Conformal Prediction with Large Language Models for Multi-Choice Question Answering

Bhawesh Kumar [*1] **Charles Lu** [*2] **Gauri Gupta** [2] **Anil Palepu** [3] **David Bellamy** [1] **Ramesh Raskar** [2] **Andrew Beam** [1]

Abstract

As large language models are widely developed, robust uncertainty quantification techniques will become crucial for safe deployment in high-stakes scenarios. This work explores how conformal prediction can quantify uncertainty in language models for multiple-choice question-answering. We find that the uncertainty estimates from conformal prediction are tightly correlated with prediction accuracy. This observation can be helpful in downstream applications such as selective classification and filtering out low-quality predictions. We also investigate the exchangeability assumption required by conformal prediction to out-of-subject questions, which may be a more realistic scenario for many practical applications. Our work contributes towards more trustworthy and reliable usage of large language models in safety-critical situations, where robust guarantees of error rate are required.

1. Introduction

Large language models (LLMs) have recently achieved impressive performance on a number of NLP tasks, such as machine translation, text summarization, and code generation. However, lingering concerns of trust and bias still limit their widespread application for critical decision-making domains such as healthcare.

One well-known issue with current LLMs is their tendency to "hallucinate" false information with seemingly high confidence. These hallucinations can occur when the model generates outputs not grounded in any factual basis or when the prompt is highly unusual or ambiguous. This behavior of LLMs may also result from how these models are trained — using statistical sampling for next-token prediction — which can progressively increase the likelihood of factual errors as the length of generated tokens increases (LeCun, 2023). Factually incorrect outputs may confuse and deceive users into drawing wrong conclusions, ultimately decreasing the overall system's trustworthiness. Decisions based on unpredictable or biased model behavior could have significant negative and socially harmful consequences in high-stakes domains such as healthcare and law.

Therefore, we seek to explore principled uncertainty quantification (UQ) techniques for LLMs that can provide guaranteed error rates of model predictions. Ideally, these UQ techniques should be model agnostic and easy to implement without requiring model retraining due to the intensive computing costs and limited API access associated with many LLMs. To this end, we investigate *conformal prediction*, a distribution-free UQ framework, to provide LLMs for the task of multiple-choice question-answering (MCQA).

Based on our experiments, we find the uncertainty, as provided by conformal prediction, to be strongly correlated with accuracy, enabling applications such as filtering out low-quality predictions to prevent a degraded user experience. We also verify the importance of the exchangeability assumption in conformal prediction (see section 2) for guaranteeing a user-specified level of errors.

To summarize, our contributions are the following:

- we adapt conformal prediction for MCQA tasks to provide distribution-free uncertainty quantification in LLMs,
- show how the uncertainty provided by conformal prediction can be useful for downstream tasks such as selective classification,
- and assess the performance of conformal prediction when the exchangeability assumption is violated for in-context learning in LLMs.

[*]Equal contribution [1]Harvard University [2]MIT Media Lab [3]Harvard-MIT Health Sciences & Technology. Correspondence to: Bhawesh Kumar <bhaweshk@mit.edu>, Charlie Lu <luchar@mit.edu>.

This work was published at the *ICML 2023 (Neural Conversational AI TEACH) workshop.* Honolulu, Hawaii, USA.

2. Conformal Prediction

Uncertainty quantification (UQ) techniques are critical to deploying machine learning in domains such as healthcare (Bhatt et al., 2021; Kompa et al., 2021b;a). Conformal prediction (Gammerman et al., 2013; Vovk et al., 2022) is a flexible and statistically robust approach to uncertainty quantification. Informally, the central intuition behind conformal prediction is to output a set of predictions containing the correct output with a user-specified probability.

By providing a more nuanced understanding of the model's confidence and a statistically robust coverage guarantee, conformal prediction paves the way for improved and more reliable applications of machine learning models across various domains (Kumar et al., 2022).

Prediction sets. Formally, let $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ be a set-valued function that generates a prediction sets over the powerset of Y given an input X. This prediction set naturally encodes the model's uncertainty about any particular input by the **size** of the prediction set.

Expressing uncertainty as the set size is an intuitive output that can be helpful in decision-making contexts (Babbar et al., 2022). For example, in medical diagnosis, the concept of prediction set is similar to a differential diagnosis, where only likely and plausible conditions are considered given the observed symptoms of a patient (Lu et al., 2022c). Indeed, conformal prediction has been utilized for uncertainty quantification in healthcare applications such as medical imaging analysis (Lu et al., 2022a;b; Lu & Kalpathy-Cramer, 2022).

Coverage guarantee. Conformal methods generate prediction sets that ensure a certain user-specified probability of containing the actual label, regardless of the underlying model or distribution. This guarantee is achieved without direct access or modification to the model's training process and only requires a held-out calibration and inference dataset. This makes conformal prediction well-suited to LLM applications when retraining is costly and direct model access is unavailable through third-party or commercial APIs.

The coverage guarantee states that the prediction sets obtained by conformal prediction should contain the true answer on average at a user-specified *level*, α. This property is called *coverage*, and the corresponding coverage guarantee is defined as:

$$1 - \alpha \leq \mathbf{P}\left(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})\right), \tag{1}$$

where $\alpha \in (0, 1)$ is the desired error rate, and $\mathcal{C}$ is the calibrated prediction set introduced above. $(X_{\text{test}}, Y_{\text{test}}) \sim \mathcal{D}_{\text{calibration}}$ is an unseen test point that is drawn from the same distribution as the data used to calibrate the prediction sets.

Conformal Calibration Procedure. As previously mentioned, conformal prediction only needs the scores of a model to calibrate and construct the prediction sets. We now describe how to calibrate the prediction sets for a specific score function.

Let $f : \mathcal{X} \to \Delta^{|\mathcal{Y}|}$ be a classifier with a softmax score, where Δ is a $|\mathcal{Y}|$-dimensional probability simplex. A common choice for the score function, *least ambiguous set-valued classifiers* (LAC) (Sadinle et al., 2019), is defined as

$$S(X, Y) = 1 - [f(X)]_Y\,, \tag{2}$$

where $[f(X)]_Y$ is the softmax score at the index of the true class.

To calibrate the prediction sets to our desired level of coverage, we need to estimate a threshold $\hat{q}_\alpha$ that is the $1 - \alpha$ quantile of the calibration scores

$$\hat{q}_\alpha = \text{Quantile}\left(\{s_1, \ldots, s_n\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right), \tag{3}$$

where $\{s_1, \ldots, s_n\}$ are the LAC scores of the calibration set.

At inference time, prediction sets can be constructed in the following manner:

$$\mathcal{C}(X) = \{y \in \mathcal{Y} : S(X, y) \leq \hat{q}_\alpha\}\,, \tag{4}$$

Exchangeability assumption. Conformal prediction assumes that the data used to calibrate the prediction sets is exchangeable with the test data at inference time. If this assumption holds, the coverage guarantee, as stated in Equation 1, will hold, and the resulting prediction sets will have the desired error rate.

Exchangeability can be viewed as weaker than the independent and identically distributed (IID) assumption (Bernardo, 1996). This assumption is often made in machine learning with regard to the training, validation, and test sets. The threshold used to determine the size of the prediction set is estimated on a held-out calibration data set that is assumed to be *exchangeable* with the test distribution.

3. Prompt Engineering

In this paper, we focus on the task of multiple-choice question answering (MCQA) and frame MCQA as a supervised classification task, where the objective is to predict the correct answer choice out of four possible options. We wish to quantify the model uncertainty over the predicted output using conformal prediction. We condition each option choice (A, B, C, and D) on the prompt and question and use the LLaMA-13B model (Touvron et al., 2023) to generate the logit corresponding to each multiple-choice answer. We

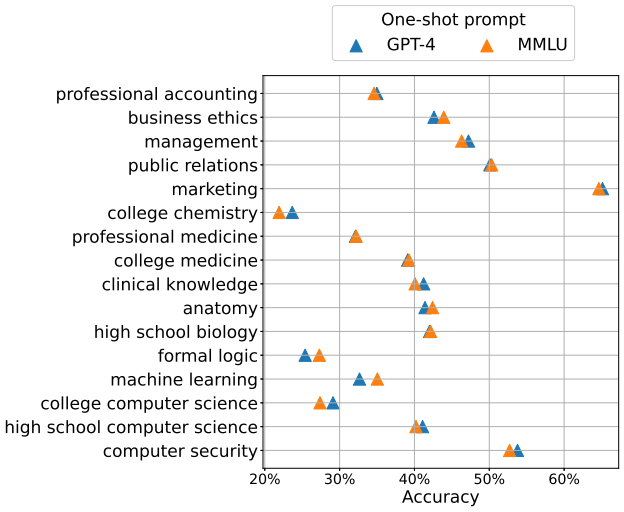


Figure 1: **LLaMA MCQA accuracy is similar for GPT-4 generated questions and real MMLU questions across subjects.** For most MMLU subjects, prediction accuracy using one-shot GPT-4 generated questions is similar to when actual MMLU questions are used in one-shot prompts. Results are averaged over ten randomly selected one-shot GPT-4 and MMLU prompts.

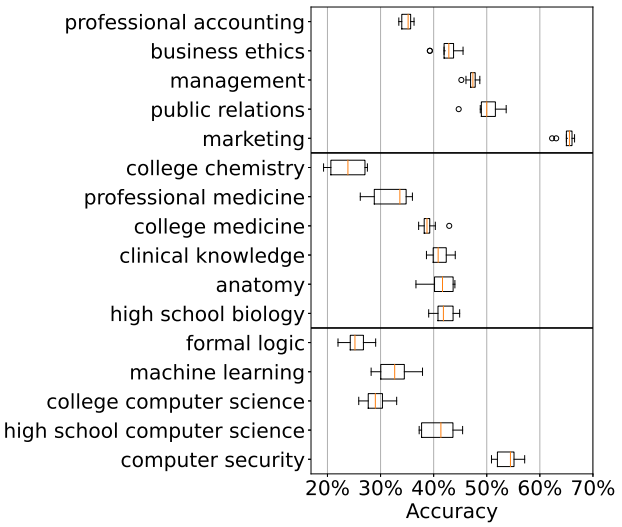


Figure 2: **The accuracy distribution across subjects for ten prompts.** We plot the distribution of accuracy for ten different one-shot prompts.

normalize the four logits using the softmax to obtain valid probabilities for each option.

One-shot prompting. LLMs are very sensitive to the exact input prompt, which has motivated a whole field of in-context learning and prompt engineering or prompt tuning (Zhou et al., 2023; Wei et al., 2023). Context learning refers to the ability of LLMs to understand and make predictions based on the context in which the input data is presented without updating the model weights. Prompt engineering methods vary significantly among tasks and require heavy experimentation and reliance on hand-crafted heuristics. For the current setup, model performance on classification tasks is often sensitive to the prompts used. Thus, we experiment with several prompting strategies before finalizing our prompts.

We use one-shot prompting by including one context example. For each subject, we use a slightly different prompt. For example, we prompt the model to assume it is the "world's best expert in college chemistry" when generating predictions for college chemistry subjects.

We also use ten different prompts for each subject to generate ten softmax probability outputs to reduce variance. We obtain the final probability outputs for a question by averaging the softmax outputs corresponding to these ten prompts. The ten prompts for a given subject only vary in terms of the one-shot question. A sample prompt for high school biology is provided below:

```
This is a question from high school
biology.

A piece of potato is dropped into a
beaker of pure water. Which of the
following describes the activity after
the potato is immersed into the water?
(A) Water moves from the potato into
the surrounding water.
(B) Water moves from the surrounding
water into the potato.
(C) Potato cells plasmolyze.
(D) Solutes in the water move into
the potato.
The correct answer is option B.

You are the world's best expert in
high school biology. Reason
step-by-step and answer the
following question.
From the solubility rules, which of
the following is true?
(A) All chlorides, bromides, and
iodides are soluble
(B) All sulfates are soluble
(C) All hydroxides are soluble
(D) All ammonium-containing compounds
are soluble

The correct answer is option:
```

GPT-4 generated examples. We explore two approaches for the one-shot example in the prompts: (1) One-shot example is one of the questions in the MMLU dataset for that subject. We then exclude this specific question for generating predictions with the resulting prompt. (2) We use GPT-4 to generate multiple-choice questions for each subject. We then cross-check the questions and answers produced by GPT-4 for correctness and select ten correct question-answer pairs.

We use the following prompt to generate MCQs for clinical knowledge from GPT-4: "*Give me 15 multiple choice questions on clinical knowledge with answers*". Specific questions and answers generated by the GPT-4 are available from our code (refer to Section 4.4.) We have also included a subset of sample GPT-4 generated questions and answers as well as MMLU-based questions and answers in the Appendix (A.1)

We generate MCQs for other subjects using similar prompts. GPT-4-based one-shot questions produce more accurate answers than MMLU-based questions, as shown in Figure 1.

After controlling for the size of the prompts (limited to 700 tokens), we find that MMLU-based and GPT-4 based one-shot questions produce similar accuracy on the sixteen subjects we evaluate. We conduct all the following experiments on prompts that use GPT-4-based one-shot questions since they are shorter on average and achieve similar performance.

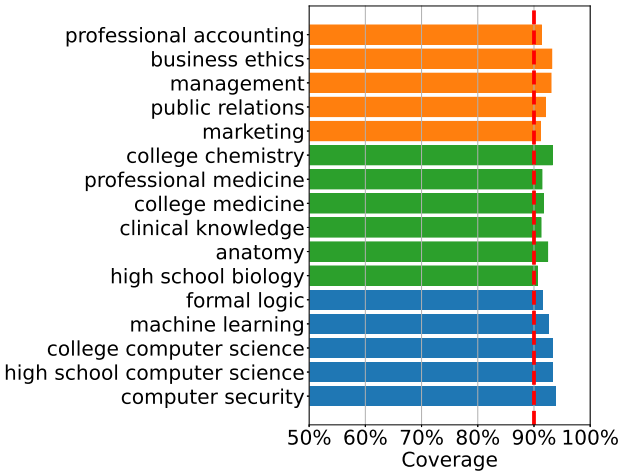


Figure 3: **Desired coverage is achieved for all subjects.** The red dashed line shows the desired coverage rate (specified at $\alpha = 0.1$), which is guaranteed by conformal prediction to be with at least $1 - \alpha$ percent of the time. The colors denote the three categories of questions.

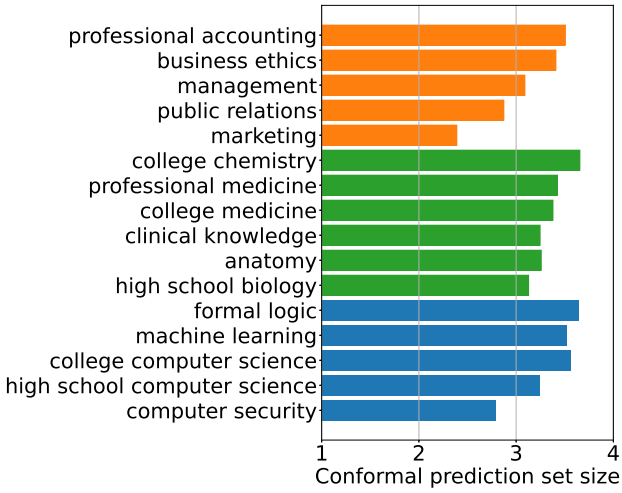


Figure 4: **Uncertainty quantification using prediction set size.** In conformal prediction, a set of predictions is generated for each question. The size of this set indicates how uncertain the model is for a particular question. Larger set sizes denote greater uncertainty, and smaller set sizes denote less uncertainty. The colors denote the three categories of questions.

4. Experiments

4.1. Model and dataset

We use the LLaMA-13B model (Touvron et al., 2023) to generate predictions for MCQA. LLaMA-13B is an open-source 13 billion parameter model trained on 1 trillion tokens and has been shown to achieve good zero-shot performance on various question-answering benchmarks. For our dataset, we use the MMLU benchmark (Hendrycks et al., 2021), which contains MCQA questions from 57 domains covering subjects such as STEM, humanities, and medicine.

For our experiments, we considered the following subset of MMLU: computer security, high school computer science, college computer science, machine learning, formal logic, high school biology, anatomy, clinical knowledge, college medicine, professional medicine, college chemistry, marketing, public relations, management, business ethics, and professional accounting. We group these domains into three broad categories: "business", "medicine", and "computer science". These 16 subjects represent diverse domains and have sufficient samples (each with at least 100 questions).

We perform classification by obtaining logit scores corresponding to option choices 'A', 'B', 'C', and 'D' conditioned on the one-shot prompt and the question. For example, for the sample prompt and question pair described in section 3, we find the logit score corresponding to the next tokens corresponding to each of the four options. We then take the

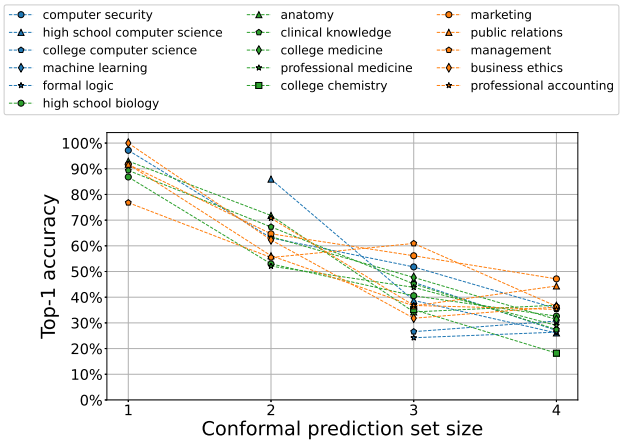


Figure 5: **Top-1 accuracy stratified by prediction set size.** For all subjects, we find a strong correlation between the prediction uncertainty (as measured by set size) and the top-1 accuracy of those predictions. Conformal prediction can be used for selective classification by filtering those predictions in which the model is highly uncertain.

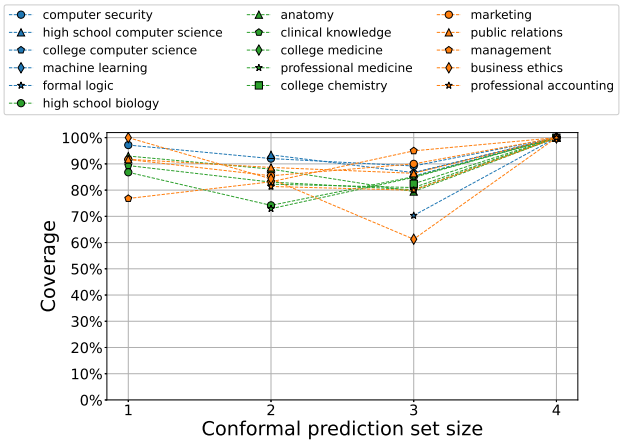


Figure 6: **Stratified coverage at each size of prediction set.** For most subjects, coverage is fairly consistent at all set sizes for prediction sets constructed with the conformal prediction procedure at $\alpha = 0.1$. This means that the true answer is one of the items in the predicted set on average about 90% of the time.

softmax over the logit scores corresponding to the options choices to obtain probability scores. The softmax scores corresponding to ten different prompts (that vary in terms of one-shot questions) are averaged to obtain final probability scores for each question-option pair.

4.2. Setup

We randomly split the data into equal-sized calibration and evaluation sets for each subject and averaged results over 100 random trials for our conformal prediction experiments. For each trial, we randomly sample 50% of data for calibration and 50% to evaluate coverage and set size. Thus, we have at least 50 samples for calibration. While the theoretical guarantee of conformal prediction holds on average for even such a small number of calibration samples, the individual 100 random trials may not always have exact coverage. A higher calibration size can reduce variance in coverage associated with the different random trials (Angelopoulos & Bates, 2021b).

4.3. Results

Naive Calibration in LLMs. Previous works have studied calibration in context to LLMs. (Si et al., 2022) looked at the limitation of traditional calibration metrics like Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). (Jiang et al., 2021) looked at T5, BART, and GPT-2 language models and found that the models are not calibrated for question-answering tasks. More recently, (Kadavath et al., 2022) found that large language models are well calibrated for various MCQA tasks. In current work, we examine the calibration error in the softmax probability output for the MCQA task for the LLaMA-13B language model. To this end, we calculate the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE), metrics that measure the average and maximum discrepancy between the confidence of the model's predictions and their accuracy. We find that the naive softmax output of the model is reasonably well calibrated across subjects on average, with ECE varying between a minimum of 1% for high school biology to a maximum of 7% for marketing (refer figure 9 in the appendix.) This aligns with previous findings on calibration error in LLMs (Kadavath et al., 2022). Nonetheless, MCE is significant for most subjects, indicating that the model is under-confident or over-confident at specific confidence levels. Additionally, there are no formal guarantees in terms of calibration errors.

Difference in coverage and set sizes between subjects. We next implement the conformal prediction procedure and compare coverage and prediction set size between subjects in Figure 3 and Figure 4 at the error rate $\alpha = 0.1$. The coverage guarantee of conformal prediction holds across all subjects (Figure 3). Comparing Figure 2 and Figure 4, we see that for each of the three categories, uncertainty — as measured by prediction set sizes — is, in general, significant for subjects with low top-1 accuracy and low for subjects with high top-1 accuracy.

For example, more challenging subjects such as formal logic and college chemistry have the most uncertainty on average, while "easier" subjects such as marketing have the lower average uncertainty. We show more results for different α

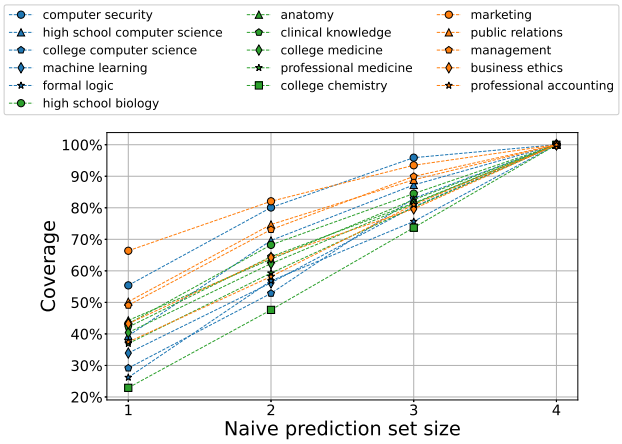


Figure 7: **Coverage of naive top-k prediction sets.** Coverage sharply falls off at smaller set sizes for naive prediction sets constructed by simply taking the top-k softmax scores for all predictions.

values in Table 1.

Selective classification with conformal prediction. Conformal Prediction framework can also be used for selective classification (Angelopoulos et al., 2022b; Angelopoulos & Bates, 2021a). In Figure 5, we analyze the correlation between uncertainty (as measured by conformal prediction) and top-1 accuracy performance. Specifically, we look at top-1 accuracy across subjects stratified by the size of the prediction set outputted by conformal prediction. We find a robust negative correlation between set size and top-1 accuracy for all subjects. This is intuitive as models with low confidence scores should correspond to less accurate predictions.

The accuracy for prediction sets with only one prediction is significantly higher than naive top-1 accuracy, as shown in Figure 7 (refer $k = 1$ accuracy). Thus, our results demonstrate that the set size obtained from conformal prediction procedure can filter low-quality predictions in downstream applications for LLMs. For example, highly uncertain predictions in a disease screening application should be flagged for manual review and not shown to the user.

Size-stratified coverage and comparison with naive top-k prediction sets. Size-stratified coverage measures error-rate guarantee across prediction sets of different sizes (Angelopoulos et al., 2022a). This experiment shows that coverage is not trivially satisfied by naively forming prediction sets by simply taking the top-k highest softmax probabilities. In Figure 7, we show the coverage when all prediction sets have a fixed set size and find that coverage decreases sharply with size. This is in contrast to prediction sets formed by conformal prediction in Figure 6, where we find that even prediction sets of size one have close to the desired level of coverage (90% when $\alpha = 0.1$) across most subjects. Indeed, we found that coverage is consistent over all set sizes for conformal prediction.

Conformal prediction can be thought of as outputting "adaptive" prediction sets that try to attain the proper level of coverage (depending on the chosen error rate α) instead of "fixed" prediction sets of size k.

Exchangeability assumption across subjects. In Figure 8, we test the exchangeability assumptions between subjects by calibrating on one subject and evaluating coverage on a different subject, grouped into three categories of subjects. Recall that the exchangeability assumption is needed for the coverage guarantee of Equation 1 to hold.

On the main diagonal, where the prediction sets are calibrated and evaluated on the same subject, we observed little deviation from the desired coverage rate of 90%. For example, prediction sets calibrated and evaluated on the same subject had close to the desired error rate of 10% when $\alpha = 0.1$. On the off-diagonal, we can see significant disparities between some subjects. For example, when prediction sets are calibrated on MCQA data from "high school computer science" and evaluated on "business ethics", coverage is only around 83%, less than the desired 90% coverage. However, for subjects from similar domains and accuracy, such as "clinical knowledge", "anatomy", and "high school biology", we find relatively more minor deviations from the targeted coverage rate when calibrated on out-of-subject data. This may result from good generalization capabilities and relatively calibrated softmax probability (Kadavath et al., 2022) outputted by the LLMs.

4.4. Code Availability

We release the code at this **Github repository**. The code repository also contains the question-answer pairs generated by GPT-4 for our prompts.

5. Discussion

As Large Language Models (LLMs) become increasingly powerful and are deployed in mission-critical systems, obtaining formal uncertainty guarantees for these models is crucial.

In this work, we investigated uncertainty quantification in LLMs in the context of multiple-choice questions using conformal prediction, a statistical framework, for generating prediction sets with coverage guarantees.

We found that naive softmax outputs of LLMs are relatively well calibrated on average but can suffer from under-confidence and over-confidence, and the extent of miscalibration varies across different subjects. To have a formal guarantee on the error rate of the model prediction, we im-

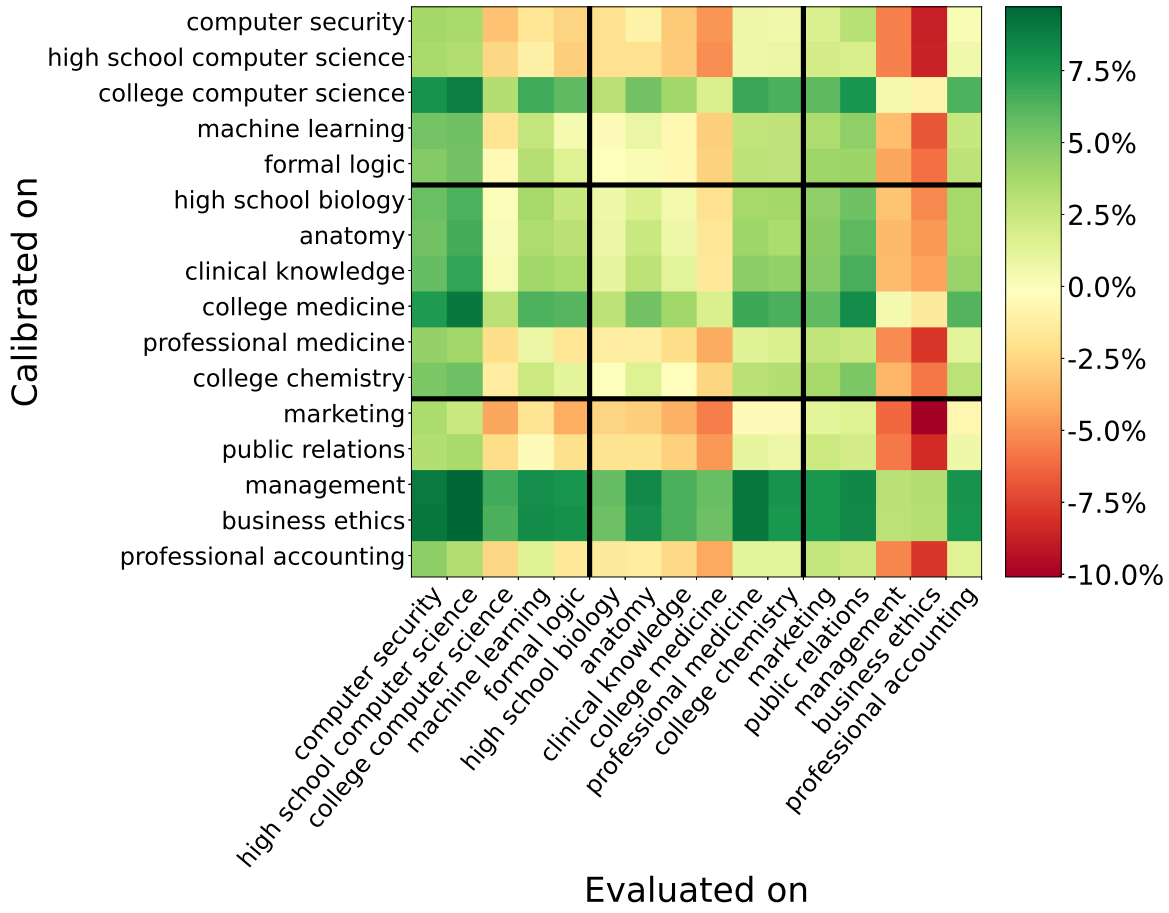


Figure 8: **Difference in coverage when calibrated on different subjects.** Deviation from 90% coverage for $\alpha = 0.1$. The off-diagonals represent entries corresponding to the cases where exchangeability conditions are violated between calibration and evaluation data sets. The subjects are grouped into the three broad categories of computer science, medicine, and business.

plemented the conformal prediction procedure on the naive softmax output of the LLM.

The conformal prediction framework produces valid prediction sets with error rate guarantees when calibration and evaluation sets come from the same distribution. We also explored the application of conformal prediction procedures for selective classification tasks. We found that conformal prediction can be used to discard predictions with unusual and low-quality outputs where the model is not confident, as indicated by the size of its prediction sets.

To summarize, our main takeaways are

- Developers of LLM systems should provide estimates of uncertainty to improve trustworthiness in their outputs to users.
- Uncertainty quantification can be useful for downstream applications such as filtering biased, unusual, or low-quality outputs.
- Conformal prediction is one approach to uncertainty quantification where a user-specified error rate can be statistically guaranteed when the calibration data is exchangeable with the test data.
- For our specific dataset (MMLU) and LLM (LLaMA-13B), we find that softmax outputs obtained as described in section 4.1 are reasonably calibrated on average. Nonetheless, models suffer from under-confidence and overconfidence, especially at the tail ends of probability distribution (refer figure 9 in the Appendix.)

Our work has some limitations. Our findings were limited to the MCQA task on the MMLU dataset using the LLaMA-13B model. Future works could extend our findings to multiple models and data sets. Further, it would be interesting to extend the conformal prediction framework

to more general settings like free-form text generation to control for inaccurate, biased, and harmful outputs from LLMs. It would also be interesting to explore exchangeability conditions in LLMs further when calibration and evaluation data sets are from different distributions (i.e., not just from MMLU), which is a more realistic scenario.

Despite these limitations, our work represents, to our knowledge, the first exploration of conformal prediction for LLMs in classification tasks. Our results contribute to the growing body of research on uncertainty estimation and generalization capabilities of LLMs and serve as a step forward in developing more robust and reliable uncertainty measures for increasingly capable large language models. Such measures are essential for ensuring LLMs' safe and responsible deployment in mission-critical applications.

Acknowledgement

We thank Prof. Yoon Kim, Abbas Zeitoun, and Anastasios Angelopoulos for helpful discussions and feedback on this work.

References

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction, 2022a.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021a. URL https://arxiv.org/abs/2107.07511.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. 2021b. doi: 10.48550/ARXIV.2107.07511. URL https://arxiv.org/abs/2107.07511.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022b.

Babbar, V., Bhatt, U., and Weller, A. On the utility of prediction sets in human-ai teams. *arXiv preprint arXiv:2205.01411*, 2022.

Bernardo, J. M. The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122, 1996.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Gammerman, A., Vovk, V., and Vapnik, V. Learning by transduction, 2013.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. 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. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022.

Kompa, B., Snoek, J., and Beam, A. L. Empirical frequentist coverage of deep learning uncertainty quantification procedures. *Entropy*, 23(12):1608, 2021a.

Kompa, B., Snoek, J., and Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021b.

Kumar, B., Palepu, A., Tuwani, R., and Beam, A. Towards reliable zero shot classification in self-supervised models with conformal prediction. *arXiv preprint arXiv:2210.15805*, 2022.

LeCun, Y. Do large language models need sensory grounding for meaning and understanding? In *Workshop on Philosophy of Deep Learning, NYU Center for Mind, Brain, and Consciousness and the Columbia Center for Science and Society*, Mar 2023. URL https://drive.google.com/file/d/1BU5bV3X5w65DwSMapKcsr0ZvrMRU_Nbi/view.

Lu, C. and Kalpathy-Cramer, J. Distribution-free federated learning with conformal predictions, 2022.

Lu, C., Angelopoulos, A. N., and Pomerantz, S. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pp. 545–554. Springer, 2022a.

Lu, C., Chang, K., Singh, P., and Kalpathy-Cramer, J. Three applications of conformal prediction for rating breast density in mammography. *arXiv preprint arXiv:2206.12008*, 2022b.

Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12008–12016, 2022c.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341. URL https://doi.org/10.1080/01621459.2017.1395341.

Si, C., Zhao, C., Min, S., and Boyd-Graber, J. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2814–2829, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.204.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer International Publishing, 2022. doi: 10.1007/978-3-031-06649-8. URL https://doi.org/10.1007%2F978-3-031-06649-8.

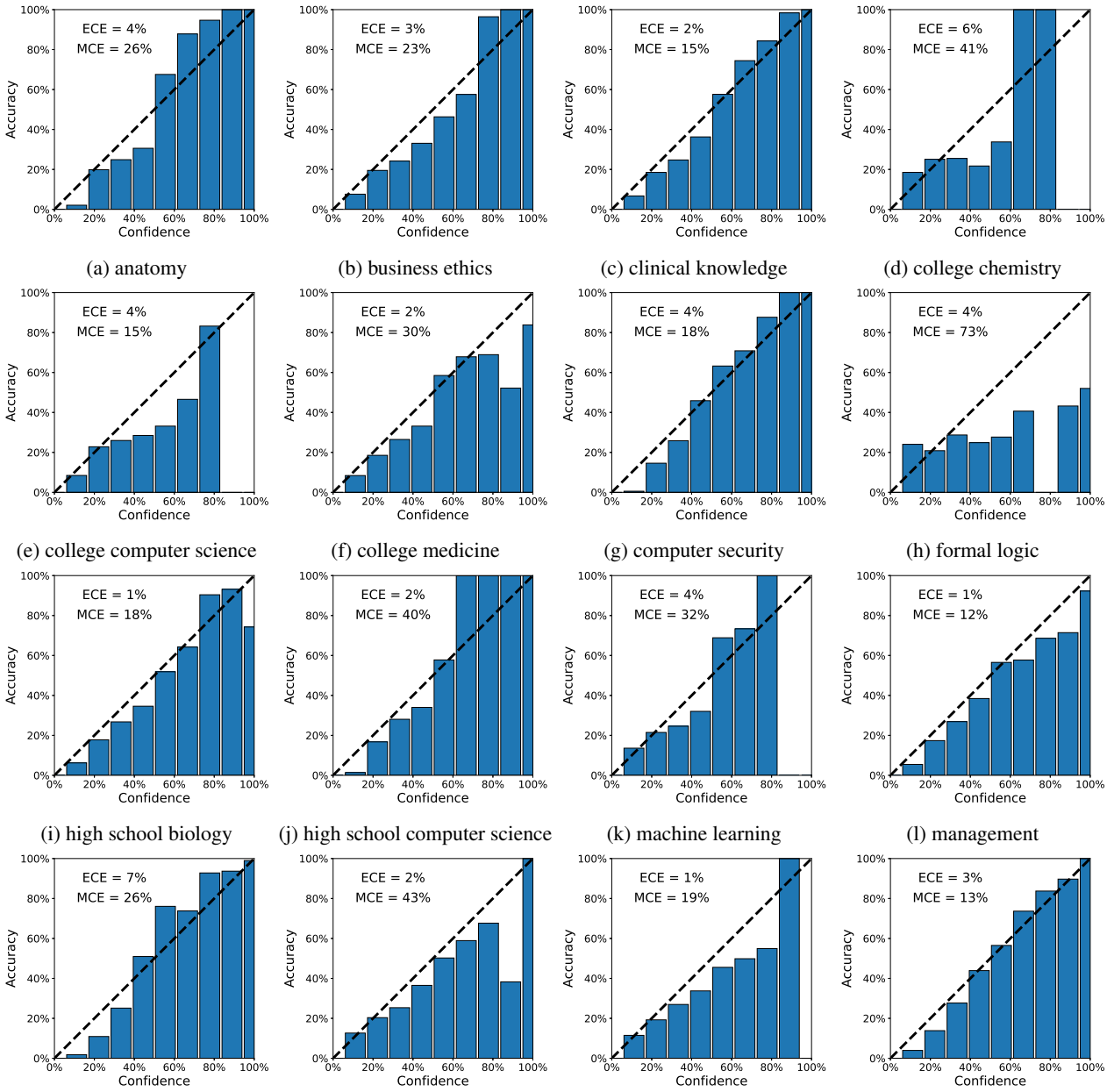
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., and Chi, E. Least-to-most prompting enables complex reasoning in large language models, 2023.

A. Appendix

Table 1: Empirical coverage and prediction set size at two specified error rates.

DATASET	$1-\alpha$	COVERAGE	SET SIZE
PROFESSIONAL ACCOUNTING	90%	$91\% \pm 3\%$	3.5 ± 0.1
	80%	$81\% \pm 3\%$	3.0 ± 0.0
BUSINESS ETHICS	90%	$93\% \pm 2\%$	3.4 ± 0.1
	80%	$82\% \pm 3\%$	2.8 ± 0.2
MANAGEMENT	90%	$94\% \pm 2\%$	3.1 ± 0.1
	80%	$83\% \pm 3\%$	2.5 ± 0.1
PUBLIC RELATIONS	90%	$93\% \pm 2\%$	3.0 ± 0.1
	80%	$83\% \pm 2\%$	2.3 ± 0.1
MARKETING	90%	$91\% \pm 1\%$	2.4 ± 0.1
	80%	$81\% \pm 1\%$	1.6 ± 0.1
COLLEGE CHEMISTRY	90%	$93\% \pm 2\%$	3.6 ± 0.1
	80%	$82\% \pm 4\%$	3.2 ± 0.1
PROFESSIONAL MEDICINE	90%	$91\% \pm 6\%$	3.4 ± 0.2
	80%	$82\% \pm 7\%$	2.9 ± 0.2
COLLEGE MEDICINE	90%	$92\% \pm 2\%$	3.4 ± 0.1
	80%	$82\% \pm 2\%$	2.8 ± 0.1
CLINICAL KNOWLEDGE	90%	$91\% \pm 3\%$	3.2 ± 0.1
	80%	$82\% \pm 3\%$	2.7 ± 0.1
ANATOMY	90%	$92\% \pm 3\%$	3.3 ± 0.1
	80%	$81\% \pm 4\%$	2.7 ± 0.1
HIGH SCHOOL BIOLOGY	90%	$91\% \pm 1\%$	3.2 ± 0.1
	80%	$81\% \pm 2\%$	2.6 ± 0.1
FORMAL LOGIC	90%	$92\% \pm 2\%$	3.7 ± 0.1
	80%	$82\% \pm 3\%$	3.2 ± 0.1
MACHINE LEARNING	90%	$93\% \pm 2\%$	3.6 ± 0.1
	80%	$82\% \pm 4\%$	3.1 ± 0.1
COLLEGE COMPUTER SCIENCE	90%	$93\% \pm 2\%$	3.5 ± 0.2
	80%	$83\% \pm 2\%$	3.1 ± 0.2
HIGH SCHOOL COMPUTER SCIENCE	90%	$93\% \pm 2\%$	3.2 ± 0.2
	80%	$82\% \pm 3\%$	2.7 ± 0.1
COMPUTER SECURITY	90%	$94\% \pm 3\%$	2.9 ± 0.1
	80%	$83\% \pm 2\%$	2.2 ± 0.1



(a) anatomy

(b) business ethics

(c) clinical knowledge

(d) college chemistry

(e) college computer science

(f) college medicine

(g) computer security

(h) formal logic

(i) high school biology

(j) high school computer science

(k) machine learning

(l) management

(m) marketing

(n) professional accounting

(o) professional medicine

(p) public relations

Figure 9: **Maximum softmax confidence does not represent true probability.** Deviation of softmax confidence from the probability of being correct for each subject. ECE is the expected calibration error, and MCE is the maximum calibration error.

A.1. Sample GPT-4- and MMLU-based one-shot questions for selected subjects.

A.1.1. College Computer Science

GPT-4 Based One-shot Questions

```
Which of the following sorting algorithms has the best average case performance?
    A. Bubble Sort
    B. Quick Sort
    C. Selection Sort
    D. Insertion Sort
The correct answer is option: B

What does the term "Big O Notation" describe in Computer Science?
    A. The speed of a computer
    B. The operating system version
    C. The size of a database
    D. The time complexity of an algorithm
The correct answer is option: D

What does HTTP stand for in terms of web technology?
    A. Hyper Text Transfer Portal
    B. Hyper Transfer Protocol
    C. Hyper Text Transfer Protocol
    D. High Transfer Text Protocol
The correct answer is option: C

In object-oriented programming, what is 'inheritance' used for?
    A. To distribute data across multiple databases
    B. To share methods and fields between classes
    C. To encrypt data before storing it
    D. To speed up program execution
The correct answer is option: B

Which of the following data structures is non-linear?
    A. Array
    B. Stack
    C. Tree
    D. Queue
The correct answer is option: C
```

MMLU Based One-shot Questions

```
An integer c is a common divisor of two integers x and y if
and only if c is a divisor of x and c is a divisor of y.
Which of the following sets of integers could possibly be
the set of all common divisors of two integers?
    A. {-6,-2, -1, 1, 2, 6}
    B. {-6, -2, -1, 0, 1, 2, 6}
    C. {-6, -3, -2, -1, 1, 2, 3, 6}
    D. {-6, -3, -2, -1, 0, 1, 2, 3, 6}
The correct answer is option: C.

You want to cluster 7 points into 3 clusters using the k-Means
Clustering algorithm. Suppose after the first iteration, clusters
C1, C2 and C3 contain the following two-dimensional points: C1
```

```
contains the 2 points: {(0,6), (6,0)} C2 contains the 3 points:
{(2,2), (4,4), (6,6)} C3 contains the 2 points: {(5,5), (7,7)}
What are the cluster centers computed for these 3 clusters?
    A. C1: (3,3), C2: (4,4), C3: (6,6)
    B. C1: (3,3), C2: (6,6), C3: (12,12)
    C. C1: (6,6), C2: (12,12), C3: (12,12)
    D. C1: (0,0), C2: (48,48), C3: (35,35)
The correct answer is option: A.

Consider the collection of all undirected graphs with 10 nodes
and 6 edges. Let M and m, respectively, be the maximum and
minimum number of connected components in any graph in the
collection. If a graph has no selfloops and there is at most
one edge between any pair of nodes, which of the following is true?
    A. M = 10, m = 10
    B. M = 10, m = 1
    C. M = 7, m = 4
    D. M = 6, m = 4
The correct answer is option: C.

Which of the following statements describe(s) properties of
a purely segmented memory system?
I. It divides memory into units of equal size.
II. It permits implementation of virtual memory.
III. It suffers from internal fragmentation.
    A. I only
    B. II only
    C. III only
    D. I and III
The correct answer is option: B.

Which of the following statements about floating-point arithmetic is NOT true?
    A. It is inherently nonassociative because some numbers have no      exact representation
    B. It is inherently nonassociative because there have to be upper
       and lower bounds on the size of numbers.
    C. Associativity can be achieved with appropriate roundoff
       conventions.
    D. Some rational numbers have no exact representation.
The correct answer is option: C.
```

A.1.2. PROFESSIONAL ACCOUNTING

GPT-4 Based One-shot Questions

```
Which of the following is used in accounting to analyze the
financial health of a business?
    A. Horizontal analysis
    B. Vertical analysis
    C. Ratio analysis
    D. All of the above
The correct answer is option: D

What does the acronym GAAP stand for in accounting?
    A. General Accepted Accounting Principles
    B. Global Accepted Accounting Procedures
```

```
    C. General Applied Accounting Procedures
    D. Global Applied Accounting Principles
The correct answer is option: A

What is the basic accounting equation?
    A. Assets = Liabilities + Owner's Equity
    B. Assets = Liabilities - Owner's Equity
    C. Assets + Liabilities = Owner's Equity
    D. Assets - Liabilities = Owner's Equity
The correct answer is option: A

What is a balance sheet used for in accounting?
    A. To record the day-to-day financial transactions
    B. To determine the company's financial position at a specific point in time
    C. To track the company's cash flows
    D. To record the company's sales revenue
The correct answer is option: B

Which of the following best describes accrual accounting?
    A. Revenue and expenses are recorded when they are received and paid
    B. Revenue and expenses are recorded when they are earned and incurred
    C. Revenue and expenses are recorded at the end of the financial year
    D. Revenue and expenses are recorded at the start of the financial year
The correct answer is option: B
```

MMLU Based One-shot Questions

```
Arno Co. did not record a credit purchase of merchandise made
prior to year end. However the merchandise was correctly included
in the year-end physical inventory. What effect did the omission
of reporting the purchase of merchandise have on Arno's balance
sheet at year end? Assets Liabilities
    A. No effect No effect
    B. No effect Understated
    C. Understated No effect
    D. Understated Understated
The correct answer is option B.

Which of the following procedures would an auditor generally perform
regarding subsequent events?
    A. Inspect inventory items that were ordered before the year end
       but arrived after the year end.
    B. Test internal control activities that were previously reported
       to management as inadequate.
    C. Review the client's cutoff bank statements for several months
       after the year end.
    D. Compare the latest available interim financial statements with
       the statements being audited.
The correct answer is option D.

The FASB makes changes to the Accounting Standards Codification by issuing
    A. Accounting Standards Updates.
    B. Emerging Issues Task Force Releases.
    C. Statements of Financial Accounting Standards.
```

```
    D. Staff Technical Bulletins.
The correct answer is option A.

On July 1 year 7 Dean Co. issued at a premium bonds with a due date
of July 1 year 12. Dean incorrectly used the straight-line method
instead of the effective interest method to amortize the premium.
How were the following amounts affected by the error at June 30
year 12? Bond carrying Retained amount earnings
    A. Overstated Understated
    B. Understated Overstated
    C. Overstated No effect
    D. No effect No effect
The correct answer is option D.

A company recently moved to a new building. The old building is
being actively marketed for sale, and the company expects to
complete the sale in four months. Each of the following statements
is correct regarding the old building, except:
    A. It will be reclassified as an asset held for sale.
    B. It will be classified as a current asset.
    C. It will no longer be depreciated.
    D. It will be valued at historical cost.
The correct answer is option D.
```

A.1.3. CLINICAL KNOWLEDGE

GPT-4 Based One-shot Questions

```
Which of the following is the most common cause of community-acquired pneumonia?
    A. Streptococcus pneumoniae
    B. Haemophilus influenzae
    C. Klebsiella pneumoniae
    D. Pseudomonas aeruginosa
The correct answer is option: A

Which hormone is primarily responsible for regulating blood calcium levels?
    A. Calcitonin
    B. Parathyroid hormone
    C. Thyroxine
    D. Insulin
The correct answer is option: B

What is the most common cause of acute pancreatitis?
    A. Gallstones
    B. Alcohol
    C. Hypertriglyceridemia
    D. Medications
The correct answer is option: A

What is the most common cause of secondary hypertension?
    A. Renal artery stenosis
    B. Pheochromocytoma
    C. Hyperaldosteronism
    D. Cushing's syndrome
The correct answer is option: A
```

```
Which of the following is a common extraintestinal manifestation
of ulcerative colitis?
    A. Erythema nodosum
    B. Gallstones
    C. Uveitis
    D. All of the above
The correct answer is option: D
```

MMLU Based One-shot Questions

```
The key attribute in successful marathon running is:
    A. strength.
    B. power.
    C. stride length.
    D. stamina.
The correct answer is option D.

Which of the following is NOT a symptom of anaphylaxis?
    A. Stridor.
    B. Bradycardia.
    C. Severe wheeze.
    D. Rash.
The correct answer is option B.

In what situation are closed pouches applied?
    A. The patient has a semi-formed or liquid output.
    B. The patient has a colostomy.
    C. In the immediate post-operative period.
    D. The patient has a urostomy.
The correct answer is option B.

With an increasing number of sprints the:
    A. anaerobic contribution progressively increases.
    B. pH of the muscle falls below 6.0.
    C. blood glucose concentration falls below 3 mmol/L.
    D. relative contribution of aerobic metabolism increases.
The correct answer is option D.

Which of the following is true in diplopia?
    A. Diplopia can never occur if one eye is covered
    B. The outer image is always the false image
    C. A fourth nerve palsy occurs when the patient looks upwards
    D. A sixth nerve palsy causes a divergent squint
The correct answer is option B.
```