The dataset contains over 113k images collected from a diverse set of sources and over 22 million questions. Only one ground-truth answer is provided for each image-question pair.

VizWiz (Gurari et al., 2018) is a benchmark dataset for visual question answering. It includes

Dataset	Sample Prompts
Mini-SDv2	What is the answer? Context:[context]; Question:[question]; If you can't find the answer, please respond "unanswerable". Answer:
	Answer the question depending on the context. Context: [context]; Question: [question]; If you can't find the answer, please respond "unanswerable". Answer:

Mini-NQ	Answer the question depending on the context without explanation.
	Context: [context];
	Question: [question];
	Answer:

Table 6: Our sample prompts in QA datasets. SQuAD-V2 were available in PromptSource (Bach et al., 2022) for prompt generation, we selected the prompt from PromptSource for Mini-SDv2, which contains two forms of prompts.

31K images, 250K questions, and answers collected through a mobile app for visually impaired users. 10 ground-truth answers are provided for each image-question pair.

Additionally, we compare the label differences of the pre-trained dataset (VQA v2 (Antol et al., 2015)) of VQA base models with task-specific datasets (GQA, VizWiz) for ensemble training. Figure 4 shows the top 7 most frequent answers and their percentages of GQA, VQA v2 and VizWiz. Four answers in GQA do not appear in the top list of VQA v2 and three for VizWiz (including the top frequent answer "unanswerable"). We also sample 3k most frequent answers from each dataset and calculate their percentage of overlapping, which is reported on the intersection in the figure. GQA and VizWiz have 32.9 % and 21.6% of overlap with VQA v2 respectively, showcasing significant differences between the pre-trained dataset and task-specific datasets.

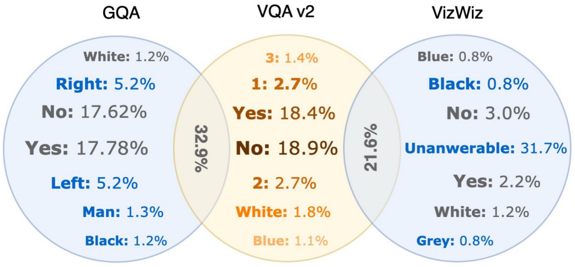

Figure 4: Top 7 most frequent answers of VQA v2 (pre-trained dataset of VQA base models), GQA and VizWiz (task-specific datasets for ensemble training). This explains why *InfoSel* performs poorly in Mini-Viz, as the new label "unanswerable" is the top frequent answer in VizWiz, but this new label has been exposed to base models that are pre-trained on VQA v2.

A.3 Base Models

LLMs		VQA Models	
Model	**#Param**	**Model**	**#Param**
LLaMA-2-70b-chat	70B	ALBEF	290M
text-davinci-003	175B	BLIP	361M
ChatGPT	175B	VLMo	182M
PairRanker	110M	PairRanker	110M
InfoSel-TT	110M	*InfoSel*-MT	115M
*InfoSel**-TT	110M	*InfoSel**-MT	115M

Table 7: Parameter size of the models used in our experiments.

ChatGPT is a large language model with 175B parameters, it allows you to have human-like conversations and much more with the chatbot. ChatGPT can generate context-based responses to user prompts (questions). However, this model is currently only accessible by cloud APIs.

GPT 3.5 text-davinci-003 is similar to ChatGPT but designed specifically for instruction-following tasks which enables it to respond concisely and more accurately. This model is deprecated after our experiments.

LLaMA-2-70b-chat (Touvron et al., 2023) is a large language model with 70B parameters. It is fine-tuned on LLaMA 2 with publicly available instruction datasets and over 1 million human annotations, while Llama 2 models are trained on 2 trillion tokens from publicly available online data sources. LLaMA 2 models are currently publicly accessible.

ALBEF (Li et al., 2021a)[6] first encodes the image and text with an image encoder (visual trans-

[6] https://github.com/salesforce/ALBEF

former (Dosovitskiy et al., 2021)) and a text encoder respectively. Then a multimodal encoder is used to fuse the image features with the text features through cross-modal attention. The V&L representation is trained with objectives of image-text contrastive learning, masked language modeling and image-text matching. Different from U-VisualBERT, ALBEF uses a 6-layer transformer decoder to generate answers for VQA tasks.

BLIP (**Li et al., 2022**)[7] uses a visual transformer as the image encoder, and a multi-task model (multimodal mixture of encoder-decoder) as a unified model with both understanding and generation capabilities. The model is jointly pre-trained with three vision-language objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling. Similarly to ALBEF, the VQA task is considered an answer generation task in this method.

VLMo (**Bao et al., 2022**)[8] is a unified vision-language pre-training method with Mixture-of-Modality-Experts. VLMO leverages large-scale image and text data to learn joint representations of vision and language. It employs a mixture model to capture diverse interactions between visual and textual information, achieving state-of-the-art performance on various vision-language tasks.

A.4 Experiment Setup

We fixed the batch size to the upper limit of the server capacity, while the learning rates and epochs are selected after a grid search on a set of values (learning rates: {e3, 5e4, e4, 5e5, e5, 5e6, e6}, epochs: {3, 5, 10, 15, 20}). Models for TQA are trained for 5 epochs using a learning rate of 5×10^{-5} and batch size of 4. Models for VQA use the same learning rate but a batch size of 16 for 20 epochs. We spent ∼74 and ∼290 seconds training 1 epoch on 1,000 samples for TQA and 4,319 samples for VQA respectively. The training was performed on 1 GPU with 16GB memory of a DGX1 server ((Pascal) Tesla P100).

A.5 Multi-modal Information Concatenation or Fusion?

We studied the impact of concatenating and fusing multi-modal input information for the VQA task. *InfoSel*-MLP is an alternative model type for *InfoSel* which processes all the input information separately with a simple Multilayer perceptron (MLP) instead of MT. A pre-trained Sentence-BERT (Reimers and Gurevych, 2019) [9] M_{qa} is used for generating question embedding R^q and answer embeddings R^a.

Model	Mini-GQA	Mini-Viz
	ACC	
InfoSel-MLP	52.35	21.12
***InfoSel*-MT**	**55.16**	**23.16**

Table 8: Comparison of using different architecture for processing input information differently. Input concatenation result is demonstrated by *InfoSel*-MLP and the fusion result is shown by *InfoSel*-MT.

$$R_i^q = M^{qa}(Q_i), R^q \in \mathbb{R}^{768}$$
$$R_{ij}^a = M^{qa}(A_{ij}^v), R_{ij}^a \in \mathbb{R}^{768}$$

MLP takes the concatenated representation of question, answer, and visual embeddings R_i^v as input and maps it to the label space. The objective function of *InfoSel*-MLP is formalized as:

$$\min_\theta \sum_{i=1}^{N} \mathcal{L}_{BCE}(MLP_\theta(\{R_i^v, [R_i^q, R_{ij}^a]\}_{j=1}^{K}), Y_i^v) \quad (1)$$

The input layer of the MLP maps the concatenated representations to a hidden layer with a size equal to 300, followed by a ReLU activation layer and then an output layer with an output size equal to the number of models.

Table 8 demonstrates the performance of the input concatenation result (*InfoSel*-MLP) and fusion result (*InfoSel*-MT). We observe that *InfoSel*-MT achieves ∼3% and ∼2% higher accuracy than *InfoSel*-MLP in Mini-GQA and Mini-Viz respectively, which proves that a fused contextual representation of inputs provides more discriminative information than a concatenation of input embeddings.

A.6 Case Study for TQA

Table 9 illustrates two insightful cases from the predictions of different models on textual Mini-SDv2 and Mini-NQ QA datasets. The first case showcases the ability of *InfoSel*-TT to select the right model (Davinci) when the rest of the models is incorrect. However, *InfoSel**-TT selects the wrong answers from the FT-TT model and underperforms

[7] https://github.com/salesforce/BLIP

[8] https://github.com/microsoft/unilm/tree/master/vlmo

[9] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

	Mini-SDv2	**Mini-NQ**
Context:	...Derrick Norman Lehmer's list of primes up to 10,006,721...	...in 2005 and the release of her eponymous debut album the following year...
Question:	How many primes were included in Derrick Norman Lehmer's list of prime numbers?	When did Taylor Swift 's first album release?
LLaMA	unanswerable	2006
Davinci	**10,006,721**	2006
ChatGPT	unanswerable	2006
FT-TT	unanswerable	**2005**
***InfoSel*-TT**	**10,006,721**	2006
***InfoSel**-TT**	unanswerable	**2005**

Table 9: Case study of our models on Mini-SDv2 and Mini-NQ test data. Answers of LLMs are shortened to keywords for better demonstration. Ground truth answers are bolded, and one incorrect ground truth answer is colored red.

InfoSel-TT. The second case illustrates the ability of LLMs to generate correct answer ("2006") despite the ground truth annotation error ("2005"). This demonstrates the advantage of ensembling highly expressive LLMs instead of relying only on fine-tuning small-size models such as FT-TT.

A.7 Robustness Analysis.

We analyze the robustness of the data efficiency property of *InfoSel* by training *InfoSel*-TT with 10 different sets of randomly sampled train data. Figure 5 demonstrates the mean and standard deviation of these 10 sets of results on Mini-SDv2 and Mini-NQ. We observe that the deviation decreases as the number of training data increases, and the deviation is less than 0.5 when using 1,000 training samples (0.29 for Mini-SDv2, 0.40 for Mini-NQ). The mean value (61.12%) when using 10 samples from Mini-SDv2 is still better than the F1-score of the best base model, while the mean (71.60%) when using 100 samples from Mini-NQ is already higher than the F1-score of the best base model which is fewer samples than the result reported in the main paper (500 samples).

A.8 Ablation Studies

Figure 6 shows the results when removing ChatGPT and Davinci models respectively. It is not recommended to use *InfoSel* to ensemble only two base models where one of the base models performs significantly worse than the other one.

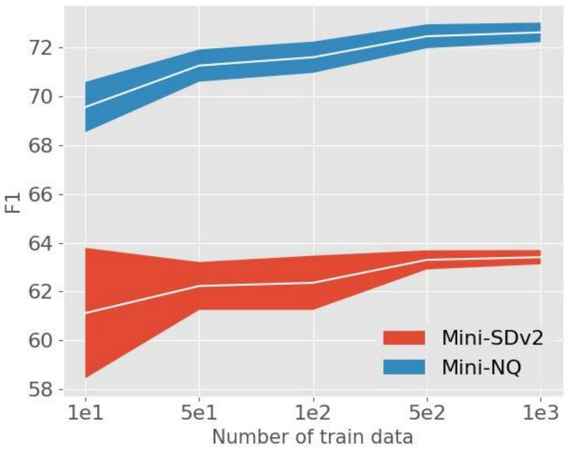

Figure 5: Mean and standard deviation of results on Mini-SDv2 and Mini-NQ when training *InfoSel*-TT with 10 sets of randomly sampled training data.

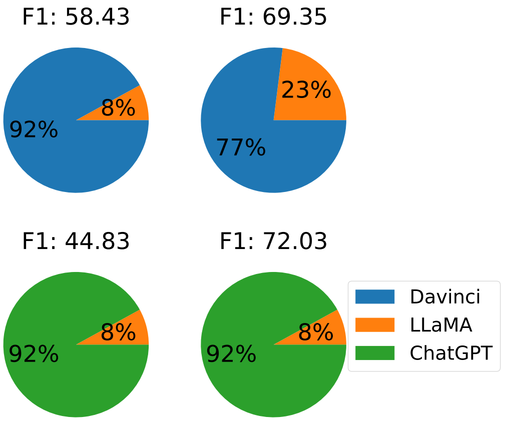

Figure 6: The portions of answers selected from different base models by *InfoSel* models on Mini-SDv2 and Mini-NQ test data. The upper row represents the results of the *InfoSel* model ensembled with Davinci and LLaMA, and the model in the lower row ensembled with ChatGPT and LLaMA.