Enhancing Healthcare LLM Trust with Atypical Presentations Recalibration

Jeremy Qin[1,2], **Bang Liu**[1,2,4], **Quoc Dinh Nguyen**[1,3]

[1]Université de Montréal [2]Mila [3]CRCHUM [4]Canada CIFAR AI Chair
{jeremy.qin@ bang.liu@ quoc.dinh.nguyen@}umontreal.ca

Abstract

Black-box large language models (LLMs) are increasingly deployed in various environments, making it essential for these models to effectively convey their confidence and uncertainty, especially in high-stakes settings. However, these models often exhibit overconfidence, leading to potential risks and misjudgments. Existing techniques for eliciting and calibrating LLM confidence have primarily focused on general reasoning datasets, yielding only modest improvements. Accurate calibration is crucial for informed decision-making and preventing adverse outcomes but remains challenging due to the complexity and variability of tasks these models perform. In this work, we investigate the miscalibration behavior of black-box and open-source LLMs within the healthcare setting. We propose a novel method, *Atypical Presentations Recalibration*, which leverages atypical presentations to adjust the model's confidence estimates. Our approach significantly improves calibration, reducing calibration errors by approximately 60% on three medical question answering datasets and outperforming existing methods such as vanilla verbalized confidence, CoT verbalized confidence and others. Additionally, we provide an in-depth analysis of the role of atypicality within the recalibration framework. The code can be found at `https://github.com/jeremy-qin/medical_confidence_elicitation`

1 Introduction

Despite recent successes and innovations in large language models (LLMs), their translational value in high-stakes environments, such as healthcare, has not been fully realized. This is primarily due to concerns about the trustworthiness and transparency of these models, stemming from their complex architecture and black-box nature. Recent studies (Xiong et al., 2024; Shrivastava et al., 2023; Tian et al., 2023; He et al., 2023; Rivera et al., 2024; Chen and Mueller, 2023) have begun to explore methods for eliciting confidence and uncertainty estimates from these models in order to enhance trustworthiness and transparency. The ability to convey uncertainty and confidence is central to clinical medicine (Banerji et al., 2023) and plays a crucial role in facilitating rational and informed decision-making. This underscores the importance of investigating and utilizing calibrated confidence estimates for the medical domain.

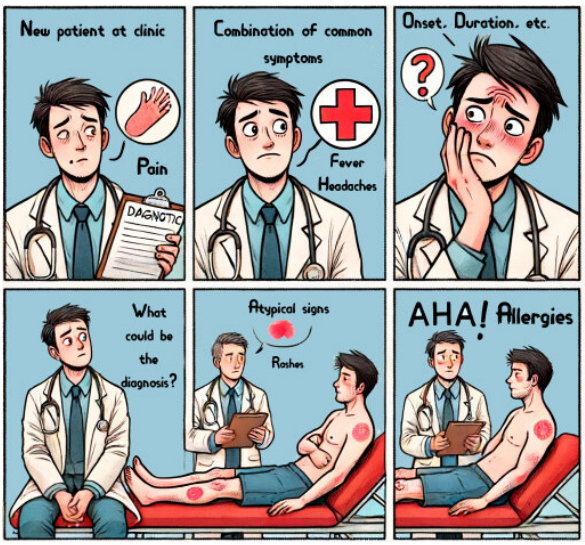

Figure 1: A physician diagnoses a patient who returned from a camping trip, presenting a combination of common symptoms and signs like fever and headaches. However, by recognizing rashes an atypical symptom, the physician ultimately identifies the condition as an allergy.

Previous work on confidence elicitation and calibration of large language models (LLMs) has mainly focused on general reasoning and general knowledge datasets for tasks such as logical reasoning, commonsense reasoning, mathematical reasoning, and scientific knowledge (Kuhn et al., 2023; Xiong et al., 2024; Tian et al., 2023; Tanneru et al.,

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2520–2537

2023; Chen and Mueller, 2023). Few studies have investigated tasks that require expert knowledge, and these have shown considerable room for improvement. Moreover, with the success of many closed-source LLMs, such as GPT-3.5 and GPT-4, which do not allow access to token-likelihoods and text embeddings, it has become prevalent to develop tailored methods for eliciting confidence estimates. However, most approaches developed consist of general prompting and sampling strategies without using domain-specific characteristics.

Traditionally, clinicians are taught to recognize and diagnose typical presentations of common illnesses based on patient demographics, symptoms and signs, test results, and other standard indicators (Harada et al., 2024). However, the frequent occurrence of atypical presentations is often overlooked (Goldrich and Shah, 2021). Failing to identify atypical presentations can result in worse outcomes, missed diagnoses, and lost opportunities for treating common conditions. Thus, awareness of atypical presentations in clinical practice is fundamental to providing high-quality care and making informed decisions. Figure 1 depicts a simplistic example of how atypicality plays a role in diagnosis. Incorporating the concept of atypicality has been shown to improve uncertainty quantification and model performance for discriminative neural networks and white-box large language models (Yuksekgonul et al., 2023). This underscores the importance of leveraging atypical presentations to enhance the calibration of LLMs, particularly in high-stakes environments like healthcare.

Our study aims to address these gaps by first investigating the miscalibration of black-box LLMs when answering medical questions using non-logit-based uncertainty quantification methods. We begin by testing various baseline methods to benchmark the calibration of these models across a range of medical question-answering datasets. This benchmarking provides a comprehensive understanding of the current state of calibration in LLMs within the healthcare domain and highlights the limitations of existing approaches.

Next, we propose a new recalibration framework based on the concept of atypicality, termed **Atypical Presentations Recalibration**. This method leverages atypical presentations to adjust the model's confidence estimates, making them more accurate and reliable. Under this framework, we construct two distinct atypicality-aware prompting strategies for the LLMs, encouraging them to consider and reason over atypical cases explicitly. We then compare the performance and calibration of these strategies against the baseline methods to evaluate their effectiveness.

Finally, our empirical results reveal several key findings. First, black-box LLMs often fail to provide calibrated confidence estimates when answering medical questions and tend to remain overconfident. Second, our proposed Atypical Presentations Aware Recalibration method significantly improves calibration, reducing calibration errors by approximately 60% on three medical question answering datasets and consistently outperforming existing baseline methods across all datasets. Third, we observe that atypicality interacts in a complex manner with both performance and calibration, suggesting that considering atypical presentations is crucial for developing more accurate and trustworthy LLMs in healthcare settings. Finally, we show that our framework is generalizable to open-source models with noteworthy improvements in confidence calibration.

2 Background and Related Work

2.1 Confidence and Uncertainty quantification in LLMs

Confidence and uncertainty quantification is a well-established field, but the recent emergence of large language models (LLMs) has introduced new challenges and opportunities. Although studies have shown a distinction between confidence and uncertainty, we will use these terms interchangeably in our work.

Research on this topic can be broadly categorized into two areas: approaches targeting closed-source models and those focusing on open-source models. The growing applications of commercial LLMs, due to their ease of use, have necessitated particular methods to quantify their confidence. For black-box LLMs, a natural approach is to prompt them to express confidence verbally, a method known as verbalized confidence, first introduced by Lin et al. (2022). Other studies have explored this approach specifically for language models fine-tuned with reinforcement learning from human feedback (RLHF) (Tian et al., 2023; He et al., 2023). Additionally, some research has proposed new metrics to quantify uncertainty (Ye et al., 2024; Tanneru et al., 2023).

Our work aligns most closely with Xiong et al. (2024), who presented a framework that combines

prompting strategies, sampling techniques, and aggregation methods to elicit calibrated confidences from LLMs. While previous studies primarily benchmarked their methods on general reasoning tasks, our study focuses on the medical domain, where accurate uncertainty quantification is critical for diagnosis and decision-making. We evaluate LLM calibration using the framework defined by Xiong et al. (2024) and propose a framework consisting of a new prompting strategy and aggregation method, termed *Atypicality Presentations Recalibration*, which shows significant improvements in calibrating LLM uncertainty in the medical domain.

2.2 Atypical Presentations

Atypical presentations have garnered increasing attention and recognition in the medical field due to their critical role in reducing diagnostic errors and enhancing problem-based learning in medical education (Vonnes and El-Rady, 2021; Kostopoulou et al., 2008; Matulis et al., 2020; Bai et al., 2023). Atypical presentations are defined as "a shortage of prototypical features most frequently encountered in patients with the disease, features encountered in advanced stages of the disease, or features commonly listed in medical textbooks" (Kostopoulou et al., 2008; Harada et al., 2024). This concept is particularly important in geriatrics, where older patients often present atypically, and in medical education, where it prompts students to engage in deeper reflection during diagnosis.

Given the increasing emphasis on atypical presentations in medical decision-making, it is pertinent to explore whether this concept can be leveraged to calibrate machine learning models. Yuksekgonul et al. (2023) were the first to incorporate atypicality into model calibration for classification tasks. Our work extends this approach to generative models like LLMs, integrating atypical presentations to achieve more accurate and calibrated confidence estimates.

3 Method

In this section, we describe the methods used to elicit confidence from large language models (LLMs) as well as our recalibration methods. Calibration in our settings refers to the alignment between the confidence estimates and the true likelihood of outcomes (Yuksekgonul et al., 2023; Gneiting and Raftery, 2007). Our experiments are based on the framework described by Xiong et al. (2024), which divides the approaches into three main components: prompting, sampling, and aggregation, and uses it as baselines. In their framework, they leverage common prompting strategies such as vanilla prompting and Chain-of-Thoughts while also leveraging the stochasticity of LLMs. In contrast, we propose an approach, *Atypical Presentation Recalibration*, that retrieves atypicality scores and use them as a recalibration method in order to have more accurate confidence estimates. Our framework is mainly divided into two parts: *Atypicality Prompting* and *Atypicality Recalibration*. We explain how each of the three components are applied to our tasks and how we integrate atypicality to develop hybrid methods that combine these elements.

3.1 Prompting methods

Eliciting confidence from LLMs can be achieved through various methods, including natural language expressions, visual representations, and numerical scores (Kim et al., 2024). We refer to these methods collectively as verbalized confidence. While there are trade-offs between these methods, we focus on retrieving numerical confidence estimates for better precision and ease of calibration. We design a set of prompts to elicit confidence estimates from LLMs.

Vanilla Prompting The most straightforward way to elicit confidence scores from LLMs is to ask the model to provide a confidence score on a certain scale. We term this method as vanilla prompting. This score is then used to assess calibration.

Chain-of-Thought (CoT) Eliciting intermediate and multi-step reasoning through simple prompting has shown improvements in various LLM tasks. By allowing for more reflection and reasoning, this method helps the model express a more informed confidence estimate. We use zero-shot Chain-of-Thought (CoT) (Kojima et al., 2023) in our study.

Atypicality Prompting Inspired by the concept of atypical presentations in medicine, we aim to enhance the reliability and transparency of LLM decision-making by incorporating atypicality into the confidence estimation process. We develop two distinct prompting strategies to achieve this goal:

- **Atypical Presentations Prompt:** This strategy focuses on identifying and highlighting atypical symptoms and features within the

Method	Prompt Template
Vanilla	Read the following question. Provide your answer and your confidence level (0% to 100%).
Atypical Scenario	Read the following question. Assess the atypicality of the scenario described with a score between 0 and 1 with 0 being highly atypical and 1 being typical. Provide your answer, atypicality score and confidence level.
Atypical Presentations	Read the following question. Assess each symptom and signs with respect to its typicality in the described scenario with a score between 0 and 1 with 0 being highly atypical and 1 being typical. Provide your answer, atypicality scores and confidence level.

Table 1: Illustrations of the vanilla prompting and Atypical Presentations Aware Recalibration prompting strategies (complete prompts in Appendix B)

medical data. The prompt is designed to guide the LLM to assess the typicality of each symptom presented in the question. By systematically evaluating which symptoms are atypical, the model can better gauge the uncertainty associated with the diagnosis. For example, the prompt might ask the model to rate the typicality of each symptom on a scale from 0 to 1, where 1 represents a typical symptom and 0 represents an atypical symptom. In the following sections of the paper, we will refer to these scores as atypicality scores where the lower the score is the more atypical it is. This information is then used to adjust the confidence score accordingly.

- **Atypical Scenario Prompt:** This strategy evaluates the typicality of the question itself. It is based on the notion that questions which are less familiar or more complex may naturally elicit higher uncertainty. The prompt asks the LLM to consider how common or typical the given medical scenario is. For instance, the model might be prompted to rate the overall typicality of the scenario on a similar scale. This approach helps to capture the inherent uncertainty in less familiar or more complex questions.

3.2 Sampling and Aggregation

While verbalized confidences provide a straightforward way to assess the uncertainty of LLMs, we can also leverage the stochasticity of LLMs (Xiong et al., 2024; Rivera et al., 2024) by generating multiple answers for the same question. Different aggregation strategies can then be used to evaluate how aligned these sampled answers are. We follow the framework defined by Xiong et al. (2024) for the sampling and aggregation methods and uses them as baselines to our *Atypical Presentations Recalibration* framework.

Self-Random Sampling The simplest strategy to generate multiple answers from an LLM is by repeatedly asking the same question and collecting the responses. These responses are then aggregated to produce a final confidence estimate.

Consistency We use the consistency of agreement between different answers from the LLM as the final confidence estimate (Xiong et al., 2024). For a given question with a reference answer $\tilde{Y}$, we generate a sample of answers $\hat{Y}_k$. The aggregated confidence $C_{consistency}$ is defined as:

$$C_{consistency} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{\hat{Y}_k = \tilde{Y}\} \quad (1)$$

Weighted Average Building on the consistency aggregation method, we can use a weighting mechanism that incorporates the confidence scores elicited from the LLM. This method weights the agreement between the different answers by their respective confidence scores. The aggregated confidence $C_{average}$ is defined as:

$$C_{average} = \frac{\sum_{k=1}^{K} \mathbb{1}\{\hat{Y}_k = \tilde{Y}\} * C_k}{\sum_{k=1}^{K} C_k} \quad (2)$$

Atypicality Recalibration To integrate the atypicality scores elicited with *Atypicality Presentations Prompting* into our confidence estimation framework, we propose a non-linear post-hoc recalibration method that combines the initial confidence score with an aggregation of the atypicality assessments. This method draws inspiration from

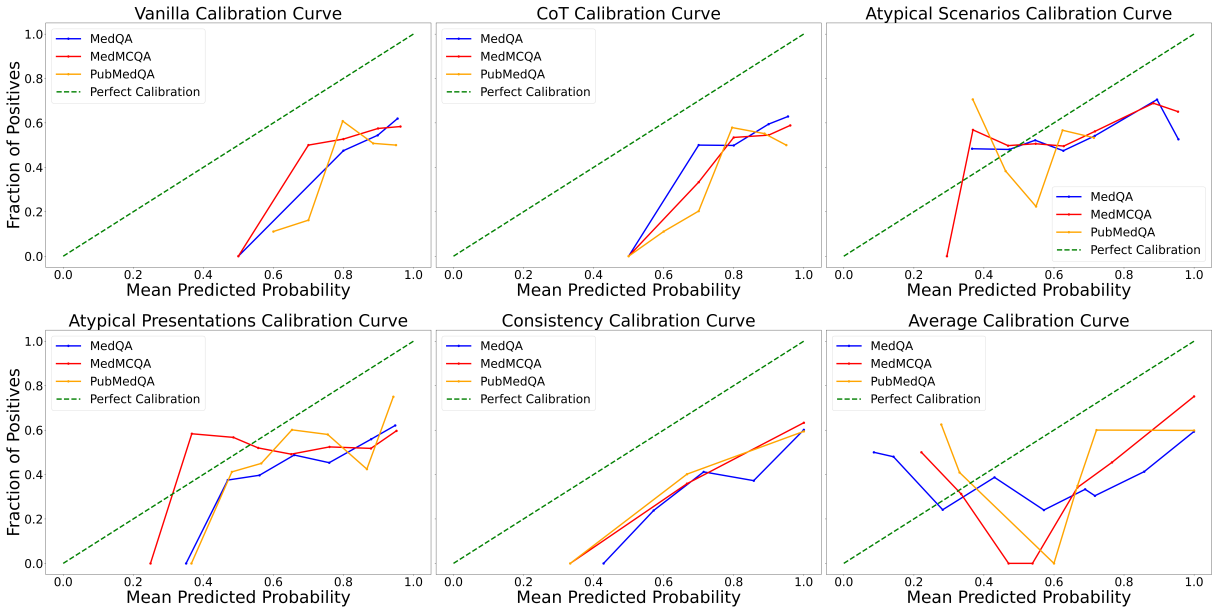

Figure 2: Calibration Curves of the different methods for GPT-3.5-turbo

economic and financial models where expert judgments are combined with varying weights and exponential utility functions to address risk aversion. Formally, for an initial confidence C_i of a given question and atypical scores A_k, the calibrated confidence CC_i is computed as follows:

$$CC_i = C_i * \left(\frac{1}{K} \sum_{k=1}^{K} e^{A_k - 1} \right) \quad (3)$$

where A_k takes values in [0,1] and a value of 1 corresponds to a typical value. For the Atypical Scenario Prompt, this equation translates to having K equal to 1. Thus, the final confidence estimate will equal the initial confidence score only if all the atypical scores are 1.

4 Experiments

4.1 Experimental Setup

Datasets Our experiments evaluate the calibration of confidence estimates across three different english medical question-answering datasets. For our experiments, we restricted on evaluating on only the development set of each dataset. **MedQA** (Jin et al., 2020) consists of 1272 questions based on the United States Medical License Exams and collected from the professional medical board exams. **MedMCQA** (Pal et al., 2022) is a large-scale multiple-choice question answering dataset with 2816 questions collected from AIIMS & NEET PG entrance exams covering a wide variety of healthcare topics and medical subjects. **PubMedQA** (Jin et al., 2019) is a biomedical question answering dataset with 500 samples collected from PubMed abstracts where the task is to answer research question corresponding to an abstract with yes/no/maybe.

Models We use a variety of commercial LLMs that includes GPT-3.5-turbo (OpenAI, 2023), GPT-4-turbo (OpenAI, 2024), Claude3-sonnet (Anthropic, 2023) and Gemini 1.0 Pro (DeepMind, 2023). To demonstrate that our framework can generalize to other models other than black-box LLMs, we also experiment on Llama3.1 8B (Llama, 2024) and Qwen2.5 7B (Qwen, 2024).

Evaluation Metrics To measure how well the confidence estimates are calibrated, we will report multiple metrics across the different datasets, methods and models. Calibration is defined as how well a model's predicted probability is aligned with the true likelihoods of outcomes (Yuksekgonul et al., 2023; Gneiting and Raftery, 2007). We measure this using *Expected Calibration Error (ECE)* (Naeini et al., 2015) and *Brier Score* (Goldrich and Shah, 2021).

To evaluate the quality of confidence estimates using ECE, we group the model's confidence into K bins and estimate ECE by taking the weighted average of the difference between confidence and accuracy in each bin (He et al., 2023). Formally, let

		MedQA (n=1272)			MedMCQA (n=2816)			PubMedQA (n=500)		
Models	Methods	Acc	ECE	Brier	Acc	ECE	Brier	Acc	ECE	Brier
gpt-3.5-turbo	Vanilla	0.526	0.351	0.363	0.555	0.323	0.350	0.544	0.251	0.304
	CoT	0.536	0.318	0.334	0.525	0.357	0.360	0.516	0.275	0.360
	Atypical scenario	0.506	**0.084**	**0.262**	0.544	**0.128**	**0.252**	0.468	**0.115**	**0.252**
	Atypical presentations	0.506	0.283	0.332	0.527	0.152	0.322	0.544	0.129	0.268
	Consistency (k=3)	0.535	0.408	0.396	**0.561**	0.350	0.356	0.544	0.335	0.370
	Average (k=3)	**0.539**	0.398	0.397	**0.561**	0.350	0.344	**0.550**	0.346	0.372
claude3-sonnet	Vanilla	0.541	0.331	0.336	0.565	0.306	0.327	0.128	0.569	0.428
	CoT	**0.638**	0.246	0.282	**0.612**	0.265	0.295	**0.246**	0.542	0.469
	Atypical scenario	0.564	**0.124**	**0.259**	0.561	**0.134**	**0.268**	0.140	**0.438**	**0.252**
	Atypical presentations	0.568	0.136	0.332	0.531	0.305	0.316	0.100	0.517	0.339
	Consistency (k=3)	0.552	0.335	0.363	0.568	0.346	0.355	0.122	0.789	0.766
	Average (k=3)	0.558	0.338	0.358	0.568	0.337	0.356	0.128	0.750	0.725
gemini-pro-1.0	Vanilla	0.472	0.369	0.385	0.551	0.297	0.338	0.492	0.342	0.362
	CoT	0.465	0.357	0.369	0.526	0.306	0.340	0.438	0.368	0.373
	Atypical scenario	0.473	**0.105**	**0.268**	0.513	**0.129**	**0.274**	0.508	**0.128**	**0.276**
	Atypical presentations	0.458	0.293	0.332	0.387	0.357	0.316	0.226	0.448	0.338
	Consistency (k=3)	0.471	0.399	0.400	**0.557**	0.337	0.343	**0.540**	0.309	0.349
	Average (k=3)	**0.477**	0.364	0.391	0.549	0.314	0.341	0.504	0.325	0.371
gpt-4-turbo	Vanilla	0.756	0.133	0.190	0.707	0.188	0.230	0.394	0.374	0.337
	CoT	**0.832**	**0.065**	**0.132**	**0.730**	0.162	**0.206**	0.358	0.445	0.402
	Atypical scenario	0.741	0.085	0.181	0.675	**0.071**	**0.206**	0.354	**0.197**	**0.249**
	Atypical presentations	0.751	0.114	0.178	0.681	0.174	0.213	0.338	0.414	0.363
	Consistency (k=3)	0.775	0.198	0.206	0.712	0.248	0.253	**0.404**	0.537	0.546
	Average (k=3)	0.767	0.194	0.205	0.708	0.249	0.255	**0.404**	0.552	0.563

Table 2: Using atypicality as post-hoc calibration brings major improvements in ECE and Brier Scores across all datasets and all models. **Atypical Scenario** outperforms all other methods in calibration in the big majority of experiments.

N be the sample size, K the number of bins, and I_k the indices of samples in the k^{th} bin, we have:

$$ECE_K = \sum_{k=1}^{K} \frac{|I_k|}{N} |acc(I_k) - conf(I_k)| \quad (4)$$

Brier score is a scoring function that measures the accuracy of the predicted confidence estimates and is equivalent to the mean squared error. Formally, it is defined as:

$$BS = \frac{1}{N} \sum_{n=1}^{N} (conf_n - o_n)^2 \quad (5)$$

where $conf_n$ and o_n are the confidence estimate and outcome of the n^{th} sample respectively.

Additionally, to evaluate if the LLM can convey higher confidence scores for correct predictions and lower confidence scores for incorrect predictions, we use the *Area Under the Receiver Operating Characteristic Curve (AUROC)*. Finally, to assess any significant changes in performance, we also report *accuracy* on the different tasks.

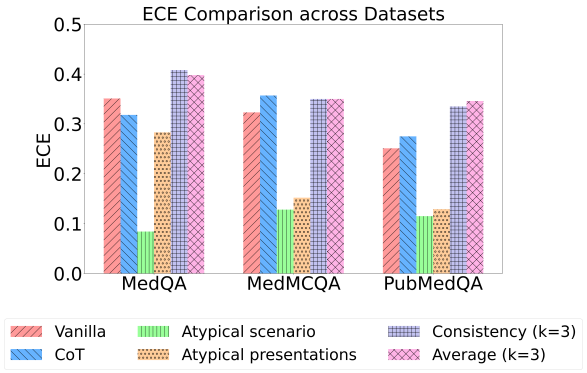

Figure 3: ECE of GPT-3.5-turbo for each method across all three datasets.

4.2 Results and Analysis

To assess the ability of LLMs to provide calibrated confidence scores and explore the use of atypical scores for calibration, we experimented with each mentioned method using four different black-box LLMs across three medical question-answering datasets. The main results and findings are reported in the following section.

		MedQA			MedMCQA			PubMedQA		
Models	Methods	Acc	ECE	Brier	Acc	ECE	Brier	Acc	ECE	Brier
Llama3.1	Vanilla	0.431	0.351	0.304	0.494	0.334	0.320	0.504	0.301	0.316
	Atypical scenario	0.454	**0.172**	**0.229**	0.378	**0.226**	**0.230**	0.388	**0.203**	**0.257**
Qwen2.5	Vanilla	0.510	0.384	0.388	0.554	0.353	0.363	0.464	0.403	0.392
	Atypical scenario	0.452	**0.324**	**0.349**	0.546	**0.303**	**0.329**	0.446	**0.305**	**0.348**

Table 3: Atypicaity recalibration generalizes well on open-source models with notable improvements in calibration errors across all datasets.

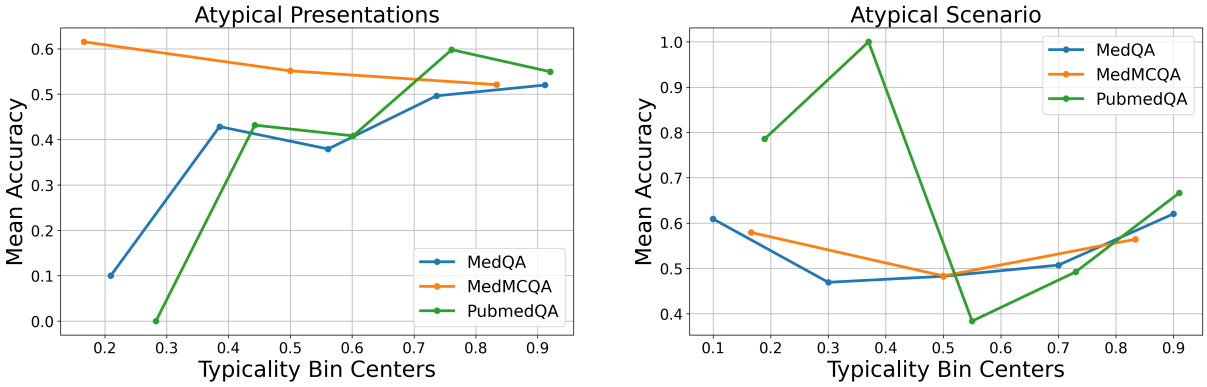

Figure 4: Accuracy by Typicality bins of GPT3.5-turbo for Atypical Presentations Aware Recalibration methods.

LLMs are miscalibrated for Medical QA. To evaluate the reliability and calibration of confidence scores elicited by the LLMs, we examined the calibration curves of GPT-3.5-turbo in Figure 2, where the green dotted line represents perfect calibration. The results indicate that the confidence scores are generally miscalibrated, with the LLMs tending to be overconfident. Although the *Atypical Scenario* and *Atypical Presentations* methods show improvements with better alignment, there is still room for improvement. Introducing recalibration methods with atypicality scores results in more variation in the calibration curves, including instances of underestimation. Additional calibration curves for the other models are provided in Appendix A.

Leveraging atypical scores greatly improves calibration. We analyzed the calibration metrics for each method and found that leveraging atypical scores significantly reduces ECE and Brier Score across all datasets, as shown in Figure 3 and Table 3. We also observe that this improvement is translated to open-source models which demonstrate the applicability of atypicality prompting across both closed and open source LLMS. In contrast, other methods show minor changes in calibration errors, with some even increasing ECE. The *Consistency* and *Average* methods do not show improvement, and sometimes degrade, due to the multiple-choice format of the datasets, which shifts confidence estimates to higher, more overconfident values. However, the *Atypical Scenario* method, which elicits an atypical score describing how unusual the medical scenario is, outperforms all other methods and significantly lowers ECE compared to vanilla confidence scores. We also note that *Atypical Scenario* outperformance translates to open-source models as well with notable improvements across all datasets and models. It is very interesting that the level of atypicality considered seems to make a significant difference. It is a hallmark of reasoning that how the LLM aggregates the atypicality from a lower level when prompted for a scenario is superior to simply aggregating the symptoms atypicality. This opens for further investigation into how LLMs reason about atypicality. We discuss and analyze the role of atypicality in calibration in the following sections. Detailed results of our experiments are reported in Table 3.

Atypicality distribution varies between Atypical Scenario and Atypical Presentations. To better understand the gap between the calibration errors of *Atypical Scenario* and *Atypical Presentations*, we first examine the distribution of the atypicality scores. In Figure 6, we observe that the distribution of *Atypical Presentations* is much more right-

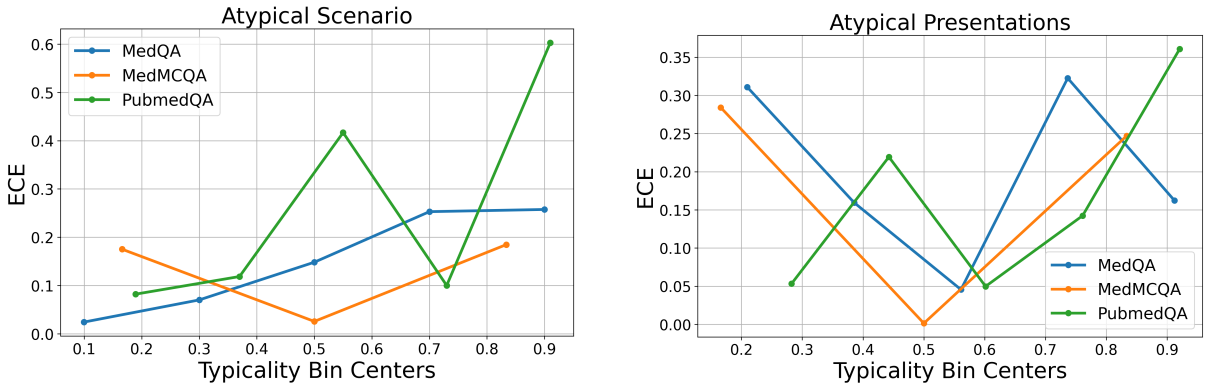

Figure 5: ECE by Typicality bins of GPT3.5-turbo for Atypical Presentations Aware Recalibration methods.

skewed, indicating a prevalence of typical scores. This is largely due to the nature of the approach. Not all questions in the datasets are necessarily diagnostic questions; for example, some may ask for medical advice or some may only touch the specific paper it is referencing (PubmedQA), where there is no atypicality associated with symptoms or presentations. In our framework, we impute the atypicality score to 1 for such cases, so it does not affect the original confidence estimate. In contrast, *Atypical Scenario* shows a more evenly distributed spread over the scores. This suggests that the LLMs can identify that some questions and scenarios are more atypical, which allows this atypicality to be considered when calibrating the confidence estimates.

Typical samples do not consistently outperform atypical samples. We now question the performance of atypical versus typical samples. The intuitive answer is that performance should be better on typical samples, which are common scenarios or symptoms, making the question easier to answer. However, as shown in Figure 4, there is no consistent pattern between accuracy and atypicality for GPT-3.5-turbo. While accuracy increases as atypicality decreases in some cases like MedQA and PubMedQA, in other cases, the accuracy remains unchanged or even decreases. This performance variation across typicality bins provides insights into how LLMs use the notion of atypicality in their reasoning process. Higher accuracy for atypical samples could suggest that unique, easily identifiable features help the LLM. Conversely, high atypicality can indicate that the question is more difficult, leading to lower accuracy. To understand this better, we also experimented with prompts to retrieve difficulty scores and analyzed their relationship with atypicality. Our results show no clear correlation between difficulty and atypicality scores. Most atypicality scores are relatively high across all difficulty levels. Although some atypical samples are deemed more difficult, the results are inconsistent and hard to interpret. Associated graphs are in Appendix A. Briefly, this inconsistent performance behavior shows there is more to explore about how LLMs use atypicality intrinsically.

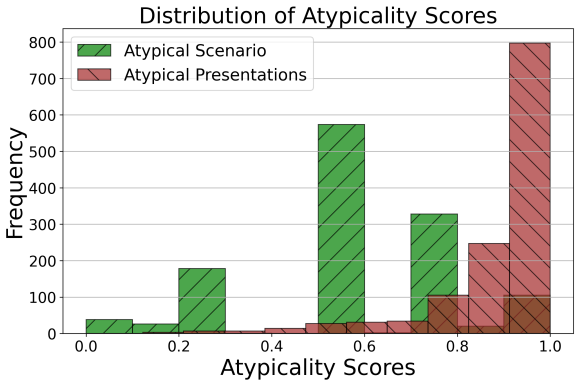

Figure 6: Distribution of atypicality scores between Atypical Presentations and Atypical Scenario of GPT-3.5-turbo on MedQA.

Atypicality does not predict LLM's calibration error. Another question we explored was whether calibration errors correlate with atypicality. We used the same approach as our performance analysis, binning the samples by atypicality scores and examining the ECE within each bin. This allowed us to evaluate how well the model's predicted confidence level aligned with actual outcomes across varying levels of atypicality. For both *Atypical Scenario* and *Atypical Presentations*, we assessed GPT-3.5-turbo's calibration. As shown in Figure 4, there are no clear patterns between atypi-

cality scores and calibration errors. The high fluctuation of ECE across different levels of atypicality suggests that the model experiences high calibration errors for both typical and atypical samples. This indicates that calibration performance is influenced by factors beyond just atypicality. Similar to the previous performance analysis in terms of accuracy, how LLMs interpret and leverage atypicality may vary between samples, leading to inconsistent behavior.

Atypicality helps in failure prediction. While ECE and Brier Score provide insights into the reliability and calibration of confidence estimates, it is also important for the model to assign higher confidences to correct predictions and lower confidences to incorrect predictions. To assess this, we used AUROC. In Table 4, we observe that incorporating atypicality into our model improves its performance across most experiments compared to the vanilla baseline. However, these improvements do not consistently outperform all other methods evaluated. This indicates that, while incorporating atypicality can improve the model's failure prediction, there remain specific scenarios where alternative methods may be more effective. This also indicates that our framework could be improved to take more into consideration failure prediction to maybe have a multi-objective method.

5 Conclusion

In our study, we have demonstrated that LLMs remain miscalibrated and overconfident in the medical domain. Our results indicate that incorporating the notion of atypicality when eliciting LLM confidence leads to significant gains in calibration and some improvement in failure prediction for medical QA tasks. This finding opens the door to further investigate the calibration of LLMs in other high-stakes domains. Additionally, it motivates the development of methods that leverage important domain-specific notions and adapting our method for white-box LLMs. We hope that our work can inspire others to tackle these challenges and to develop methods for more trustworthy, explainable and transparent models, which are becoming increasingly urgent.

Acknowledgements

This work is supported by the Canada CIFAR AI Chair Program and the Canada NSERC Discovery Grant (RGPIN-2021-03115). Quoc Dinh Nguyen is funded by a research grant from the Canadian Institutes of Health Research (Funding Reference Number: 180570), the Fonds de recherche Québec - Santé, and the Saputo Foundation.

Limitations This study present a first effort into assessing black-box LLMs calibration and the use of atypicality in the healthcare domain. Several aspects of the study can further be improved for a better assessment. While we restricted ourselves to three modical question-answering datasets, we can expand it to more datasets with questions that are more open-ended or even different tasks such as clinical notes summarization which could also benefit a lot from having trusted confidence estimates. While our method generalizes to both closed and open source models like GPT4 and Llama, it does not leverage information that we can potentially get from open source models such as the internal representations, token embeddings and token-likelihoods. An intereting future work could focus on how atypicality is represented internally in a LLM, and if we can potentially read and control its representations of atypicality for better calibration. Interpretability approaches such as mechanistic interpretability or representation engineering are interesting directions to consider. Morever, our approach is still dependent on a prompt, and since LLMs are quite sensitive to how we prompt them, there could be even more optimal prompts for retrieving atypicality scores. Lastly, the notion of atypicality is not only seen and leverage in healthcare, but it is also present in other domains such as law. Adapting our methodology for other domains could further improve LLMs calibration performance.

Ethical considerations In our work, we focus on the medical domain with the goal of enhancing the calibration and accuracy of confidence scores provided by large language models to support better-informed decision-making. While our results demonstrate significant improvements in calibration, it is imperative to stress that LLMs should not be solely relied upon without the oversight of a qualified medical expert. The involvement of a physician or an expert is essential to validate the model's recommandations and ensure a safe and effective decision-making process.

Moreover, we acknowledge the ethical implications of deploying AI in healthcare. It is crucial to recognize that LLMs are not infallible and can produce erroneuous outputs. Ensuring transparency

in how these models reach their conclusions, and incorporating feedback from healthcare professionals are vital steps in maintaining the integrity and safety of medical practice. Thus, our work is a step towards creating reliable tools, but it must be integrated thoughtfully within the existing healthcare framework to truly benefit patient outcomes.

References

Anthropic. 2023. Claude 3. https://www.anthropic.com. Accessed: 2024-06-02.

S Bai, L Zhang, Z Ye, D Yang, T Wang, and Y Zhang. 2023. The benefits of using atypical presentations and rare diseases in problem-based learning in undergraduate medical education. *BMC Med Educ*, 23(1):93.

C. R. S. Banerji, T. Chakraborti, C. Harbron, et al. 2023. Clinical ai tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, 29:2996–2998.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *Preprint*, arXiv:2308.16175.

Google DeepMind. 2023. Gemini 1.0 pro. https://www.deepmind.com. Accessed: 2024-06-02.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Michael Goldrich and Amit Shah. 2021. *Atypical Presentations of Illness*. McGraw-Hill Education, New York, NY.

Y Harada, R Kawamura, M Yokose, T Shimizu, and H Singh. 2024. Definitions and measurements for atypical presentations at risk for diagnostic errors in internal medicine: Protocol for a scoping review. *JMIR Res Protoc*, 13:e56933.

Guande He, Peng Cui, Jianfei Chen, Wenbo Hu, and Jun Zhu. 2023. Investigating uncertainty calibration of aligned language models under the multiple-choice setting. *Preprint*, arXiv:2310.11732.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. *Preprint*, arXiv:2405.00623.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

O Kostopoulou, BC Delaney, and CW Munro. 2008. Diagnostic difficulty and error in primary care–a systematic review. *Fam Pract*, 25(6):400–413. Epub 2008 Oct 7.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Llama. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

JC Matulis, SN Kok, EC Dankbar, and AJ Majka. 2020. A survey of outpatient internal medicine clinician perceptions of diagnostic error. *Diagnosis (Berl)*, 7(2):107–114.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2015, pages 2901–2907.

OpenAI. 2023. Gpt-3.5-turbo. https://www.openai.com. Accessed: 2024-06-02.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Qwen. 2024. Qwen2.5: A party of foundation models.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. *Preprint*, arXiv:2401.08694.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don't show: Surrogate models for confidence estimation. *Preprint*, arXiv:2311.08877.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Quantifying uncertainty in natural language explanations of large language models. *Preprint*, arXiv:2311.03533.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Cassandra Vonnes and Rosalie El-Rady. 2021. When you hear hoof beats, look for the zebras: Atypical presentation of illness in the older adult. *The Journal for Nurse Practitioners*, 17(4):458–461.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Preprint*, arXiv:2401.12794.

Mert Yuksekgonul, Linjun Zhang, James Zou, and Carlos Guestrin. 2023. Beyond confidence: Reliable models should also consider atypicality. *Preprint*, arXiv:2305.18262.

A Additional Results

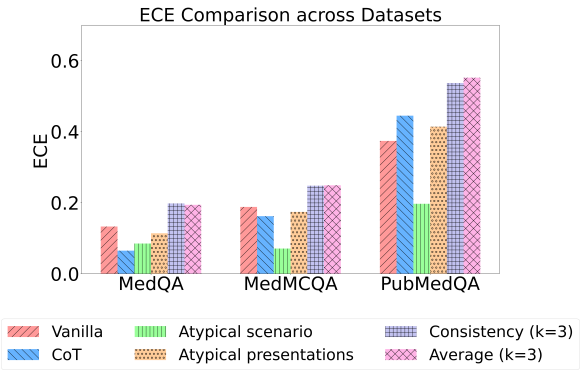

Figure 7: GPT4 ECE comparison across datasets

In the main sections of the paper, we presented figures for GPT3.5-turbo. Here we provide additional results for GPT3.5-turbo and the other three models to support the claims and findings discussed above. We show calibration and performance metrics for all methods used and across all three datasets: MedQA, MedMCQA and PubmedQA. Furthermore, we provide additional graphs to support the analysis of the distributions of atypicality scores across the different datasets as well as the distribution of atypicality scores by difficulty levels.

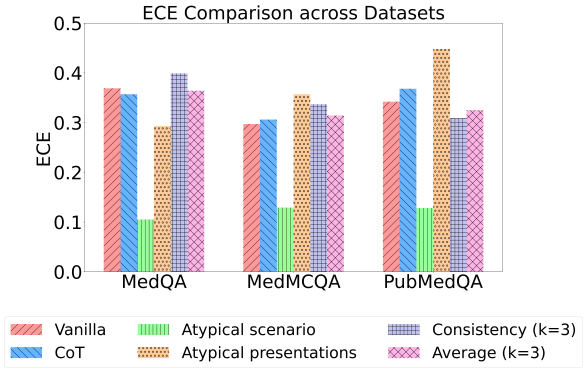

Figure 8: Gemini ECE comparison across datasets

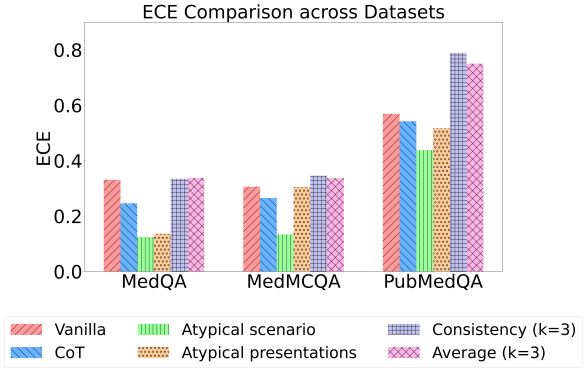

Figure 9: Claude3-sonnet ECE comparison across datasets

The findings and conclusions from these additional figures are already discussed in the main sections of the paper. These supplementary figures are included here to demonstrate that the findings are consistent across multiple models, ensuring that the conclusions drawn are robust and not based solely on one model.

B Prompt templates

We provide the full prompt used for *Atypical Scenario* and *Atypical Presentations*. Note that for completeness, the version of prompts provided contains the component of difficulty scores. This component is optional and is only used for analyzing the relationship between difficulty and atypicality. The prompt templates can be found at Table 5.

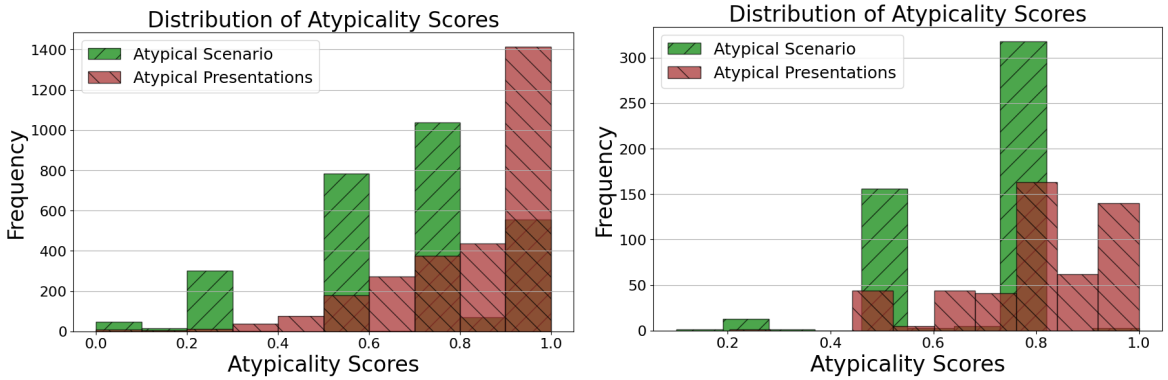

Figure 10: Atypicality Distributions of GPT3.5

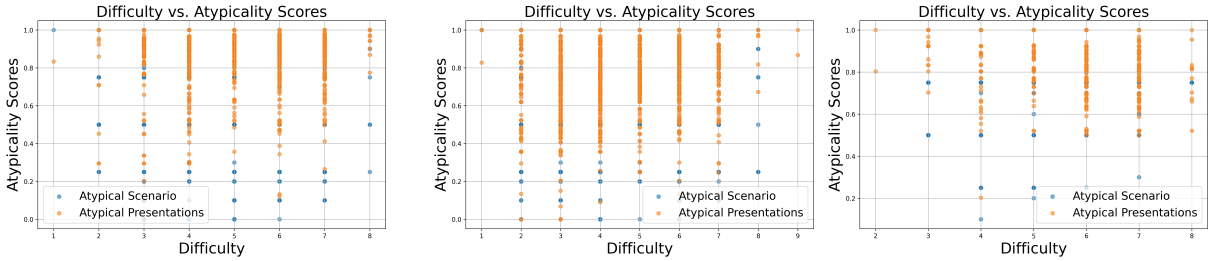

Figure 11: Atypicality by Difficulty for GPT3.5

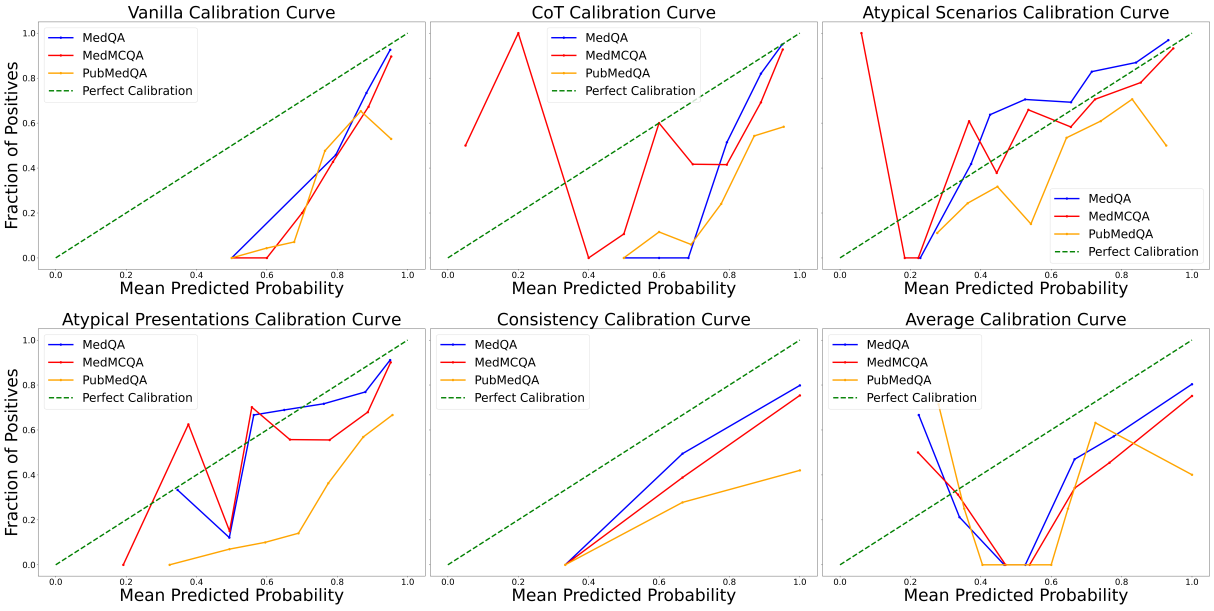

Figure 12: GPT4 Calibration curves across all methods

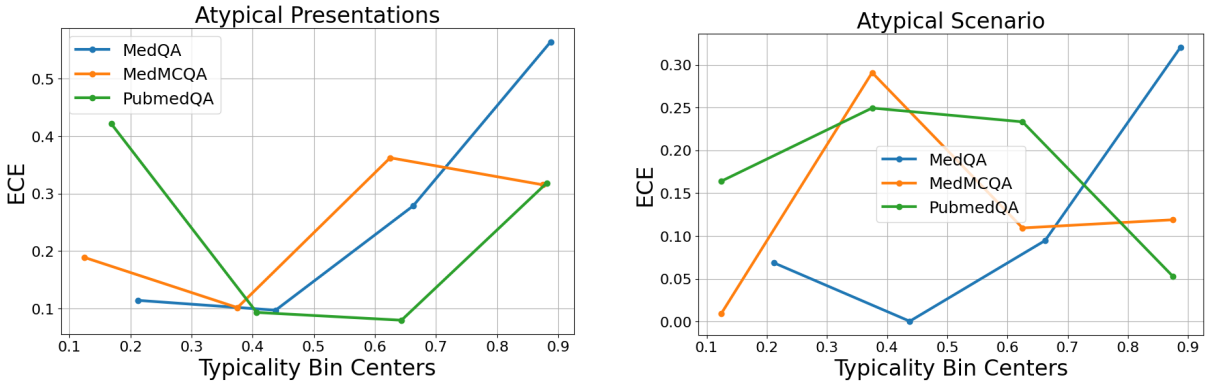

Figure 13: ECE by Typicality bins of GPT4-turbo for Atypical Presentations Aware Recalibration methods

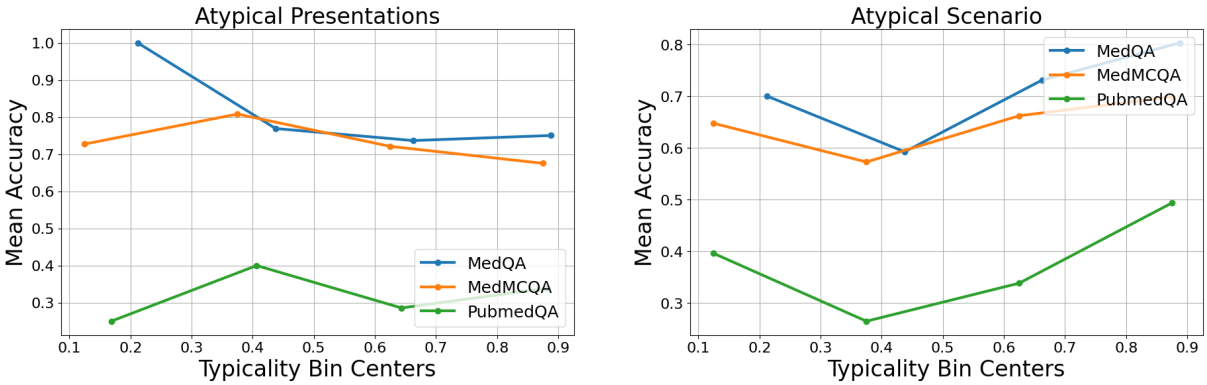

Figure 14: Accuracy by Typicality bins of GPT4-turbo for Atypical Presentations Aware Recalibration methods

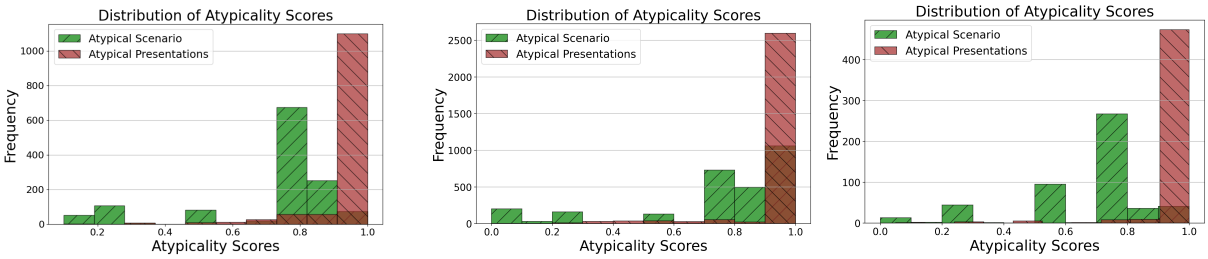

Figure 15: Atypicality Distribution of GPT4

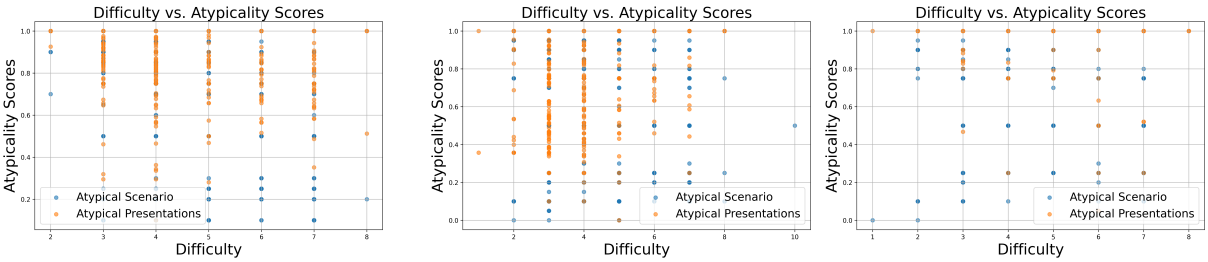

Figure 16: Atypicality by Difficulty of GPT4

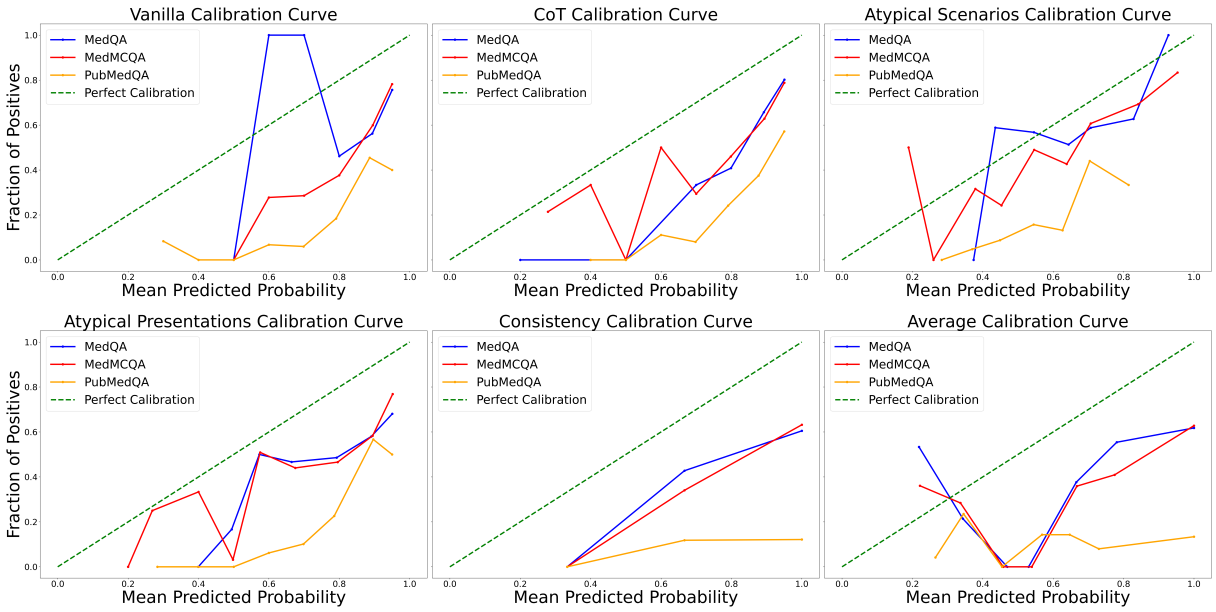

Figure 17: Claude3-sonnet Calibration curves across all methods

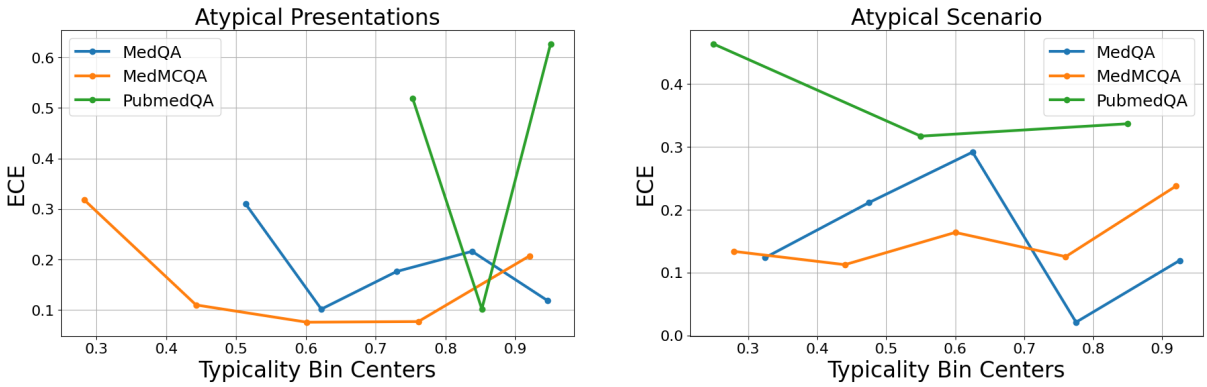

Figure 18: ECE by Typicality bins of Claude3-sonnet for Atypical Presentations Aware Recalibration methods

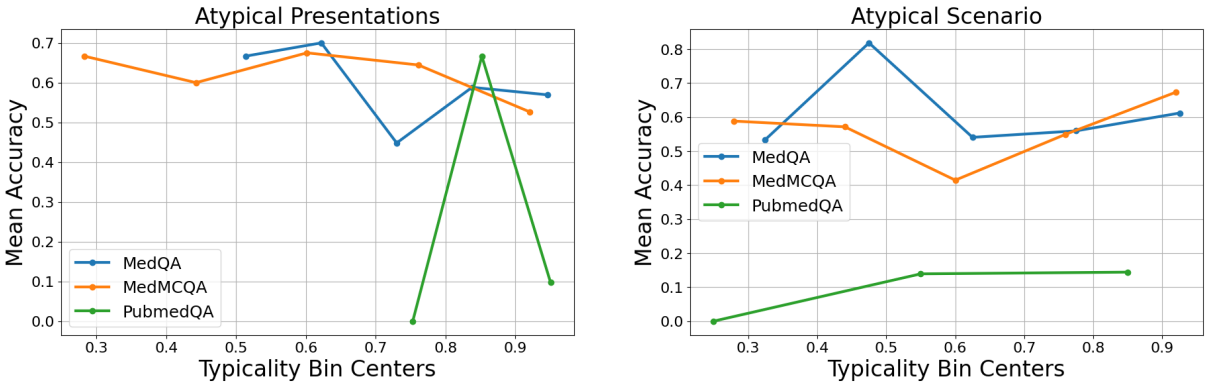

Figure 19: Accuracy by Typicality bins of Claude3-sonnet for Atypical Presentations Aware Recalibration methods

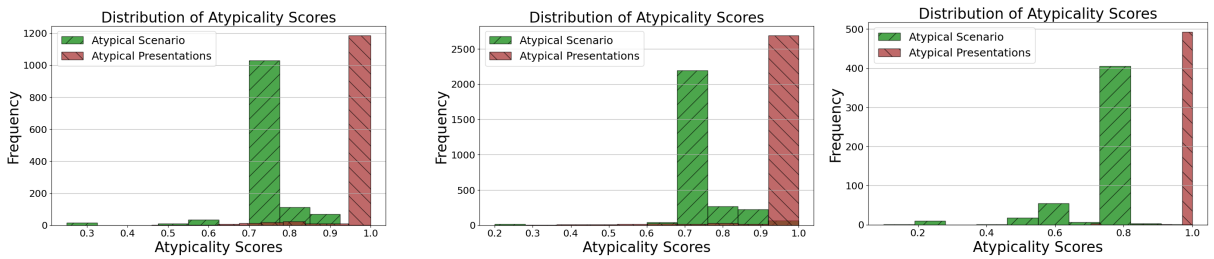

Figure 20: Atypicality Distribution of Claude

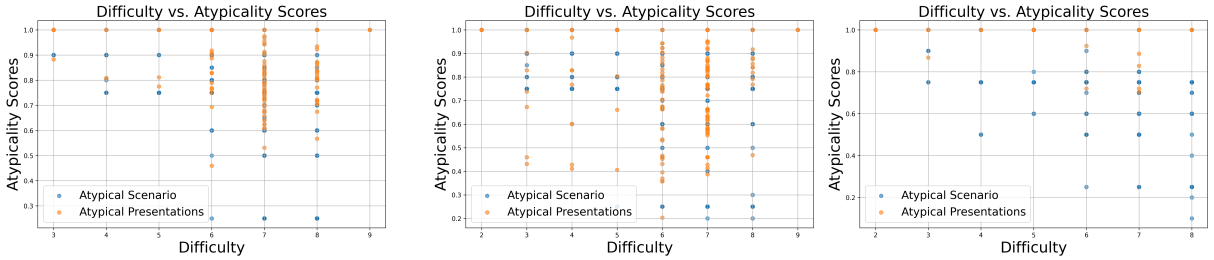

Figure 21: Atypicality by Difficulty of Claude

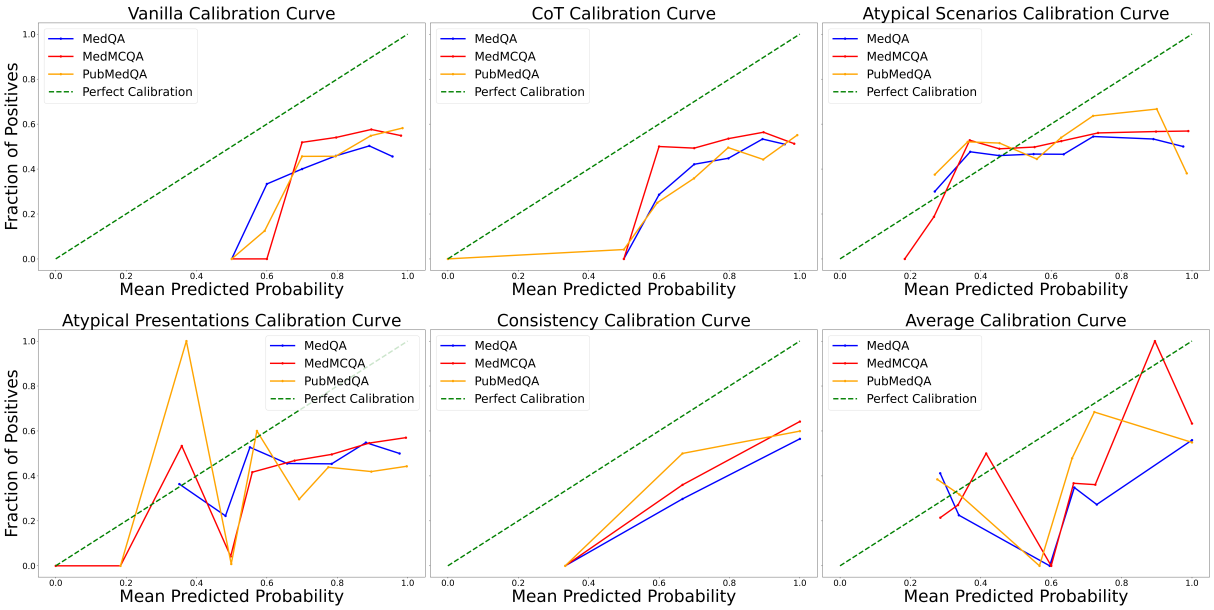

Figure 22: Gemini Calibration curves across all methods

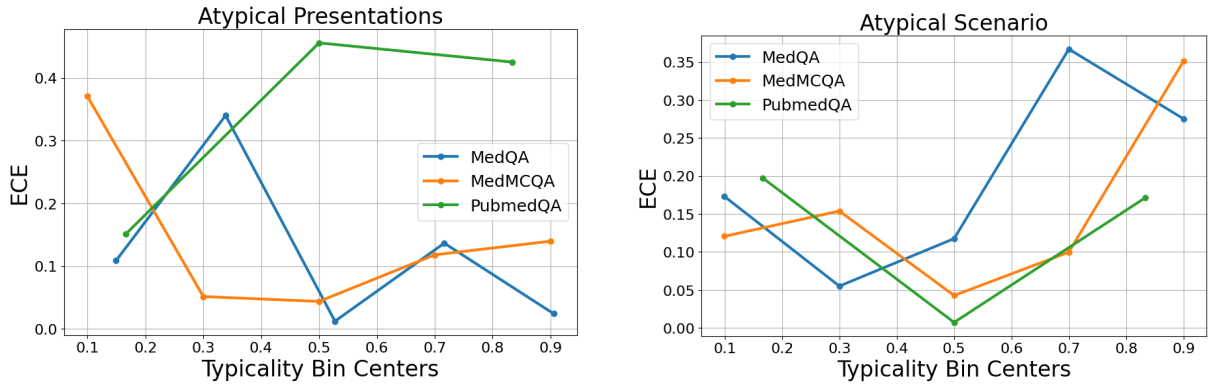

Figure 23: ECE by Typicality bins of Gemini for Atypical Presentations Aware Recalibration methods

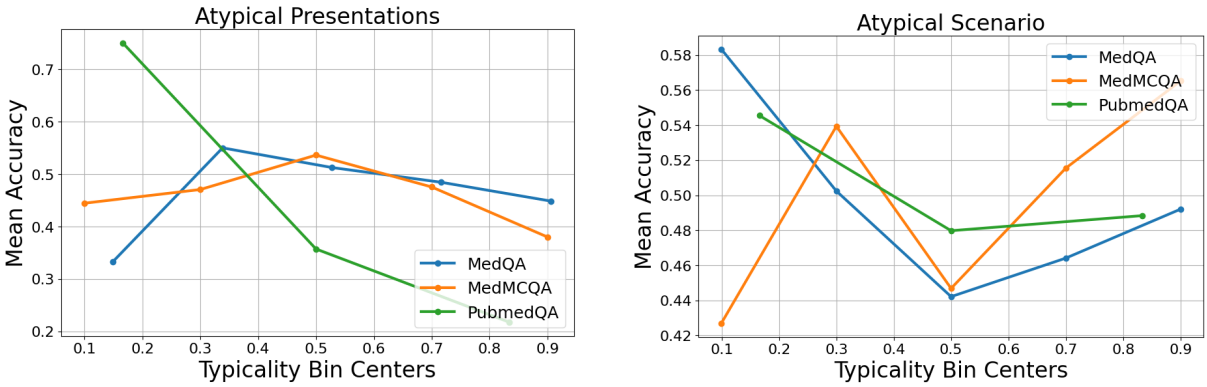

Figure 24: Accuracy by Typicality bins of Gemini for Atypical Presentations Aware Recalibration methods

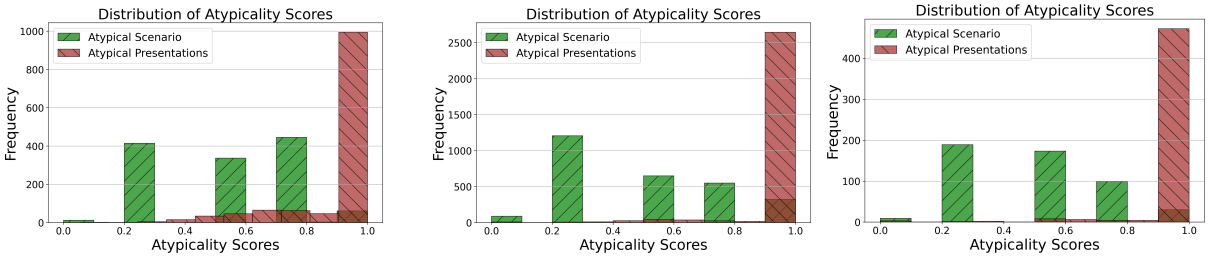

Figure 25: Atypicality Distribution of Gemini

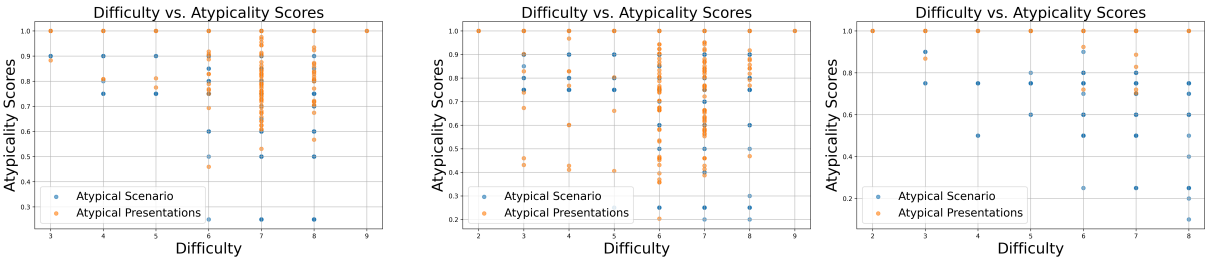

Figure 26: Atypicality by Difficulty of Gemini

		MedQA (n=1272)				MedMCQA (n=2816)				PubMedQA (n=500)			
Models	Methods	Acc	ECE	Brier	AUC	Acc	ECE	Brier	AUC	Acc	ECE	Brier	AUC
gpt-3.5-turbo	Vanilla	0.526	0.351	0.363	0.553	0.555	0.323	0.350	0.530	0.544	0.251	0.304	0.562
	CoT	0.536	0.318	0.334	**0.608**	0.525	0.357	0.360	0.588	0.516	0.275	0.360	**0.588**
	Atypical scenario	0.506	**0.084**	**0.262**	0.530	0.544	**0.128**	**0.252**	0.549	0.468	**0.115**	**0.252**	0.581
	Atypical presentations	0.506	0.283	0.332	0.557	0.527	0.152	0.322	0.515	0.544	0.129	0.268	0.540
	Consistency (k=3)	0.535	0.408	0.396	0.567	**0.561**	0.350	0.356	**0.613**	0.544	0.335	0.370	0.524
	Average (k=3)	**0.539**	0.398	0.397	0.555	**0.561**	0.350	0.344	**0.613**	**0.550**	0.346	0.372	0.536
claude3-sonnet	Vanilla	0.541	0.331	0.336	**0.613**	0.565	0.306	0.327	0.630	0.128	0.569	0.428	0.743
	CoT	**0.638**	0.246	0.282	0.599	**0.612**	0.265	0.295	0.615	**0.246**	0.542	0.469	0.663
	Atypical scenario	0.564	**0.124**	**0.259**	0.547	0.561	**0.134**	**0.268**	0.604	0.140	**0.438**	**0.252**	0.634
	Atypical presentations	0.568	0.136	0.332	0.564	0.531	0.305	0.316	**0.666**	0.100	0.517	0.339	**0.880**
	Consistency (k=3)	0.552	0.335	0.363	0.555	0.568	0.346	0.355	0.591	0.122	0.789	0.766	0.443
	Average (k=3)	0.558	0.338	0.358	0.565	0.568	0.337	0.356	0.585	0.128	0.750	0.725	0.491
gemini-pro-1.0	Vanilla	0.472	0.369	0.385	0.530	0.551	0.297	0.338	0.520	0.492	0.342	0.362	0.572
	CoT	0.465	0.357	0.369	0.578	0.526	0.306	0.340	0.537	0.438	0.368	0.373	0.590
	Atypical scenario	0.473	**0.105**	**0.268**	0.510	0.513	**0.129**	**0.274**	0.517	0.508	**0.128**	**0.276**	0.495
	Atypical presentations	0.458	0.293	0.332	0.568	0.387	0.357	0.316	**0.712**	0.226	0.448	0.338	**0.782**
	Consistency (k=3)	0.471	0.399	0.400	**0.613**	**0.557**	0.337	0.343	0.624	**0.540**	0.309	0.349	0.591
	Average (k=3)	**0.477**	0.364	0.391	0.599	0.549	0.314	0.341	0.635	0.504	0.325	0.371	0.529
gpt-4-turbo	Vanilla	0.756	0.133	0.190	0.670	0.707	0.188	0.230	0.673	0.394	0.374	0.337	**0.792**
	CoT	**0.832**	**0.065**	**0.132**	**0.710**	**0.730**	0.162	**0.206**	0.729	0.358	0.445	0.402	0.743
	Atypical scenario	0.741	0.085	0.181	0.693	0.675	**0.071**	**0.206**	0.672	0.354	**0.197**	**0.249**	0.679
	Atypical presentations	0.751	0.114	0.178	0.673	0.681	0.174	0.213	**0.739**	0.338	0.414	0.363	0.763
	Consistency (k=3)	0.775	0.198	0.206	0.555	0.712	0.248	0.253	0.587	**0.404**	0.537	0.546	0.490
	Average (k=3)	0.767	0.194	0.205	0.573	0.708	0.249	0.255	0.590	**0.404**	0.552	0.563	0.458

Table 4: Using atypicality as post-hoc calibration brings major improvements in ECE and Brier Scores across all datasets and all models. **Atypical Scenario** outperforms all other methods in calibration in the big majority of experiments.

Prompts	
Atypical Scenario	Question and Options: {question} First, assess the situation described in the question and assign an atypicality score between 0 and 1, where: - 0 indicates a highly atypical situation, uncommon or rare in such scenarios. - 1 indicates a very typical situation, commonly expected in such scenarios. - Scores between 0 and 1 (such as 0.25, 0.5, 0.75) indicate varying degrees of typicality. Situation Atypicality: [Atypicality score] Then, provide your response in the following format: Response: - Answer (letter): [Letter of the choice] - Difficulty: [Score on a scale from 1 to 10 with 10 being the hardest] - Confidence: [Percentage score between 0 and 100%] Answer, Difficulty, and Confidence:
Atypical Presentations	Question and Options: {question} First, assess each symptom and signs with respect to its typicality in the described scenario. Assign an atypicality score between 0 and 1, where: - 0 indicates a highly atypical situation, uncommon or rare in such scenarios. - 1 indicates a very typical situation, commonly expected in such scenarios. - Scores between 0 and 1 (such as 0.25, 0.5, 0.75) indicate varying degrees of typicality. Symptoms and signs: - Symptom 1: [Atypical score] - Symptom 2: [Atypical score] - Symptom 3: [Atypical score]- - ... Then, provide your response in the following format: Response: - Answer (letter): [Letter of the choice] - Difficulty: [Score on a scale from 1 to 10 with 10 being the hardest] - Confidence: [Percentage score between 0 and 100%] Answer, Difficulty, and Confidence:

Table 5: Complete prompts used for Atypical Presentations Aware Recalibration framework