What to do if language models disagree? Black-box model ensembling for textual and visual question answering

Anonymous ACL submission

Abstract

A diverse range of large language models (LLMs), e.g., ChatGPT, and visual question answering (VQA) models, e.g., BLIP, has been developed for addressing text and visual question answering tasks. However, both LLMs and VQA models encounter challenges when applied to out-domain datasets. Fine-tuning these models for domain adaptation is either impossible (only accessible by APIs as black-box models) or computationally expensive (big model size), and often only limited labeled out-domain data is available. Under these constraints, ensemble techniques provide a compelling alternative. In this paper, we aim to improve out-domain model performance by utilizing the capabilities of existing black-box models with limited computational cost and labeled data. To address this challenge, we introduce a novel data-efficient ensemble method, ***InfoSel***, which trains small-size (<120M parameters) ensemble models to select the best answers without relying on prediction confidences for both text and visual question answering tasks. Our results demonstrate that ***InfoSel*** improves the performance compared to the ensembled base models over four mini datasets sampled from SQuAD-V2, NQ-Open, GQA and VizWiz.

1 Introduction

Large language models (LLMs) have demonstrated remarkable proficiency across a wide range of tasks, predominantly attributed to their ability to comprehend instructions and tap into vast repositories of high-quality data (Bubeck et al., 2023; Laskar et al., 2023). A representative model – ChatGPT[1] finds extensive utilization in daily question answering (QA) tasks, rendering substantial convenience to a myriad of users (Malik et al., 2023). 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, Laskar et al. (2023); Kocoń et al. (2023) evaluate state-of-the-art LLMs and conclude that ChatGPT solves various tasks to some degree but consistently falls short of state-of-the-art performance, highlighting its limitations to specific datasets. Similarly, the same issue applies to VQA models (Li et al., 2022, 2021a,b; Bao et al., 2022). These models, when trained on in-domain data and tasks, can encounter challenges in generalizing to out-domain data due to variations in format or structure (Arora et al., 2018). Unfortunately, fine-tuning on out-domain data is not an option, as ChatGPT[2] and its similar models (e.g., GPT-3.5 text-davinci-003[3]) are proprietary and only accessible via APIs (black-box models) to users, thereby limiting our access to detailed insights regarding their architectural intricacies, model weights, training data and even prediction confidences (Jiang et al., 2023). Besides, even though few models such as LLaMA-2-70b-chat (Touvron et al., 2023) are recently accessible through online platforms[4], it is computationally expensive to fine-tune due to its large model size (70B parameters).

In the context of possessing limited computational resources and labeled data, a reliable and robust strategy for maximizing the utility of existing black-box models is to obtain predictions from multiple models and subsequently ensemble the predictions (Dietterich, 2000). Figure 1 demonstrates our motivation for developing an ensemble method to help users select the best answers from all the answers generated by different black-box models. However, standard ensemble methods like stacking, weighted averaging (Sagi and Rokach, 2018), or recent LLM-Blender (Jiang et al., 2023)

[1] https://chat.openai.com/

[2] https://chat.openai.com/

[3] https://platform.openai.com/docs/introduction

[4] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

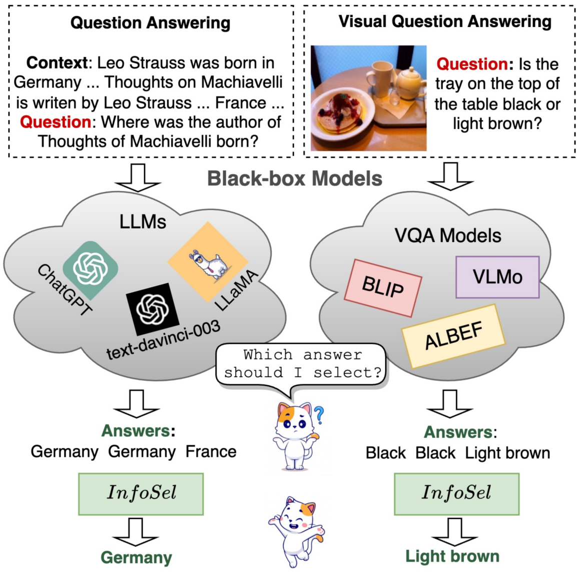


Figure 1: ***InfoSel*** learns to select the best answer from the predicted answers of black-box models for new domain datasets.

are not applicable in this case, since they either require to train their own base models independently (have access to the model architecture) or demand prediction scores and thus do not fulfill the black-box setting (where only the predicted answer is available). Majority voting, on the other hand, is applicable but provides limited performance improvement (Chan and van der Schaar, 2022).

To address the limitations of previous methods, we propose our new ensemble method named ***InfoSel*** (*Informed Selection*), a sample-level approach that trains an ensemble model to select the best answer regarding different input samples with a limited computational cost and labeled data in a black-box setting. Specifically, the ensemble model learns to solve a multiple choice text or visual QA task by considering all the predicted answers as choices and performing it as a classification task. Three LLMs (ChatGPT, LLaMA-2-70b-chat and GPT3.5 text-davinci-003) and three VQA models (ALBEF (Li et al., 2021a), BLIP (Li et al., 2022) and VLMo (Bao et al., 2022)) are used as ensemble base models to provide answers for text and visual QA task respectively.

To simulate a realistic application scenario, we sample limited labeled data from public datasets for (out-domain) training and/or ensembling, and test the ensemble of (pre-trained, black-box, in-domain) models on the corresponding (out-domain) test dataset. We refer to this setting with limited labeled data in the out-domain as "mini-*". For text QA task, we created mini-SDv2 and mini-NQ by randomly sampling 1k samples from SQuAD v2 (Rajpurkar et al., 2018) and NQ-Open (Kwiatkowski et al., 2019) train dataset respectively; mini-GQA and mini-Viz for VQA task contain only the development dataset of GQA (Hudson and Manning, 2019) and VizWiz (Gurari et al., 2018)).

Specifically, two different architectures are applied for text and visual QA tasks respectively. ***InfoSel***-BERT simply uses BERT-Base (110M parameters) (Devlin et al., 2019) as the backbone to process the question with predicted answers as a multiple choice textual QA task. Differently, ***InfoSel***-MT employs a multimodal transformer (MT) (115M parameters) (Li et al., 2019) to create fused contextual representations of input data (image, question, and the predicted answers). The fused representations are then used to train a dense layer for selecting the best answer. To address the limitation of the max capability of base models, we introduce ***InfoSel***$^+$, which further ensemble the trained ***InfoSel*** model with a fine-tuned model using BERT or MT with the same amount of labeled data.

Our results demonstrate that ***InfoSel*** and ***InfoSel***$^+$ improve the performance in mini-SDv2 (58.44% to 63.71%) and mini-NQ (71.54% to 73.37%) for textual QA task, and also mini-GQA (50.60% to 55.16%) and mini-Viz (21.28% to 52.91%) for VQA task compared to the ensembled base models.

Our contributions are: (1) We propose, ***InfoSel***, a new approach to ensembling black-box question answering models. Our approach is the first that does not rely on access to model architecture, weights or prediction confidences. ***InfoSel*** is lightweight in parameters and data-efficient. (2) We study ***InfoSel*** in textual and viual question answering and demonstrate its effectiveness on four benchmark datasets; (3) Analysis shows that on some datasets ***InfoSel*** already achieves better performance than the best of the base models with only as little as 10 samples; (4) We investigate the impact of selecting different modality of input information for ensemble training in the VQA task.

2 Related Work

Domain adaptation methods aim to improve the performance of a model on a target domain by leveraging knowledge from a source domain (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 individual models and thus require access to the model architecture, weights, or in-domain training data.

Ensemble learning entails the generation and combination of multiple learners (ML models) to address a particular machine learning 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 with numerous high-quality training data and are thus computationally expensive. Snapshot ensemble method (Huang et al., 2017) uses several local minima from one single model for ensembling, which requires full access to model weights and architecture. Stacking methods (Wolpert, 1992; Pascanu et al., 2014) uses a meta-learner to learn the predictions from base models and provides the final output. However, the predictions usually consist of probability scores generated by base models.

LLMs/VQA models ensembling methods proposed by Jiang et al. (2023) uses a PairRanker to rank the best top K answers generated by LLMs. (Puerto et al., 2021) introduces MetaQA to combine models from different domains. (Han et al., 2021) and (Clark et al., 2019) aim to avoid dataset biases, while (Xu et al., 2019) learns joint feature embeddings across different domains. However, these methods either rely on the model's prediction confidences or have access to in-domain training data and model architecture.

Ensembling black-box models can be achieved by majority voting which selects a final answer with the most votes, but it can only provide limited improvement in performance (Chan and van der Schaar, 2022). To address the limitation of the above methods, ***InfoSel*** provides a computation- and data-efficient ensemble solution for black-box models without relying on knowledge of model architecture, weights, in-domain training data, and model prediction confidences.

3 Informed Selection Ensemble Training

The left half of Figure 2 illustrates the ***InfoSel*** framework for ensemble LLMs in QA tasks, while the right half presents the ensemble framework for VQA models in VQA tasks.

3.1 *InfoSel* Training for Textual QA

Data Preparation. We randomly sampled N (N=10^3) content-question pairs $\{(C_i, Q_i)\}_{i=1}^{N}$ from training data of different benchmark datasets. (C_i, Q_i) is then formed as a prompt with certain rules (explained in Table 8 in Appendix) $P_i = R(C_i, Q_i)$ for getting high-quality answers from LLMs. $\widetilde{A}_i^l$ denotes the ground-truth answer of P_i. [5] K (K=3) state-of-the-art black-box LLMs $\{M_j^l(P_i) \rightarrow A_{ij}^l\}_{j=1}^{K}$ are chosen to predict on the N prompts, and thereby provide $N * K$ candidate answers. We calculate the token-based $F1$ scores (Rajpurkar et al., 2018) of all candidate answers $\{A_{ij}^l\}_{j=1}^{K}$ predicted on P_i and use it as the target label Y_i^l for training answer-selection.

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

P_i is later concatenated with $\{A_{ij}^l\}_{j=1}^{K}$ respectively as input $\{X_{ij}^l\}_{j=1}^{K}$ for ensemble training, while Y_i^l contribute as label for the training optimization. We denote the concatenation of vectors or strings by the notation $[\cdot, \cdot]$.

$$X_{ij}^l = [P_i, A_{ij}^l]$$

***InfoSel*-BERT.** BERT-Base is used as the backbone of *InfoSel*-BERT to generate K sentence representations $\{h_{ij}^x\}_{j=1}^{K}$ of $\{X_{ij}^l\}_{j=1}^{K}$ respectively.

$$h_{ij}^x = BERT(X_{ij}^l), h_{ij}^x \in \mathbb{R}^{768}$$

A dense layer (DL) is followed to classify $\{h_{ij}^x\}_{j=1}^{K}$ to label Y_i^l with a binary cross entropy loss BCE. We denote θ to be the set of trainable parameters and formulate the training objective of *InfoSel*-BERT as:

$$min_\theta \sum_{i=1}^{N} BCE(DL_\theta([BERT_\theta(X_{ij}^l)]_{j=1}^{K}), Y_i^l) \tag{1}$$

FT-BERT. Motivated by a situation where a tunable QA model (BERT-Base) and limited labeled data are available, we fine-tune the BERT Base model with the same amount of the labeled data (10^3) and name the fine-tuned model as FT-BERT. In particular, FT-BERT aims to locate the start and end token position of the answer from the context C. Therefore, the start token and end token

[5] We distinguish components in textual QA with **l**anguage models and **v**isual QA with superscripts l and v.

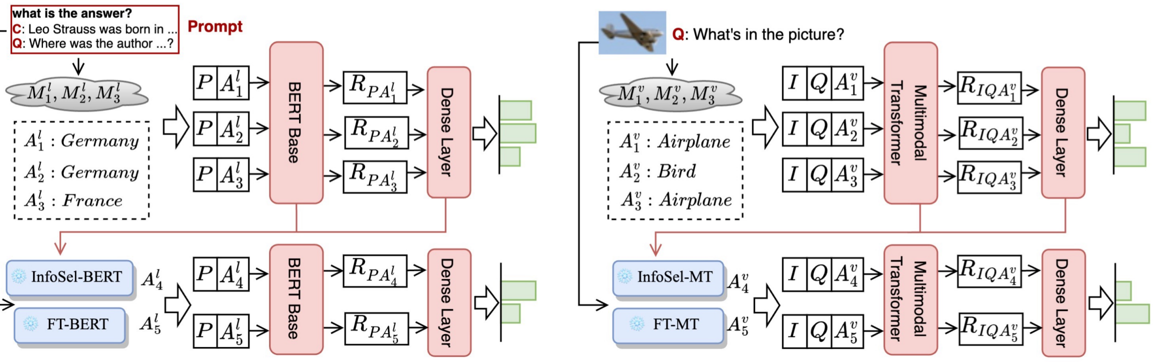


Figure 2: ***InfoSel*** framework. Trainable models are in red color, while blue represents the frozen models.

position of $\widetilde{A}_i^l$ is provided as the label for token classification optimization.[6]

***InfoSel*$^+$-BERT.** To address the limitation of the max capability of base models, *InfoSel*$^+$-BERT performed a further ensemble training of FT-BERT and *InfoSel*-BERT with the same training scheme as *InfoSel*-BERT. We expect *InfoSel*$^+$-BERT can capture the unseen labels of base models from FT-BERT and thus improve the overall performance.

3.2 *InfoSel* Training for VQA

Data Preparation. Assume we have N image-question pairs $\{(I_i, Q_i)\}_{i=1}^N$ from development data of VQA benchmark datasets. K (K=3) pre-trained VQA models $\{M_j^v((I_i, Q_i)) \rightarrow A_{ij}^v\}_{j=1}^K$ learned to predict $N * K$ candidate answers over $\{(I_i, Q_i)\}_{i=1}^N$. $\widetilde{A}_i^v$ is the ground-truth answer of (I_i, Q_i). A binary vector, i.e., label Y_i^v, is then constructed by the accuracy scores of the K candidate answers.

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

A concatenation of question and answer denotes as text segment $T_{ij} = [Q_i, A_{ij}^l]$. Text embeddings h_{ij}^t are generated by the BERT embedding layer, which means each subword embedding is the sum of its token, position, and segment embedding.

$$h_{ij}^t = embedding(T_{ij}), h_{ij}^t \in \mathbb{R}^{768}$$

Visual embeddings h_i^v generated by a pre-trained R-CNN model (Anderson et al., 2018) include the image region embeddings h_i^I and the detector tag (i.e., object labels of the image) embeddings h_i^{tag}. Each region embedding is the sum of a visual feature vector from the detector and a spatial box coordinate embedding (Tan and Bansal, 2019; Li et al., 2021b). We linearly map the size of h_i^I from 2048 to 768 for concatenation with tag embeddings.

$$h_i^I, tags = RCNN(I_i), h_i^I \in \mathbb{R}^{2048}$$

$$h_i^{tag} = embedding(tags), h_i^{tag} \in \mathbb{R}^{768}$$

$$h_i^v = [Linear(h_i^I), h_i^{tag}], h_i^v \in \mathbb{R}^{768}$$

In summary, The text and visual embeddings $\{(h_{ij}^t, h_i^v)\}_{j=1}^K$ are served as inputs for ensemble training, while Y_i^v is used as the label for training optimization.

***InfoSel*-MT.** A Multimodal Transformer (MT) (Li et al., 2021b) is employed as the backbone for *InfoSel*-MT to generate a fused contextual representation h_{ij}^c of (h_{ij}^t, h_i^v). Finally, a dense layer (DL) is followed for the classification by mapping $\{h_{ij}^c\}_{j=1}^K$ to label Y_i^v.

$$h_{ij}^c = MT(h_{ij}^t, h_i^v), h_{ij}^c \in \mathbb{R}^{768}$$

The training objective function can be formalized as follows:

$$min_\theta \sum_{i=1}^{N} BCE(DL_\theta([MT_\theta(h_{ij}^t, h_i^v)]_{j=1}^K), Y_i^v) \quad (2)$$

FT-MT. Similar to FT-BERT, the trainable MT in this framework is also fine-tuned with the development dataset as a VQA model which is able to predict answers. Different from *InfoSel*-MT, the input only contains the question embedding h_i^q (instead of text embeddings) and visual embedding h_i^v.

$$h_i^q = embedding(Q_i), h_i^q \in \mathbb{R}^{768}$$

[6]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.

Dataset	Source Dataset	Num.
mini-SDv2 train	SQuAD-V2 train	800
mini-SDv2 validation	SQuAD-V2 train	200
mini-SDv2 test	SQuAD-V2 dev	11,873
mini-NQ train	NQ-Open train	800
mini-NQ validation	NQ-Open train	200
mini-NQ test	NQ-Open dev	3,499
mini-GQA train	GQA dev	105,640
mini-GQA validation	GQA dev	26,422
mini-GQA test	GQA test	12,578
mini-Viz train	VizWiz dev	3,456
mini-Viz validation	VizWiz dev	863
mini-Viz test	VizWiz test	8,000

Table 1: Details of datasets used for *InfoSel* ensemble training.

Specifically, FT-MT solves a multi-label classification task by mapping the fused question and visual representation to a label vector formalized by the accuracy of a list of frequent answers extracted from the training data. The training scheme is adapted from (Li et al., 2021b).

***InfoSel*$^{+}$-MT.** Similar to *InfoSel*$^{+}$-BERT, *InfoSel*$^{+}$-MT ensembles the predictions from FT-MT and *InfoSel*-MT using the same training scheme of *InfoSel*-MT.

4 Experiments

Datasets. To address the constraint of having a limited amount of labeled data, we created smaller QA datasets by randomly sampling 1,000 samples from public benchmark datasets for QA. Specifically, we established Mini-SDv2 and Mini-NQ, containing samples from the SQuAD-V2(Rajpurkar et al., 2018) and NQ-Open (Kwiatkowski et al., 2019) training datasets respectively. For Mini-NQ, we used the long answer as the context and the short answer as the ground-truth answer like (Fisch et al., 2019). The 1,000 samples of each dataset were divided into train and validation data using an 8:2 ratio, while the test data was set to the dev data of the original datasets (since the original test data is not publicly available). For VQA tasks, we constructed Mini-GQA and Mini-Viz datasets using only the development dataset of GQA (Hudson and Manning, 2019) and VizWiz (Gurari et al., 2018)) respectively. These dev data were divided into train and validation data using an 8:2 ratio, while the test data remained the same as the test data of the original datasets. Table 1 demonstrates the details of these datasets. More descriptions about the datasets are shown in Appendix A.1.

Base Models. ChatGPT, LLaMA-2-70b-chat (Touvron et al., 2023) and GPT3.5 text-davinci-003, which are state-of-the-art LLMs, are chosen to provide candidate answers for our QA ensemble training. Three 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) with different architectures are selected as base models for VQA ensemble training. All base models either can only return predictions without any logits or scores, or this restriction is assumed for the purpose of our study. More details about the model description are shown in Appendix A.2.

The **Oracle** represents the maximum capability of a combination of base models. Specifically, for each input, the oracle always selects the best answer, i.e., the answer with the highest agreement with the ground truth, among all the candidate answers predicted by base models. Thus, the oracle score represents the performance of an ideal ensemble model.

Baselines. **Majority voting (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, a random one is picked. Similar to (Schick and Schütze, 2020), which uses the model accuracy of the training set before training as the weight for average weighting, we use the model's corresponding out-domain accuracy as the weight for **weighted voting (WV)**.

Evaluation Metric. LLMs intend to generate contextual answers which lead to lower scores in exact match (EM) even when with high recall scores (number of common tokens / number of ground truth answer tokens). Therefore, we mainly use the $F1$ score as the main evaluation metric for QA performance. The base VQA models are not trained for unanswerable visual questions and thus perform badly on the VizWiz dataset, which contains $\sim$28% of visual questions that are deemed unanswerable. Therefore, we consider data samples with the ground-truth answers and predicted answers not equal to "unanswerable", "unknown" or " " as relevant samples and retrieved samples respectively. Precision, recall and $F1$ score are reported on relevant and retrieved samples.

Setup. We use a learning rate of 5×10^{-5} and batch size of 4 for training *InfoSel*-BERT and FT-BERT over 5 epochs. For *InfoSel*-MT and FT-MT, we use a learning rate of 5×10^{-5} and batch size of 16, the models are trained over 20 epochs. Experiments are run on Nvidia DGX-1 with 1 GPU.

	mini-SDv2				mini-NQ			
	EM	**P**	**R**	**F1**	**EM**	**P**	**R**	**F1**
LLaMA-2-70b-chat	0.24	7.20	52.70	11.34	28.07	43.20	**79.21**	46.47
text-davinci-003	**52.37**	**56.86**	63.58	58.44	52.24	69.96	77.50	69.44
ChatGPT	30.89	40.53	**68.54**	44.95	57.53	74.15	75.81	71.54
Oracle	58.61	64.04	77.98	66.20	64.02	80.54	87.97	79.21
MV	26.95	34.23	61.22	37.75	46.07	62.56	77.66	62.43
WV	**52.37**	**56.86**	63.58	58.44	57.53	74.15	75.81	71.54
FT-BERT	46.80	47.70	48.86	47.68	36.52	42.81	43.46	40.60
***InfoSel*-BERT**	**52.36**	**56.85**	63.59	**63.71**	**58.45**	**75.99**	77.75	**73.37**
***InfoSel*$^{+}$-BERT**	52.12	52.74	53.47	52.68	46.61	55.08	54.63	52.49

Table 2: Model performance on textual QA tasks. The best results are bolded.

5 Results and Analysis

5.1 Main Result of *InfoSel* for Textual QA

Table 2 shows the main results of *InfoSel*-BERT and the comparison with base models, baselines and FT-BERT for textual QA tasks. LLaMA-2-70b-chat performed the worst in F1 score among the base models, the main reason is that it usually provides a longer explanation text for the generated answers compared to the other two LLMs. All the models perform better in mini-NQ as mini-SDv2 test data contains ∼50% of unanswerable questions which increases the difficulty of the QA task. The oracle of the base model indicates an ideal ensemble method can only improve the $F1$ score of mini-SDv2 and mini-NQ from 58.44 to 66.20 and 71.54 to 79.21. The results of the LLMs can be different from (Laskar et al., 2023) or (Kocoń et al., 2023) because we do not apply any post-processing, human evaluation or output constraints for the generated answers. Another factor is that LLMs are updating over time and thus can provide different responses for different users.

Weight voting always selects the best model (in F1 score). Majority voting can randomly capture the answers from a model with a lower $F1$ score but a higher recall score (LLaMA-2-70b-chat), which is showcased by achieving a higher recall than weight voting in mini-NQ.

With only 1,000 samples, *InfoSel*-BERT achieves 96.24% (63.71/66.20) of the oracle in mini-SDv2 and 93.06% (73.37/79.21) on mini-NQ. In contrast, FT-BERT falls obviously (more than 10%) from *InfoSel*-BERT even when it outperforms two of the base models in mini-SDv2. *InfoSel*$^{+}$ does not bring an obvious improvement here due to the poor performance of FT-BERT.

We studied the impact of training *InfoSel*-BERT and FT-BERT with different amounts of training

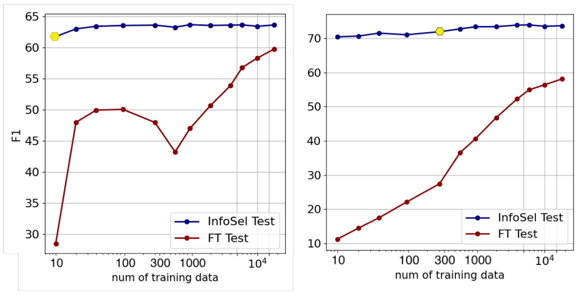


Figure 3: Test performance of *InfoSel*-BERT (referred to *InfoSel* in the figure) and FT-BERT (referred to FT) over an increasing number of training data from SQuAD-V2 (left) and NQ-Open (right). The yellow dot highlights the point when *InfoSel* outperforms base models.

data from SQuAD-V2 and NQ-Open and demonstrated the result in Figure 3. We observe that *InfoSel*-BERT can achieve a higher F1 score than base models even when only 10 samples from SQuAD-V2 are used for training, while 300 samples are needed from NQ-Open to get a better result than base models. Additionally, we find that a larger training data size benefits FT-BERT more than *InfoSel*-BERT. The F1 score of FT-BERT increased ∼200% and ∼500% from 10 to 10,000 training samples on SQuAD-V2 and NQ-Open respectively, while *InfoSel*-BERT only increased only ∼3% and ∼4%. However, the result also confirmed that fine-tuning requires numerous training data for getting a comparable performance with *InfoSel*.

5.2 Main Result of *InfoSel* for VQA

Table 3 demonstrates the performance of base models, baselines and our methods for VQA task. All the base models achieve close performance on both datasets. mini-Viz contains ∼28% unanswerable questions and thus gets worse scores than mini-GQA. Fine-tuning (FT-MT) leads to overfitting on GQA as the highest validation accuracy (68.86%) does not guarantee any improvement on test data.

Model	mini-GQA		mini-Viz		
	Val	Test	Val		Test
	Acc	Acc	Acc	F1	Acc
ALBEF	54.82	50.60	21.92	20.51	21.28
BLIP	52.94	48.08	22.64	20.08	20.80
VLMo	54.00	48.21	21.95	20.10	19.77
Oracle	70.30	65.03	28.76	24.87	-
MV	55.85	51.05	23.64	21.48	21.47
WV	56.45	52.10	23.82	21.59	19.43
FT-MT	68.86	50.48	51.71	20.66	51.76
***InfoSel*-MT**	63.00	**55.16**	25.13	22.60	23.16
***InfoSel*$^+$-MT**	**70.06**	52.54	**55.92**	**32.18**	**52.91**

Table 3: Validation and test performance on VQA tasks, more details of the precision, recall, and $F1$ score are shown in Table 7 in Appendix A. The test data annotation of mini-Viz dataset is not accessible and thus the oracle score on test data can not be reported.

While *InfoSel*-MT overcame this problem with an improvement of 9% ((55.16-50.60)/50.60) from base models and achieving 84.81%(55.16/65.03) of the oracle. However, FT-MT enhanced $\sim 240\%$ (51.76/21.28) accuracy on mini-Viz, this is because fine-tuning introduced new labels (e.g., "unanswerable") for FT-MT which base models have not seen during training. This statement is showcased by the higher F1 score of *InfoSel*-MT when compared with FT-MT. Finally, *InfoSel*$^+$-MT perfectly blends the strengths of *InfoSel* and fine-tuning by ensembling *InfoSel*-MT and FT-MT, which improved upon both models from 51.76 to 52.91.

Figure 4 demonstrates that *InfoSel*-MT can outperform base models with only 5% (6603 samples) of training data from mini-GQA and 20% (864 samples) from mini-Viz. Additionally, we notice that the increase in training data size does not guarantee a performance improvement in fine-tuning (showcased by mini-GQA).

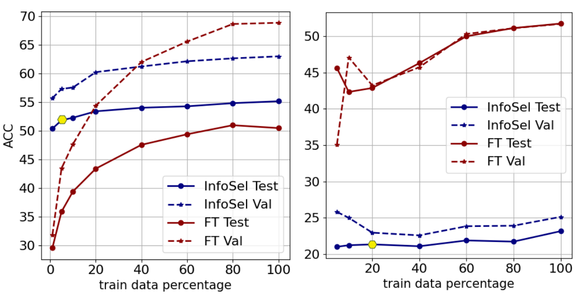


Figure 4: Validation and test performance of *InfoSel*-MT (referred to *InfoSel* in the figure) and FT-MT (referred to FT) over an increasing percentage of training data from mini-GQA (left) and mini-Viz (right). The yellow dot highlights the point when *InfoSel* outperforms base models.

Model	mini-GQA		mini-Viz	
	Val	Test	Val	Test
InfoSel-MT(V)	55.33	50.56	22.73	20.79
InfoSel-MT(Q)	57.70	51.11	23.23	21.21
InfoSel-MT(VQ)	57.75	50.83	23.33	20.06
InfoSel-MT(VA)	59.25	52.38	24.47	22.66
InfoSel-MT(QA)	62.84	54.76	25.02	22.89
***InfoSel*-MT(VQA)**	**63.00**	**55.16**	**25.20**	**23.26**

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

5.3 Analysis of Model Disagreements

Figure 5 demonstrates the model disagreement over different datasets. The number in the tables presents the number of samples that column models provide better predictions (with higher evaluation scores) than the row models. That is model pairs with dark cells have many disagreements and can potentially benefit from ensembling. In particular, for a dark cell, the row model provides many good answers that the column model does not find. Hence, the column for the oracle contains all 0's when compared to the base models, but fine-tuning (FT, *InfoSel*$^+$) can find some answers that the base models cannot find.

This analysis sheds light on the quality of models and the effect of fine-tuning in the different settings. For the textual QA datasets, LLaMA is clearly outperformed in all comparisons (dark LLaMA column and light LLaMA row), but fine-tuning (FT, *InfoSel*$^+$) has difficulties contributing substantial amounts of valuable answers.

For mini-GQA, the different models are able to contribute more evenly. mini-Viz is the only setting where fine-tuning finds substantial amounts of answers not found by the base models (dark rows for FT and *InfoSel*$^+$).

5.4 Ablation Study

In an ablation experiment (Table 4), we compared the effect of providing different information to *InfoSel*-MT, and found that the best setting is to combine the image, question and answer (V+Q+A) information, and the second most useful is Q+A information. The worst setting is to apply only the image as the signal. The reason can be that a single image usually has multiple corresponding questions on GQA, and thus hard for the model to learn discriminative features.

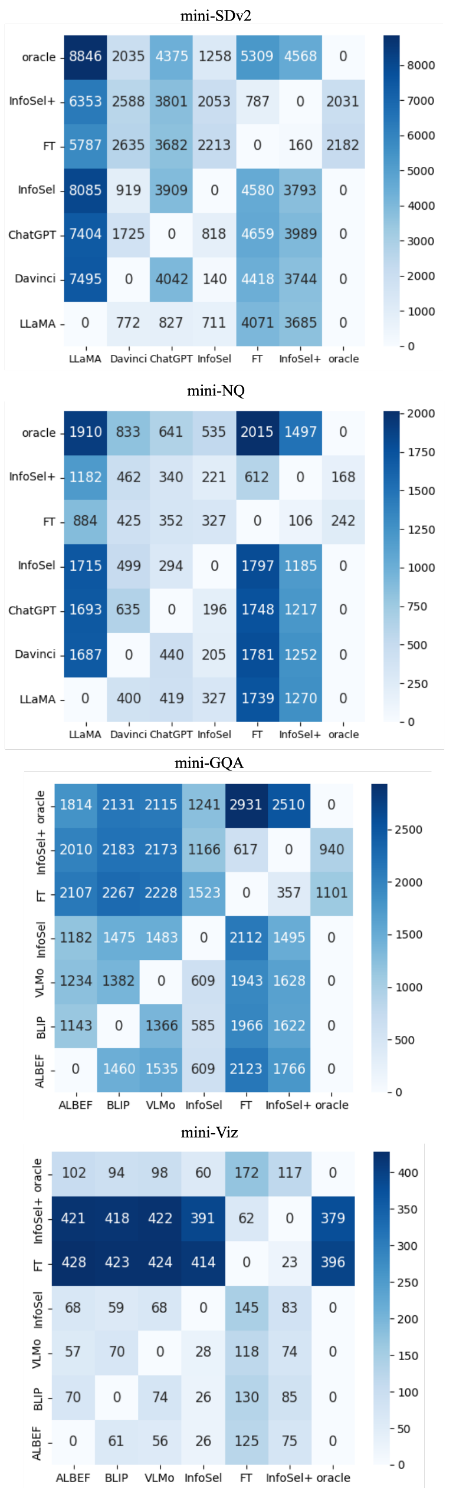


Figure 5: Model disagreements over different datasets.

5.5 Case Study

Table 5, 6 (in Appendix) demonstrate several interesting cases from the predictions of different models for textual and visual QA tasks. We observe from Table 5 that *InfoSel*-BERT selects answers from different language models. However, *InfoSel*$^+$-BERT may select wrong answers from the overfitted FT-BERT model and underperforms *InfoSel*-BERT in those instances. The last case showcases a wrong ground-truth answer provided by the original dataset. However, LLMs are still able to generate the right answer with their contextual comprehension ability, while FT-BERT limited to classification tasks can only extract answer tokens from context and thus cannot provide the right answer. Therefore, ensembling LLMs to utilize their powerful comprehension ability can benefit users more than fine-tuning small-size models.

Table 6 shows that *InfoSel* and *InfoSel*$^+$ are able to capture the right answer even though only one of the base models provides the right answer. The last case demonstrates that *InfoSel*$^+$captures the new label "unanswerable" introduced by FT-MT, which can never be predicted by *InfoSel*-MT as the base models always predict an answer. Therefore, it is essential to include FT-MT for ensembling training when out-domain datasets contain a high percentage of new labels.

6 Conclusion

The rise of black-box AI services and hosted models demands for methods to choose an answer from such systems when their responses disagree. Previous methods such as weighted voting are too simplistic since they do not capture sample-specific patterns that can help in determining which model is the most reliable for one particular example type; and/or they need access to components that cannot be assumed to be available, such as prediction confidences or tunable model parameters.

In this paper we propose *InfoSel*, a lightweight method to select an answer from several distinct base models, considering question-, context-/image- and predicted answer-information (but not based on predicted answer confidences). In *InfoSel*, only a small-size transformer for answer selection is fine-tuned, and *InfoSel* consistently improves over always choosing the answer from the overall best model.

Extensive analysis, comparing *InfoSel* to an oracle ensemble score, and to a fine-tuned similar-size QA model, highlights the robustness of *InfoSel*. *InfoSel* reaches (depending on the dataset) between 84% and 96% of the oracle in textual and visual question answering tasks.

7 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 H. 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 Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Alex J Chan and Mihaela van der Schaar. 2022. Synthetic model combination: An instance-wise approach to unsupervised ensemble learning. *arXiv preprint arXiv:2210.05320*.

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.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. 2017. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.

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.

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. ArXiv:2201.12086 [cs].

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021a. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. ArXiv:2107.07651 [cs].

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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. ArXiv:2010.12831 [cs].

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Tegwen Malik, Yogesh Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, et al. 2023. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. How to construct deep recurrent neural networks.

Haritz Puerto, Gözde Gül Şahin, and Iryna Gurevych. 2021. Metaqa: Combining expert agents for multi-skill question answering. *arXiv preprint arXiv:2112.01922*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers.

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.

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

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2019. Open-ended visual question answering by multi-modal domain adaptation. *arXiv preprint arXiv:1911.04058*.

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.

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 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 by Google AI Language and made available for research purposes.

In order to get high-quality answers from LLMs, we use the prompts consisting of the question and context from these two datasets. The details about the prompts are demonstrated in Table 8.

GQA 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 is a benchmark dataset for visual question answering. It includes 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 in-domain dataset (VQA v2 (Antol et al., 2015)) with out-domain datasets (GQA, VizWiz) for VQA base models. Figure 6 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. 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 in-domain dataset and out-domain datasets.

A.2 Base Models

ChatGPT also named chat Generative pre-trained Transformer, is a natural language processing model developed and released by OpenAI. It utilizes OpenAI's GPT foundation models – GPT-3.5

	mini-SDv2		mini-NQ	
Context:	... The building was designed by architects Marek Budzyński and Zbigniew Badowski...	... Derrick Norman Lehmer's list of primes up to 10,006,721 ...	Dwight David Howard ... player for the Charlotte Hornets ...	... in 2005 and the release of her epony-mous debut album the following year ...
Question:	What profession does Zbigniew Marek have?	How many primes were included in Derrick Norman Lehmer's list of prime numbers?	who did Dwight Howard play for last year?	when did Taylor Swift 's first album release?
LLaMA-2-70b-chat	architect	unanswerable	**Charlotte Hornets**	2006
text-davinci-003	Architect	**10,006,721**	The Houston Rockets	2006
ChatGPT	**unanswerable**	unanswerable	Washington Wizards	2006
FT-BERT	architects Marek Budzyński and Zbigniew Badowski	unanswerable	Dwight David Howard	**2005**
***InfoSel*-MT**	**unanswerable**	**10,006,721**	**Charlotte Hornets**	2006
***InfoSel*$^+$-MT**	**unanswerable**	unanswerable	Dwight David Howard	**2005**

Table 5: Case study of our models on mini-SDv2 test and mini-NQ test data. Answers of LLMs are shortened to keywords for better demonstration. Ground-truth answers are bolded, and one suspicious ground-truth answer is colored red.

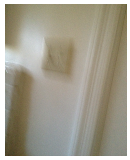
	mini-GQA		mini-Viz	
Image:				
Question:	What appliance is it?	Is the tall tree on the right?	What kind of food is in this can?	What is this pro-duct?
ALBEF	blender	yes	fruit salad	refrigerator
BLIP	**toaster**	yes	**vegetable soup**	toilet
VLMo	microwave	yes	fruit	door
FT-MT	coffee maker	**no**	soup	**unanswerable**
***InfoSel*-MT**	**toaster**	yes	**vegetable soup**	toilet
***InfoSel*$^+$-MT**	coffee maker	**no**	**vegetable soup**	**unanswerable**

Table 6: Case study of our models on mini-GQA test and mini-Viz validation data. Ground-truth answers are bolded.

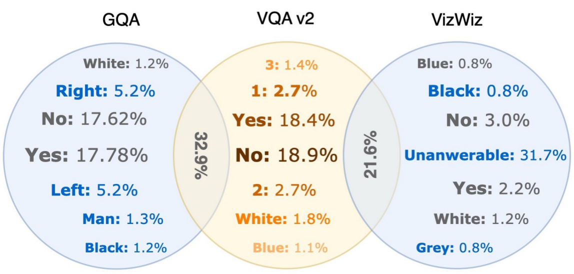


Figure 6: Top 7 most frequent answers of VQA v2 (in-domain dataset of VQA models), GQA and VizWiz (out-domain datasets).

and GPT-4 – to generate context-based responses to user prompts.

LLaMA-2-70b-chat (Touvron et al., 2023) is a 70B parameter generative text model developed by Meta and launched as part of the LLaMA 2 collection of fine-tuned large language models in July 2023. It was pre-trained on 2 trillion tokens of publicly available data and has a context length of 4096 tokens (i.e., twice the context length of LLaMA 1 models).

GPT 3.5 text-davinci-003 is part of the GPT 3.5 family of large language models introduced by OpenAI in 2022. It has a capacity of 175 billion parameters, a context window of 4097 tokens and was trained on a dataset that contains data up to June 2021.

ALBEF (Li et al., 2021a)[7] first encodes the image and text with an image encoder (visual transformer (Dosovitskiy et al., 2020)) 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. Differnet from U-VisualBERT, ALBEF uses a 6-layer transformer decoder to generate answers for VQA task.

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

	Model	GQA Val P	GQA Val R	GQA Val F1	GQA Test P	GQA Test R	GQA Test F1	VizWiz Val P	VizWiz Val R	VizWiz Val F1
BASE	ALBEF	54.82	54.82	54.82	50.60	50.60	50.60	14.68	34.00	20.51
	BLIP	52.94	52.94	52.94	48.08	48.08	48.08	14.35	33.43	20.08
	VLMo	57.12	54.00	55.52	52.87	48.21	50.43	14.40	33.24	20.10
	Oracle	70.30	70.30	70.30	65.03	65.03	65.03	17.81	41.24	24.87
	MV	56.56	55.85	56.21	52.24	51.05	51.64	15.37	35.65	21.48
	WV	56.45	56.45	56.45	52.10	52.10	52.10	15.43	35.95	21.59
TF	FT-MT	68.86	68.86	68.86	50.48	50.48	50.48	29.26	15.97	20.66
	***InfoSel*-MT**	63.00	63.00	63.00	**55.16**	**55.16**	**55.16**	16.16	**37.59**	22.60
	***InfoSel*$^+$-MT**	**70.06**	**70.06**	70.06	52.54	52.54	52.54	**39.07**	27.35	**32.18**

Table 7: Validation and test performance of different models on new domain datasets.

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 8: 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.

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

Table 9: Parameter size of models.

BLIP (Li et al., 2022)[8] 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, VQA task is considered as an answer generation task in this method.

VLMo (Bao et al., 2022)[9] 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.

The model parameter sizes are shown in Table 9.

A.3 Multi-modal Information Concatenation or Fusion?

We studied the impact of concatenating and fusing multi-modal input information for VQA task. *InfoSel*-MLP is an alternative model type for *InfoSel* which processes all the input information separately with a simple multi-layer perceptron (MLP) instead of MT. A pre-trained Sentence-BERT (Reimers and Gurevych, 2019) [10] M_{qa} is used for generating question embedding h^q and answer embeddings h^a.

$$h_i^q = M_{qa}(Q_i), h^q \in \mathbb{R}^{768}$$

$$h_i^{a_j} = M_{qa}(A_{ij}), h_i^{a_j} \in \mathbb{R}^{768}$$

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

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

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

Model	mini-GQA Val	mini-GQA Test	mini-Viz Val	mini-Viz Test
InfoSel-MLP	57.87	52.35	22.68	21.12
***InfoSel*-MT**	**63.00**	**55.16**	**25.13**	**23.16**

Table 10: Comparison of using different architecture for processing input information in a different way. Input concatenation result is demonstrated by *InfoSel*-MLP and the fusion result is shown by *InfoSel*-MT.

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

$$min_\theta \sum_{i=1}^{N} BCE(MLP_\theta([h_i^q, h_i^v, [h_i^{a_j}]_{j=1}^{K}]), Y_i^v) \tag{3}$$

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 10 demonstrates the performance of input concatenation result (*InfoSel*-MLP) and fusion result (*InfoSel*-MT). We observe that *InfoSel*-MT achieves $\sim 3\%$ and $\sim 2\%$ 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.