FINETUNING LANGUAGE MODELS TO EMIT LINGUISTIC EXPRESSIONS OF UNCERTAINTY

Arslan Chaudhry, Sridhar Thiagarajan, Dilan Gorur *
Google DeepMind
{arslanch,sthiagarajan,dilang}@google.com

ABSTRACT

Large language models (LLMs) are increasingly employed in information-seeking and decision-making tasks. Despite their broad utility, LLMs tend to generate information that conflict with real-world facts, and their persuasive style can make these inaccuracies appear confident and convincing. As a result, end-users struggle to consistently align the confidence expressed by LLMs with the accuracy of their predictions, often leading to either blind trust in all outputs or a complete disregard for their reliability. In this work, we explore supervised fine-tuning on uncertainty-augmented predictions as a method to develop models that produce linguistic expressions of uncertainty. Specifically, we measure the calibration of pre-trained models and fine-tune language models to generate *calibrated* linguistic expressions of uncertainty. Through experiments on various question-answering datasets, we demonstrate that LLMs are well-calibrated in assessing their predictions, and supervised fine-tuning based on the model's own confidence leads to well-calibrated expressions of uncertainty, particularly for single-claim answers.

1 INTRODUCTION

Large Language Models (LLMs) are emerging as powerful tools that can absorb internet-scale data in the parametric knowledge of a neural network (Brown, 2020; Rae et al., 2021; Hoffmann et al., 2022). These models are the foundational blocks of several conversational agents (Achiam et al., 2023; Team et al., 2023; Anthropic, 2024; Dubey et al., 2024) that people are increasingly relying on for information seeking and decision making tasks. Owing to their natural language interface, these models are more easily accessible to and interacted by the general public than any other machine learning models that existed before[1]. This widespread utility naturally raises questions around the truthfulness and factuality of the predictions made by these models.

Despite being state-of-the-art on several natural language processing (NLP) tasks (Brown, 2020; Team et al., 2023), LLMs occasionally produce incorrect predictions – especially on queries that are outside the training distribution of the model. These inaccuracies are tricky to deal with as models do not express uncertainty in their generations, making their statements sound highly confident (Huang et al., 2023; Ji et al., 2023). As such, the end-users have no way of associating the confidence the model is expressing in its predictions to its correctness. This limits the utility of these models in many safety-critical applications, such as medicine (Thirunavukarasu et al., 2023; Saab et al., 2024) and law (Dahl et al., 2024), and precludes the users from reliably using the predictions made by these models in many information seeking tasks (Passi and Vorvoreanu, 2022; Vasconcelos et al., 2023). We illustrate this with a mock scenario, as shown in **Figure 1**.

In this work, we study how to augment an LLM's prediction with a linguistic expression of uncertainty, where an uncertainty expression reflects the likelihood that the model's predictions are accurate, aggregated across all samples with similar uncertainty levels. For instance, if a model indicates an uncertainty expression such as 'it is unlikely' and this corresponds to approximately 30% confidence, we expect around 30% of those predictions to be correct. In other words, we aim for the uncertainty-augmented answers to be well-calibrated (Guo et al., 2017). Previous approaches have used similar

---

*The first two authors contributed equally.

[1]As evidenced by the surge of ChatGPT users into millions in the first few weeks of its launch.
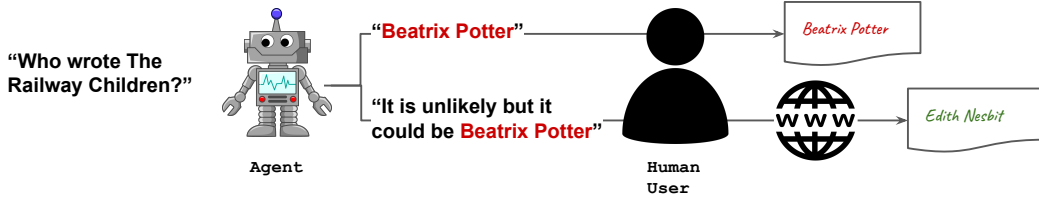
Figure 1: **Motivation**: The agent provides an incorrect response to a given query. In the response at the bottom, however, the agent includes an uncertainty expression. Without this uncertainty expression, as seen in the response at the top, the human user might form an incorrect belief about the world. In contrast, with the uncertainty-augmented response at the bottom, the human user is prompted to consult additional resources, leading to a more accurate understanding of the world.

measures to avoid answering uncertain questions (Cole et al., 2023). In contrast, our focus is on always providing an answer while appropriately conveying the associated uncertainty. This enables the end-user to decide whether to act on the model's information, request further clarification from the model, or consult additional sources (see **Figure 1**). Additionally, we believe that for many user-facing applications, conveying model uncertainty linguistically is often more effective than presenting it numerically. This is because **a)** linguistic expressions, such as 'I am certain' or 'It is highly unlikely,' are more intuitive for humans to interpret than raw numbers like 90% or 10%; **b)** linguistic expressions integrate more naturally with predictions; **c)** language can indicate the source of uncertainty, whether it arises from the model's limitations or inherent uncertainty (Hüllermeier and Waegeman, 2021); and **d)** linguistic expressions are more adaptable for downstream processing, such as text-to-speech conversion.

We focus on supervised fine-tuning (SFT) with curated datasets as the primary method for equipping models with linguistic expressions of uncertainty in their predictions. To achieve this, we first assess the model's confidence in its predictions by querying whether each prediction is true or false. The confidence corresponds to the normalized[2] probability assigned to the 'true' token. Building on the work of Kadavath et al. (2022), we observe that this self-evaluation score is fairly well-calibrated for single-claim answers across the pre-trained models we tested (**Figures 4 and 7**). Moreover, applying mild post-processing, such as isotonic regression on a small calibration set, achieves near-perfect calibration of both base and instruction-tuned models across various sizes (**Figure 4**). Next, we map the confidence scores to linguistic expressions using the framework from Fagen-Ulmschneider (2019), which was developed based on a survey where human subjects associated probability ranges with different uncertainty expressions. These linguistic expressions are then combined with the corresponding predictions to create a fine-tuning dataset. **Figure 2** illustrates the overall process of curating the finetuning dataset. When fine-tuned on this dataset, the resulting models generate predictions with well-calibrated linguistic expressions of uncertainty (**Figure 5**).

Overall, our contributions include,

1. We provide a finetuning receipe for equipping models with linguistic expressions of uncertainty.

2. We present the calibration plots of Gemini 1.0 small and medium sized models after pre-training and alignment phases on three Q/A datasets – TriviaQA (Joshi et al., 2017), AmbigQA (Min et al., 2020) and Truthful QA (Lin et al., 2021).

3. We find that pre-trained models are better calibrated than models fine-tuned for alignment, and that calibration improves with larger model sizes, consistent with existing literature (Kadavath et al., 2022; Achiam et al., 2023).

4. We explore various methods for incorporating linguistic expressions of uncertainty into predictions. Our findings indicate that finetuning with predictions augmented by adding uncertainty expressions *after* the actual answer results in the most well-calibrated finetuned models.

---

[2]The normalization constant is the sum of the probablities assigned to the 'true' and 'false' tokens.

2 RELATED WORK

There is a long body of work which studies the ability of machine learning models to express uncertainty in their predictions (Gawlikowski et al., 2023; Guo et al., 2017). The two broad sources of uncertainty typically in focus are epistemic uncertainty; i.e uncertainty stemming from the model's lack of knowledge, and aleatoric uncertainty, uncertainty as a result of inherent ambiguity in the task (Hüllermeier and Waegeman, 2021). We focus on only the former in this work. For epistemic uncertainty, in the case of language models, there have been studies (Kadavath et al., 2022; Jiang et al., 2021; Desai and Durrett, 2020; Tian et al., 2023) which look at the ability of language models to assess the confidence in the factuality of their predictions, looking at both in-domain and out-of-domain datasets. Kadavath et al. (2022) show that pre-trained models are reasonably well calibrated at predicting whether their own output is factual or not in a few-shot setting. We leverage their few-shot prompting strategy to elicit numerical confidence of the model in their own predictions. Tian et al. (2023) find that for instruction tuned models, verbalized probabilities are better-calibrated than conditional probabilities in a setting when models are prompted to output the uncertainty in their predictions. We refer the reader to (Geng et al., 2024) for a more comprehensive overview of methods to elicit epistemic uncertainty.

Lin et al. (2022) introduce the notion of linguistic uncertainty in language models, teaching a pre-trained GPT-3 model to express its uncertainty linguistically along with its prediction on mathematics tasks. A key difference of our work from theirs is that our uncertainty targets are obtained in a pointwise manner, whereas they assign the targets based on average task performance on questions in the same sub-task. We also study the effect of the placement of the uncertainty estimate (prefix/suffix) on the quality of the estimates. Zhou et al. (2023) study the effect of prefixed expression of uncertainty on a pre-existing language model's generation, finding that expression of high certainty hurt the accuracy compared to weaker expressions. Mielke et al. (2022) also work on making models emit linguistic expression of uncertainty, but their process is a two-stage pipeline where a separate calibrator predicts the numerical probability of correctness, and the language model then adds a linguistic marker based on this. Most similar to our work is (Band et al., 2024), who work on obtaining calibrated long form predictions from LLMs. Their method of obtaining uncertainty targets relies on a form of self-consistency extended to long form answers, and they also perform an additional reinforcement learning step after their supervised finetuning stage. They do not study the quality effect of position of the uncertainty estimates relative to the answer (prefix/postfixed).

3 SETUP AND METHOD

We investigate the question-answering (Q/A) task where an LLM, specifically Gemini 1.0, incorporates expressions of uncertainty while generating answers. In a Q/A task, we work with a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, consisting of n examples where x represents questions and y represents ground-truth answers. The sets of all questions and answers are denoted by $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^n$, respectively. Given a question x, the LLM $M(\cdot)$ produces a prediction $\hat{y} = M(x)$ through autoregressive decoding. The dataset $\mathcal{D}$ is divided into three non-overlapping subsets: few-shot ($\mathcal{D}_{fs}$), calibration ($\mathcal{D}_{cal}$), and training ($\mathcal{D}_{tr}$). These subsets are used in different phases of finetuning dataset curation – $\mathcal{D}_{fs}$ is used for computing model confidence in the predictions, $\mathcal{D}_{cal}$ is used for fitting the isotonic regressor that is then used for post-processing the confidence scores, and $\mathcal{D}_{tr}$ is used for finetuning (see **Algorithm 1** for more details). For each dataset $\mathcal{D}$, we reserve a separate held-out subset $\mathcal{D}_{te}$ for evaluation purposes.

3.1 METHOD

How can we train models to produce accurate expressions of uncertainty? One approach is few-shot prompting with uncertainty-augmented examples. However, recent literature suggests that LLMs struggle to generate well-calibrated linguistic expressions of uncertainty through in-context learning (Zhou et al., 2023). Additionally, in-context learning through few-shot prompting incurs extra inference costs for processing prompts with each query, making it less optimal from a latency perspective. To address these issues, we investigate supervised fine-tuning (SFT) on datasets augmented with uncertainty expressions as an alternative approach. A critical aspect of curating such datasets is ensur-
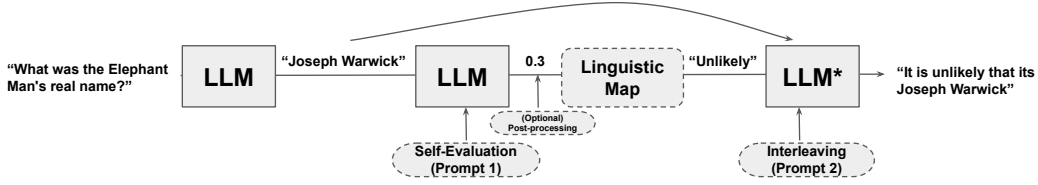
Figure 2: **Finetuning dataset curation process**: Here `LLM` refers to the language model that we are interested in finetuning. `LLM*` refers to an operation that mixes the uncertainty expression with the model prediction – it can be a prompted language model (interleaved case) or simply an operation which prefixes/post-fixes the answer with the expression of uncertainty. Given a question on the left, the LLM produces a raw prediction and then computes its own confidence on that prediction. The confidence score is converted to a linguistic expression and augmented with the raw prediction. **Prompt 1** and **Prompt 2** are given in the appendix.
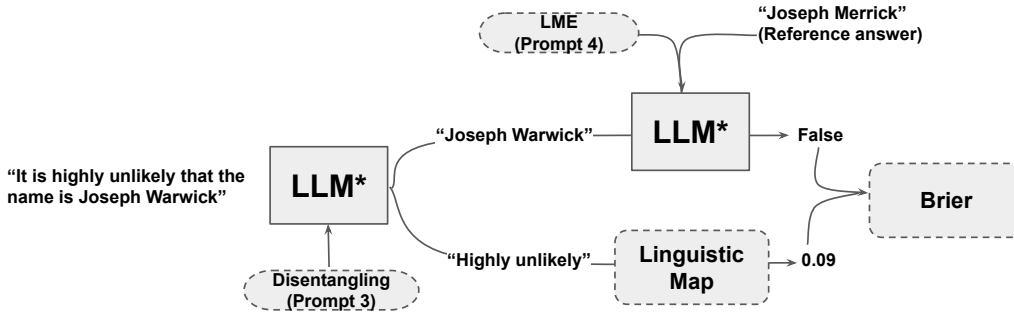

Figure 3: **Evaluation process**: Finetuned LLM produces an answer with the expression of uncertianty on the left that is split by a prompted LLM* into the raw answer ('Joseph Warwick') and expression of uncertainty ('Highly unlikely') using **Prompt 3**. LLM* then judges the correctness of the raw answer using the LME **Prompt 4** and uncertainty expression is converted to a float equal to the average of the probability range the uncertainty expression belongs to. Based on the correctness and uncertainty score, the final metric is computed.

ing that the uncertainty markers align with the model's knowledge about the questions. Inconsistent uncertainty expressions during fine-tuning can lead to hallucinated expressions during testing.

To ensure linguistic expressions of uncertainty are consistent with the model's knowledge, we first obtain confidence scores for model-generated samples through self-evaluation, as described by Kadavath et al. (2022). Specifically, we use a True/False self-evaluation task, where the model assesses the correctness of its own predictions (see **Prompt 1**). With the confidence scores in hand, we apply isotonic regression (Barlow and Brunk, 1972) using a small calibration set, $\mathcal{D}_{cal}$, to achieve nearly perfect calibration of uncertainty estimates. We then map these calibrated confidence scores to linguistic expressions of uncertainty based on human perception of uncertainty and probabilities (Fagen-Ulmschneider, 2019), as detailed in **Table 1**. The final step involves integrating uncertainty expressions with model predictions in the dataset. We explore three methods for this integration: **(1) Prefixed:** placing the uncertainty expression before the prediction (`expression, prediction`), **(2) Postfixed:** placing the uncertainty expression after the prediction (`prediction, expression`), and **(3) Interleaved:** incorporating the uncertainty expression within the prediction (`expression prediction`) using a prompted language model (see **Prompt 2**). The overall process of dataset curation is illustrated in **Figure 2** and **Algorithm 1**. After curating the dataset $\mathcal{D}^M$, we fine-tune the model (M) to generate linguistic expressions of uncertainty.

3.2 EVALUATION

We compute the accuracy (`Acc`) of a prediction by language model evaluation (`LME`) whereby a prompted language model compares the prediction with the ground truth answer (see **Prompt 4**). To measure calibration, we plot calibration charts between confidence scores and accuracy (**Figure 4**) where the x-axis is binned according to the probability ranges in the linguistic expressions map (Fagen-Ulmschneider, 2019). Further, to summarize the calibration error into a single scalar, we track,
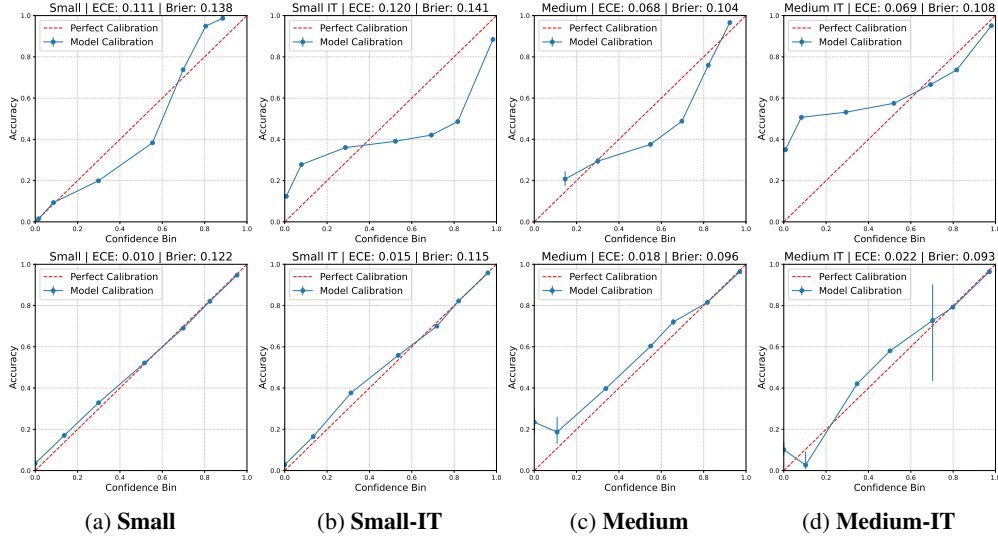
Figure 4: **TriviaA Calibration Chart**: The **top-row** shows raw calibration scores at temperature=1.0 without any post-processing. The **bottom row** shows post-processed calibration scores with isotonic regression. In each plot, the x-axis is the $p_{model}(true)$ of the generated prediction (shown here as `Confidence Bin`) and y-axis is probability of that prediction being actually correct (shown here as `Accuracy`). Expected Calibration Error (ECE) and Brier Score are reported at the top of each plot. The error bars show the variance of accuracy in each bin.

Expected Calibration Error (ECE): weighted average of the difference between the confidence assigned to examples in the bin, and accuracy of the predictions in the bin.

$$ece = \sum_{m=1}^{|B|} \frac{|B_m|}{n} \big| \texttt{Acc}(B_m) - \texttt{C}(B_m) \big|,$$

where $|B|$ are the total bins,

Brier Score: mean squared error between the confidence and correctness verdicts across all examples

$$\texttt{brier} = \frac{1}{n} \sum_{i=1}^{n} \big(\texttt{C}(x_i, \hat{y}_i) - \texttt{LME}(y, \hat{y}) \big)^2.$$

The lower these scores the better.

Once the models are finetuned to generate linguistic expressions of uncertainty, we test the calibration of their predictions on held-out test sets for each dataset. For this, we first extract the uncertainty expression from the rest of the prediction using a prompted language model (see **Prompt 3**). We then convert these uncertainty expressions into probability estimates using the same mapping employed for converting probabilities into linguistic expressions in the previous section. We measure calibration using Expected Calibration Error (ECE) and Brier Score and plot calibration charts based on these probability estimates. The complete evaluation procedure is outlined in **Figure 3** and **Algorithm 2**.

4 EXPERIMENTS

4.1 DATASETS

We use standard Q/A datasets – TriviaQA (Joshi et al., 2017), AmbigQA (Min et al., 2020) and TruthfulQA (Lin et al., 2021). For **TriviaQA**, we use the `wikipedia` version. From the `train` split, we take 16 examples for the $\mathcal{D}_{fs}$, 2000 examples for the $\mathcal{D}_{cal}$, and remaining examples for the $\mathcal{D}_{tr}$. The test set ($\mathcal{D}_{te}$) is taken from the `dev` set. For **AmbigQA**, we only take the unambiguous set, as we are focused only on conveying epistemic uncertainty. From the `train` split, we take
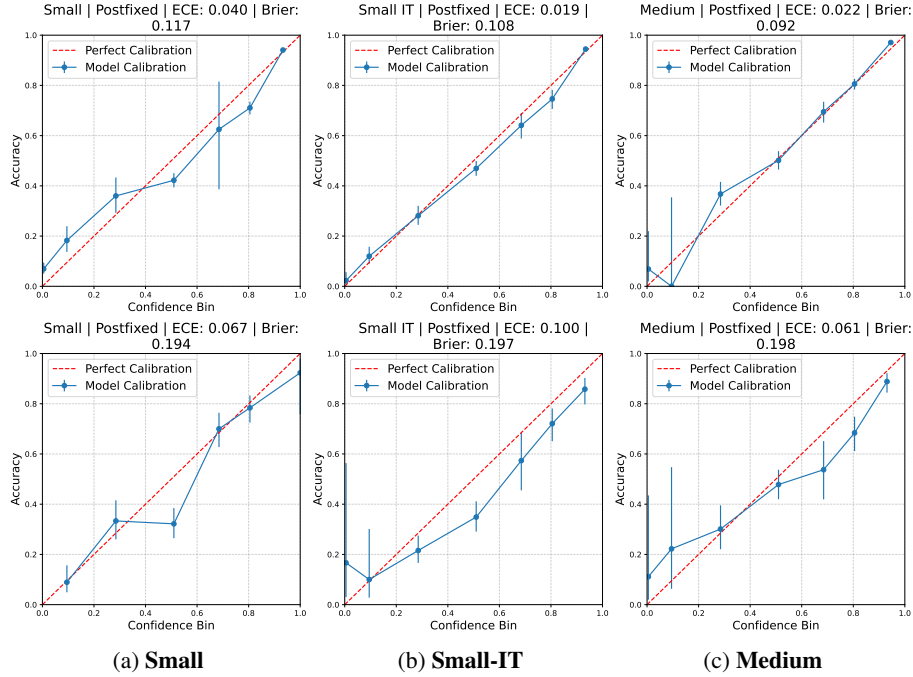
Figure 5: **Calibration Charts of Finetuned Models**: Top-row is **TriviaQA**. Bottom-row is **AmbigQA**. The model generates **post-fixed** uncertainty expressions. The x-axis is the $p_{model}(true)$ obtained by converting the linguistic expression of uncertainty to a float using **Table 1** (shown here as `Confidence Bin`) and y-axis is probability of that prediction being actually correct (shown here as `Accuracy`). No post-processing is done on the $p_{model}(true)$. Expected Calibration Error (ECE) and Brier Score are reported at the top of each plot. The error bars show the variance of accuracy in each bin.

16 examples for the $\mathcal{D}_{fs}$, 1000 examples for the $\mathcal{D}_{cal}$, and remaining examples for the $\mathcal{D}_{tr}$. The test set ($\mathcal{D}_{te}$) is taken from the `validation` set. For **TruthfulQA**, we split the `validation` set into a $\mathcal{D}_{fs}$ of size 8, $\mathcal{D}_{cal}$ of size 100, $\mathcal{D}_{te}$ of size 100 and $\mathcal{D}_{tr}$ of size ~600. The prediction is deemed correct if the `LME` (**Prompt 4**) judges it to be similar to *any* of the possible ground truth answers. We evaluate the accuracy of the interleaving (**Prompt 2**), disentangling (**Prompt 3**), and LME (**Prompt 4**) prompts on a uniformly sampled subset. We manually verify the performance of these prompts on small subsets of the data, and find that all of these independent tasks are performed at high accuracy, exceeding 95%.

4.2 MODELS

We conduct experiments with the Gemini 1.0 family of models (Team et al., 2023). Specifically, we use two variants, referred to as `Small` and `Medium`, which represent different architecture sizes. Our study includes both the pre-trained models and those that have been post-trained (aligned). The post-trained models are denoted with the '`IT`' marker.

4.3 FINETUNING DETAILS

For each question, we independently generate four model samples with a temperature setting of 1.0. We ensure balance in the curated dataset by setting a maximum number of examples per probability bin (see **Table 1**). Although the curated datasets contain model-generated samples with probability expressions, filtering predictions based on their correctness relative to the ground truth did not yield a significant performance difference. As a result, we do not apply correctness-based filtering to the datasets in the experiments discussed in this paper. The statistics for the final curated datasets are provided in **Table 2**.

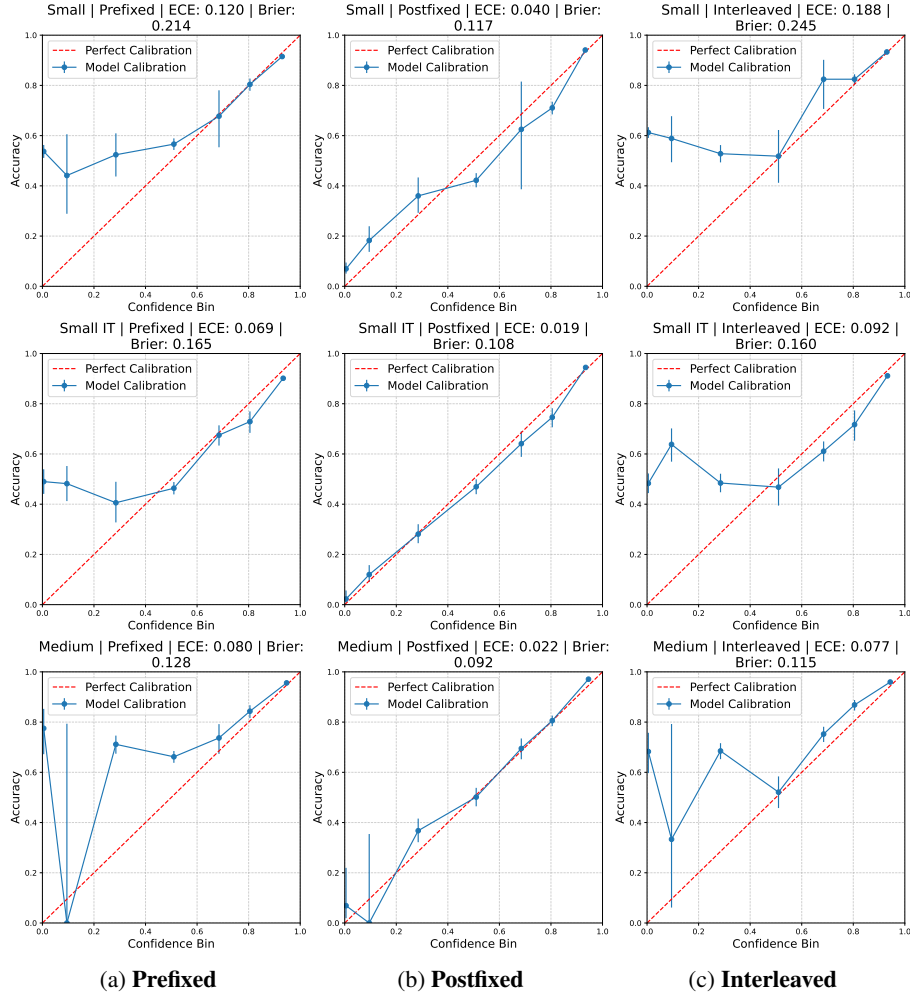(a) **Prefixed**   (b) **Postfixed**   (c) **Interleaved**

Figure 6: **TriviaQA uncertainty augmentation method**: Top-row is **Gemini Small**, middle-row is **Gemini Small-IT** and the bottom row is **Gemini Medium**. The models generate **postfixed** uncertainty expressions. The x-axis is the $p_{model}(true)$ obtained by converting the linguistic expression of uncertainty to a float using **Table 1** (shown here as Confidence Bin) and y-axis is probability of that prediction being actually correct (shown here as Accuracy). No post-processing is done on the $p_{model}(true)$. Expected Calibration Error (ECE) and Brier Score are reported at the top of each plot. The error bars show the variance of accuracy in each bin.

We use a batch size of 32 and train on each dataset for 3 epochs. The Adam optimizer (Kingma, 2014) with a cosine learning rate schedule is employed, where the learning rate is first linearly warmed up to $5e-7$ and then decayed to $5e-8$.

4.4 RESULTS

Calibration of Gemini models on the self-evaluation tasks: Figure 4 presents the calibration charts for Gemini 1.0 models on the TriviaQA dataset. Several key observations can be made:

1. The Gemini base models exhibit good calibration on the self-evaluation task (**Prompt 1**).
2. Calibration improves with larger model sizes, and pre-trained models demonstrate better calibration than instruction-tuned models, aligning with findings reported in the literature for other LLMs (Kadavath et al., 2022; Achiam et al., 2023).
3. Post-processing confidence scores using isotonic regression leads to significantly improved calibration.

Results for AmbigQA and TruthfulQA are shown in **Figures 7 and 8**, respectively. Notably, Gemini models do not achieve good calibration on TruthfulQA using the self-evaluation task. Although post-processing improves calibration somewhat, it is insufficient for effective uncertainty-augmented dataset curation. Consequently, we exclude the TruthfulQA dataset from our fine-tuning process.

Calibration of finetuned models: We now examine the calibration of linguistic expressions of uncertainty in finetuned models on held-out test sets. **Figure 5** displays the calibration charts for fine-tuned models that produce post-fixed uncertainty expressions. The figure shows that these fine-tuned models generate well-calibrated linguistic expressions of uncertainty. This result is consistent across different model sizes and applies equally to both pre-trained and instruction-tuned models.

Comparing different methods of uncertainty augmentation: Finally, we compare different methods for augmenting model predictions with uncertainty expressions – prefixed, postfixed, and interleaved – as previously described. **Figure 6** illustrates the performance of these augmentation methods on the TriviaQA dataset, with similar results for AmbigQA shown in **Figure 9**. The figure reveals that post-fixed uncertainty expressions, where the uncertainty is added after the main answer, result in the lowest calibration error. We hypothesize that this approach simplifies fine-tuning because the uncertainty expression does not influence the sampling of the main answer during autoregressive decoding. In contrast, prefixed or interleaved uncertainty expressions, where the uncertainty is added before or within the answer, can impact the answer's sampling distribution, as discussed by Zhou et al. (2023), leading to poorer calibration.

5   DISCUSSION AND CONCLUSION

In this work, we investigated supervised fine-tuning on the model's own uncertainty as a post-training step to enable models to generate linguistic expressions of uncertainty. We assessed the calibration of various Gemini 1.0 models and found them to be well-calibrated on the self-evaluation task. We then used these uncertainty scores to augment model predictions with linguistic expressions of uncertainty. Our findings show that fine-tuning with these augmented predictions results in models that produce well-calibrated linguistic expressions of uncertainty on held-out test sets. This fine-tuning approach can be employed as an independent post-training step between supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). Alternatively, uncertainty-augmented datasets can be integrated into the SFT process with appropriate system instructions that can guide models to express uncertainty.

Models capable of generating well-calibrated uncertainty expressions enable users to make informed inferences about the model's predictions. With linguistic expressions of uncertainty, users can reliably decide when to trust the model's predictions and when to seek additional information. This allows users greater control over how to utilize the model's outputs, unlike methods that rely on uncertainty estimates to determine when to abstain from giving a response. By abstaining, these methods may deprive users of potentially valuable information, even when the model's responses are uncertain. Therefore, the ability to produce well-calibrated expressions of uncertainty should be considered a key objective in the development of user facing foundational models.

REFERENCES

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

A. Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024.

N. Band, X. Li, T. Ma, and T. Hashimoto. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=rJVjQSQ8ye.

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

T. B. Brown. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.

J. R. Cole, M. J. Zhang, D. Gillick, J. M. Eisenschlos, B. Dhingra, and J. Eisenstein. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*, 2023.

M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

S. Desai and G. Durrett. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

W. Fagen-Ulmschneider. Perception of probability words, 2019. URL https://waf.cs.illinois.edu/visualizations/Perception-of-Probability-Words. Accessed: 2024-08-16.

J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, 2024.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Z. Jiang, J. Araki, H. Ding, and G. Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

S. Lin, J. Hilton, and O. Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

S. J. Mielke, A. Szlam, E. Dinan, and Y.-L. Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10: 857–872, 2022.

S. Min, J. Michael, H. Hajishirzi, and L. Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

S. Passi and M. Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 2022.

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

K. Zhou, D. Jurafsky, and T. Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. *arXiv preprint arXiv:2302.13439*, 2023.

A APPENDIX

Algorithm 1 Dataset curation with linguistic expressions of uncertainty

---

1: **Input:** Main Model (M), Augmentation Model (F), Confidence Routine (`Confidence`), Confidence Score to Linguistic Expression Map (`LinguisticMap`), Isotonic Regression Routine (`IsoReg`), Dataset ($\mathcal{D} = \{\mathcal{D}_{tr}, \mathcal{D}_{cal}, \mathcal{D}_{fs}\}$)
2: **Output:** A curated dataset ($\mathcal{D}_{tr}^M$)
3: $\hat{Y} = M(X)$ // Compute model predictions.
4: $C_{cal} = \texttt{Confidence}(X_{cal}, \hat{Y}_{cal}, \mathcal{D}_{fs})$ // Compute confidence scores on calibration set.
5: $C_{tr} = \texttt{Confidence}(X_{tr}, \hat{Y}_{tr}, \mathcal{D}_{fs})$ // Compute confidence scores on train set.
6: $R = \texttt{IsoReg}(Y_{cal}, \hat{Y}_{cal}, C_{cal})$ // Fit an isotonic regressor on the calibration set.
7: $C_{tr}^* = R(C_{tr})$ // Post-process confidence scores using the regressor.
8: $E_{tr} = \texttt{LinguisticMap}(C_{tr}^*)$ // Convert confidence scores to linguistic expressions of uncertainty.
9: $\mathcal{D}_{tr}^M = F(X_{tr}, \hat{Y}_{tr}, E_{tr})$ // Augment model predictions with linguistic expressions of uncertainty.
10: **Return** $\mathcal{D}_{tr}^M$

Algorithm 2 Evaluation of uncertainty augmented predictions

1: **Input:** Dataset containing uncertainty augmented predictions $\mathcal{D}_{te} = (X, \tilde{Y})$, Deaugmentation Model (`DeAug`), a map to convert uncertainty statements to floats (`InvMap`), Routine to compute ECE (`ECE`), Routine to compute Brier Score (`Brier Score`)
2: **Output:** $ece, brier$
3: $\hat{Y}, E = F(\tilde{Y})$ // Strip the uncertainty statements and main answers.
4: $C = \texttt{InvMap}(E)$ // Map uncertainty statements to floats.
5: $ece = \texttt{ECE}(Y, \hat{Y}, C)$
6: $brier = \texttt{Brier Score}(Y, \hat{Y}, C)$
7: **Return** $ece, brier$

Probability Range	Linguistic Expressions
[1.0-1.0]	('certain')
[0.87-0.99]	('almost certain', 'highly likely')
[0.75-0.86]	('very good chance', 'likely')
[0.63-0.74]	('i believe', 'probably', 'probable')
[0.40-0.62]	('chances are about even')
[0.18-0.39]	('i doubt', 'unlikely')
[0.02-0.17]	('little chance','chances are slight','improbable','highly unlikely')
[0.00-0.01]	('impossible')

Table 1: **Probability to Linguistic Expression Map**

```
You are a highly knowledgeable assistant who is always factual and careful. For a {QUESTION} check if the
    {PROPOSED_ANSWER} is True
or False.

{QUESTION}: Who was the first president of the US?
{PROPOSED_ANSWER}: George Washington
Is the Proposed Answer:
a) True
b) False
The proposed Answer is: True

...

{QUESTION}: <query>
{PROPOSED_ANSWER}: <model prediction>
Is the Proposed Answer:
a) True
b) False
The proposed Answer is:
```

Prompt 1: **Few-shot prompt for self-evaluation.**

```
You are a powerful agent that understands probabilities and language. Given a {QUESTION}, {ANSWER} and
    {PROBABILITY_EXPRESSION}, you
```

```
modify the {ANSWER} using the appropriate probability expression. Do not modify the core content of the
    {ANSWER}, and the only
change made is to include the {PROBABILITY_EXPRESSION} to the {ANSWER} to produce the {MODIFIED_ANSWER}.

Here are a few examples:

Example 1:
{QUESTION}: Which English author, who died in 1998, published Katie Hannigan in 1950 and The Bill Bailey
    Trilogy between 1986 and
1997?
{ANSWER}: richard adamson
{PROBABILITY_EXPRESSION}: little chance
{MODIFIED_ANSWER}: There is little chance that the author is Richard Adamson.

Example 2:
{QUESTION}: What chemical element has the symbol S?
{ANSWER}: sulphur
{PROBABILITY_EXPRESSION}: very good chance
{MODIFIED_ANSWER}: There is a very good chance that it is sulphur.

Example 3:
{QUESTION}: What is secreted by the lachrymal glands?
{ANSWER}: tears
{PROBABILITY_EXPRESSION}: certain
{MODIFIED_ANSWER}: It is certainly tears.

Example 4:
{QUESTION}: How many days is the gestation period of a domestic dog (5 days either way)?
{ANSWER}: 58-63
{PROBABILITY_EXPRESSION}: chances are about even
{MODIFIED_ANSWER}: Chances are about even that it is 58-63 days.

Example 5:
{QUESTION}: Which record label signed the Rolling Stones in 1991?
{ANSWER}: umg
{PROBABILITY_EXPRESSION}: i doubt
{MODIFIED_ANSWER}: I doubt that it is umg.

Example 6:
{QUESTION}: What is the capital of India?
{ANSWER}: Delhi.
{PROBABILITY_EXPRESSION}: impossible
{MODIFIED_ANSWER}: It is impossible that it is Delhi.

Example 7:

{QUESTION}: {{THE_QUESTION}}
{ANSWER}: {{THE_ANSWER}}
{PROBABILITY_EXPRESSION}: {{THE_PROBABILITY_EXPRESSION}}
{MODIFIED_ANSWER}:
```

Prompt 2: **Prompt to interleave uncertainty expressions with answers.**

```
You are presented with a statement, which may contain a notion of uncertainty expressed linguistically within
    it. Your task is to
extract the {UNCERTAINTY_PHRASE} separately, and remove the uncertainty component from the answer.

Here is the list of valid uncertainty expressions.

[certain, almost certain, highly likely, very good chance, likely, i believe, probably, probable, chances are
    about even,
i doubt, unlikely, little chance, chances are slight, improbable, highly unlikely, impossible].

List the Answer removed of uncertainty in the {ANSWER_WITHOUT_UNCERTAINTY} field. List the uncertainty
    expression used in the
Uncertainty field.

{ANSWER}: X was certainly not born in 1985.
{ANSWER_WITHOUT_UNCERTAINTY}: X was not born in 1985.
{UNCERTAINTY_PHRASE}: certainly

{ANSWER}: The capital of France almost certainly might be paris.
{ANSWER_WITHOUT_UNCERTAINTY}: The capital of France is Paris.
{UNCERTAINTY_PHRASE}: almost certainly

{ANSWER}: There is little chance but the fact is correct.
{ANSWER_WITHOUT_UNCERTAINTY}: The fact is correct.
{UNCERTAINTY_PHRASE}: little chance

{ANSWER}: It is about even that the coin will be heads.
{ANSWER_WITHOUT_UNCERTAINTY}: The coin will be heads.
{UNCERTAINTY_PHRASE}: about even

{ANSWER}: It is impossible that the Sun rises in the West.
{ANSWER_WITHOUT_UNCERTAINTY}: The Sun rises in the West.
{UNCERTAINTY_PHRASE}: impossible

{ANSWER}: It is highly unlikely that the coin will be heads.
{ANSWER_WITHOUT_UNCERTAINTY}: The coin will be heads.
{UNCERTAINTY_PHRASE}: highly unlikely
```

```
36  {ANSWER}: There is a very good chance that tomato might not be a vegetable.
37  {ANSWER_WITHOUT_UNCERTAINTY}: Tomato is not a vegetable.
38  {UNCERTAINTY_PHRASE}: very good chance
39
40  Here is the {ANSWER} which needs to be separated into {ANSWER_WITHOUT_UNCERTAINTY}, and {UNCERTAINTY_PHRASE}.
41
42  {ANSWER}:
```

Prompt 3: **Prompt to disentangle uncertainty statements from answers.**

```
1   Your task is to determine if two answers to a question are semantically equivalent. Two answers are
        semantically equivalent if their
2   answer to the question is the same, even if they are rephrases of each other.
3
4   Even if one contains more information than the other, as long as their answer to the question is the same,
        the answers are considered
5   semantically equivalent.
6
7   For example, for the question 'Tell me a number', the answers 'five' and '6' are not semantically equivalent
        as they are different
8   numbers. However, the answers '5' and 'A number is five' or 'five' are semantically equivalent, since they
        convey the same thing.
9
10  For a question 'Tell me the capital of France', 'Its Venice' and 'Venice' are semantically equivalent, as
        they give the same answer.
11  The Answers 'Paris' and 'Venice' are not semantically equivalent. For questions asking factual information,
        if one answer responds
12  with information disagreeing with the other, they will not be semantically equivalent.
13
14  Another example of not semantically equivalent answers would be 'Miles' and 'Kilometers' for the question
        'Tell me an unit of
15  distance', as these two units are not the same.
16
17  Here is the question, and Answer A and Answer B to be compared.
18
19  Respond with 'YES' if they are semantically equivalent, 'NO' otherwise.
20
21  Question: {THE_QUESTION}
22  Answer A: {GOLD_ANSWER}
23  Answer B: {PRED_ANSWER}
24  Semantically equivalent:
```

Prompt 4: **Language model evaluation.**

Model	Dataset	Max Examples Per Bin	Train Set Size	Evaluation Set size
Small	TriviaQA	2000	12556	7993
	AmbigQA	1000	8935	898
	TruthfulQA	-	-	-
Small - IT	TriviaQA	2000	11571	7993
	AmbigQA	1000	8487	898
	TruthfulQA	200	581	100
Medium	TriviaQA	2000	9283	7993
	AmbigQA	1000	4800	898
	TruthfulQA	200	440	100
Medium - IT	TriviaQA	2000	7590	7993
	AmbigQA	1000	3969	898
	TruthfulQA	200	518	100

Table 2: **Statistics of the datasets used for finetuning** $\mathcal{D}_{tr}^{M}$

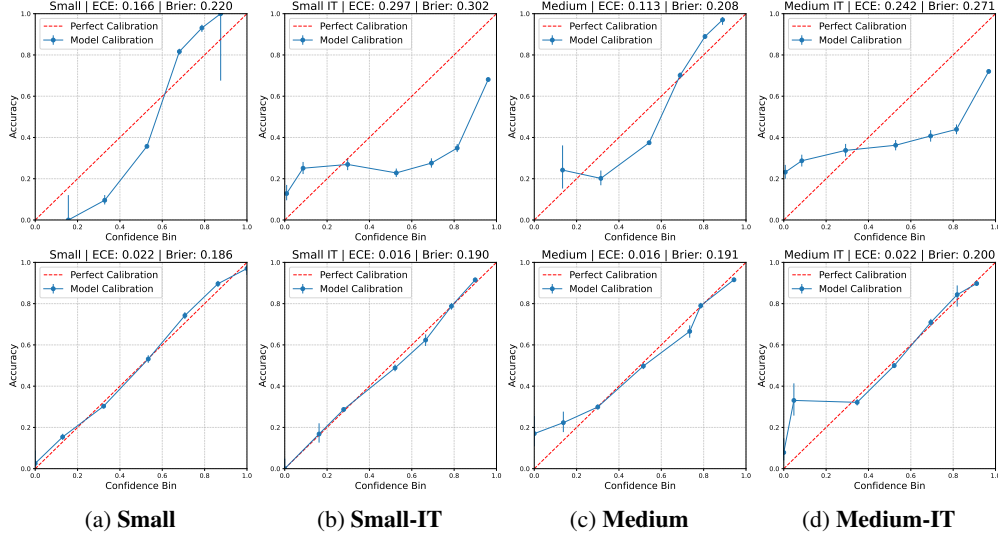(a) **Small**   (b) **Small-IT**   (c) **Medium**   (d) **Medium-IT**

Figure 7: **AmbigQA Calibration Chart**: The **top-row** shows raw calibration scores at temperature=1.0 without any post-processing. The **bottom row** shows post-processed calibration scores with isotonic regression. In each plot, the x-axis is the $p_{model}(true)$ of the generated prediction (shown here as `Confidence Bin`) and y-axis is probability of that prediction being actually correct (shown here as `Accuracy`). Expected Calibration Error (ECE) and Brier Score are reported at the top of each plot.

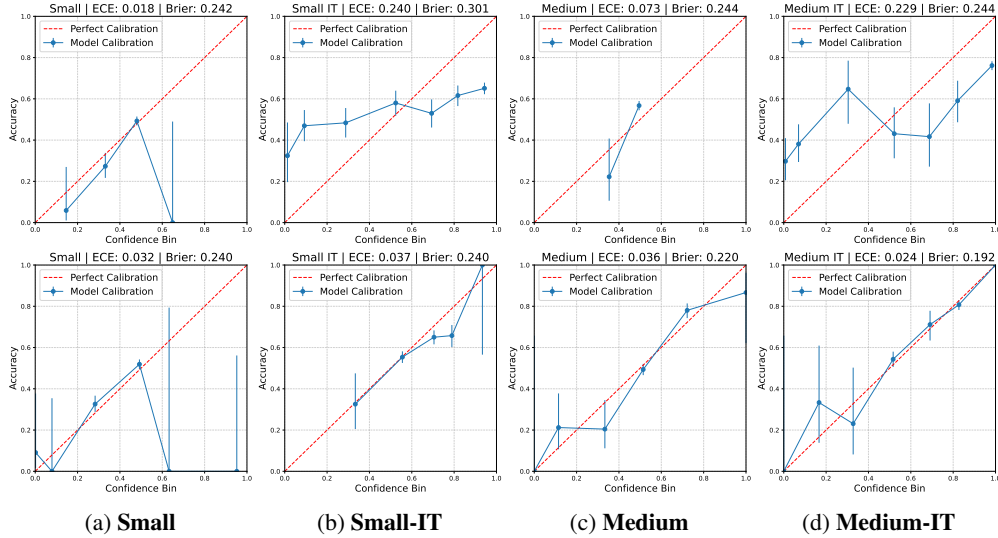(a) **Small**   (b) **Small-IT**   (c) **Medium**   (d) **Medium-IT**

Figure 8: **TruthfulQA Calibration Chart**: The **top-row** shows raw calibration scores at temperature=1.0 without any post-processing. The **bottom row** shows post-processed calibration scores with isotonic regression. In each plot, the x-axis is the $p_{model}(true)$ of the generated prediction (shown here as `Confidence Bin`) and y-axis is probability of that prediction being actually correct (shown here as `Accuracy`). Expected Calibration Error (ECE) and Brier Score are reported at the top of each plot.

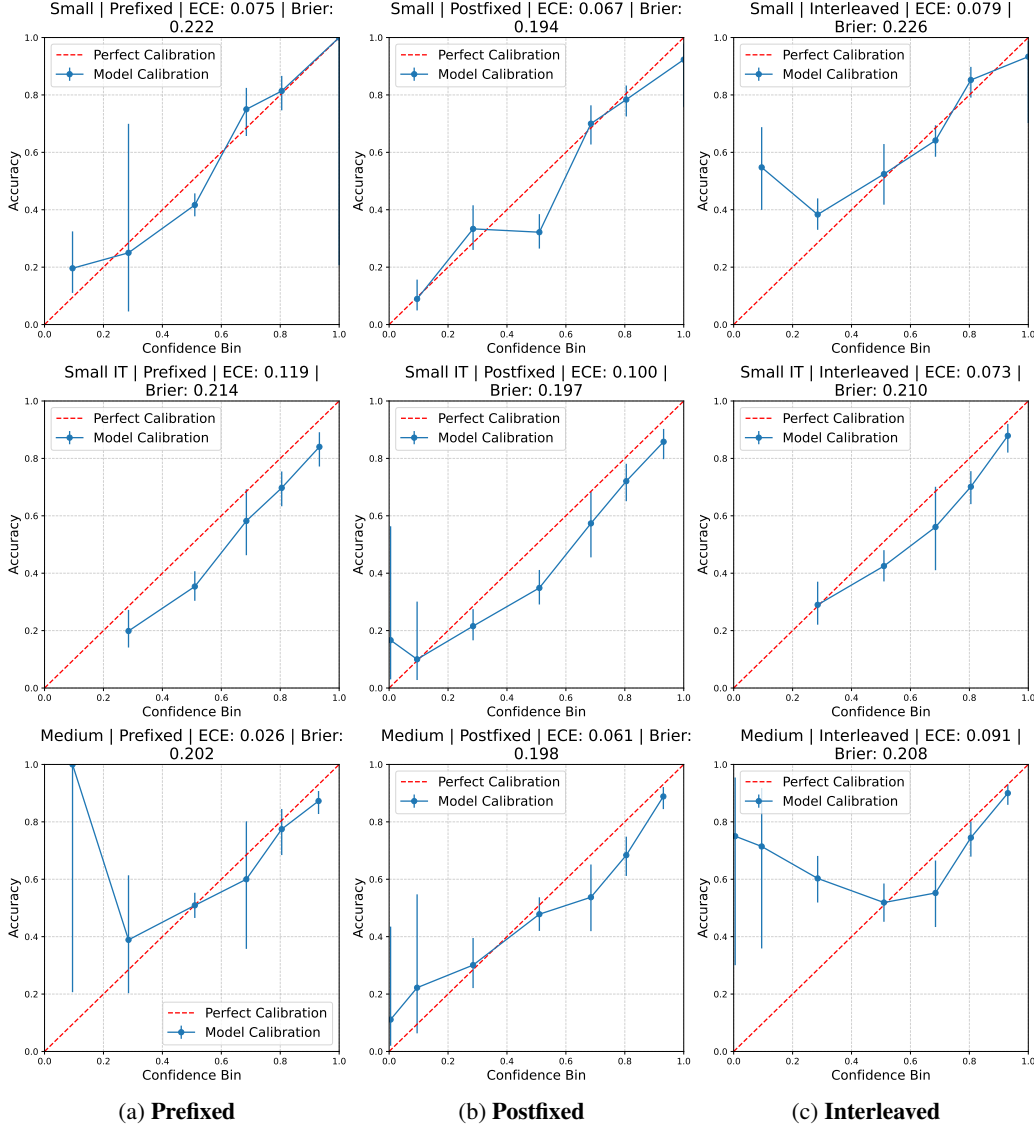(a) **Prefixed**   (b) **Postfixed**   (c) **Interleaved**

Figure 9: **AmbigQA uncertainty augmentation method**: Top-row is **Gemini Small**, middle-row is **Gemini Small-IT** and the bottom row is **Gemini Medium**. The models generate **post-fixed** uncertainty expressions. The x-axis is the $p_{model}(true)$ obtained by converting the linguistic expression of uncertainty to a float using **Table 1** (shown here as `Confidence Bin`) and y-axis is probability of that prediction being actually correct (shown here as `Accuracy`). No post-processing is done on the $p_{model}(true)$. Expected Calibration Error (ECE) and Brier Score are reported at the top of each plot. The error bars show the variance of accuracy in each bin.