

Clustering Zero-Shot Uncertainty Estimations to Assess LLM Response Accuracy for Yes/No Q&A

Christopher T. Franck^{1*}, Amy Vennos¹, W. Graham Mueller², Daniel Dakota²

¹Department of Statistics, Virginia Tech

²Leidos Holdings, Inc.

{chfranck, avennos}@vt.edu

{william.g.mueller, daniel.d.dakota}@leidos.com

Abstract

The power of Large Language Models (LLMs) in user workflows has increased the desire to access such technology in everyday work. While the ability to interact with models provides noticeable benefits, it also presents challenges in terms of how much trust a user should put in the system's responses. This is especially true for external commercial and proprietary models where there is seldom direct access and only a response from an API is provided. While standard evaluation metrics, such as accuracy, provide starting points, they often may not provide enough information to users in settings where the confidence in a system's response is important due to downstream or real-world impact, such as in Question & Answering (Q&A) workflows. To support users in assessing how accurate Q&A responses from such black-box LLMs scenarios are, we develop an uncertainty estimation framework that provides users with an analysis using a Dirichlet mixture model accessed from probabilities derived from a zero-shot classification model. We apply our framework to responses on the BoolQ Yes/No questions from GPT models, finding the resulting clusters allow a better quantification of uncertainty, providing a more fine-grained quantification of accuracy and precision across the space of model output while still being computationally practical. We further demonstrate its generalizability and reusability of the uncertainty model by applying it to a small set of Q&A collected from U.S. government websites.

1 Introduction

Large Language Models (LLMs) have substantially influenced a multitude of workflow applications, such as question and answering (Q&A) systems. While the expansive knowledge and response capabilities of generative models (e.g., GPT4) has been impressive, it also presents unique challenges

in workflow integration, namely user trust and certainty in answers and responses. This is especially pertinent when a Q&A system is designed for non-subject matter experts who will not be familiar with the response quality of the domain.

This need has resulted in growing research in uncertainty estimation to better assess the quality of a response an LLM (Shelmanov et al., 2021). Recent methods have been developed to quantify and reduce uncertainty focused on classification tasks (Gal, 2016; Kuzmin et al., 2023) and text classification models (He et al., 2020; Zhang et al., 2019; Xin et al., 2021). However, obtaining such uncertainty estimates for many generative applications (e.g., responses in a Q&A system) accessing proprietary models, such as GPT4, is not straightforward, since the uncertainty cannot be meaningfully characterized without access to the underlying probabilities.

We quantify uncertainty in terms of the predicted probability of responses. Since many current LLMs, especially proprietary models (e.g., GPT4), do not automatically furnish probabilities in their responses for a specific task or classification (e.g., Yes/No Q&A), we use a GPT-BART pipeline (see section 3) as a proxy for LLM uncertainty. The proposed method only requires probability predictions and labeled training data and thus could be implemented on future LLMs that do directly provide probabilities for tasks.

To support users in assessing responses from such models, we develop a framework which uses probability distributions from a zero-shot classification (BART-MultiLNI (Williams et al., 2018)) with a Dirichlet Mixture Model Clustering approach based on a customized version of the Expectation Maximization algorithm (EM; Dempster et al., 1977). We apply our framework to Yes/No Q&A, which remains a surprisingly difficult task subject to lower-than-expected accuracy (Clark et al., 2019). An analysis of the clusters of questions us-

*Corresponding author.

ing conformal prediction show support for users in better understanding the level of confidence an LLM so that the user can trust its responses, especially in a black-box LLM scenario. We subsequently apply our fitted general Wikipedia model to a specific questions relevant to government domains and still obtain a usable clustering analysis.

2 Related Work

2.1 Accuracy

LLM accuracy is widely studied. Metrics to quantify accuracy in LLMs for different applications include Exact Match (EM; Chang et al., 2024), F1 score (Koike et al., 2024), and ROUGE (Mishra et al., 2023). Specifically, work has been done to evaluate the accuracy and performance of specific LLMs on task specific tests. For example, ChatGPT was shown to pass the United States Medical Licensing Exam (USMLE; Kung et al., 2023) and performed well on a neurology board exam with an accuracy rate of 85% (Erdogan, 2024), in addition to showing an 86.8% overall accuracy rate when asked questions related to bariatric surgery (Samaan et al., 2023).

In terms of evaluating the accuracy of Yes/No questions, Clark et al. (2019) extensively discusses the accuracy of different models on the BoolQ dataset, with a BERT model additionally pretrained on MultiNLI producing the most accurate results at 80.4% (Clark et al., 2019). Additionally, the developers of the BoolQ_{3L} dataset provide a thorough discussion comparing the accuracy of LLMs on the BoolQ versus BoolQ_{3L} datasets (Sulem et al., 2022).¹

2.2 Uncertainty

There is a need to look for methods for black-box LLM uncertainty estimations (Xiong et al., 2024), with LLM verbalization (Lin et al., 2022), probing (Harsha Tanneru et al., 2024) and semantic sampling (Aichberger et al., 2024) having been explored. For Yes/No question, uncertainty in responses is a known problem (de Marneffe et al., 2009), as often the response itself does not take

¹The BoolQ_{3L} is composed by remapping the original BoolQ questions to corresponding passages that do not contain sufficient information to answer the question. While it does provide the addition of “I Don’t know” as an answer, we only focus on sending the questions to the model and not the corresponding passages, thus the dataset does not provide additional benefits over the standard BoolQ for our experimental setup.

form of Yes/No and requires inferences.² The recent rise in datasets created to allow uncertain responses highlights the importance of examining uncertainty in question-answering LLMs (Rajpurkar et al., 2018; Rogers et al., 2020; Wang et al., 2020). Analyzing how LLMs quantify uncertainty is motivated by several factors, one being to decrease the rate and effects of hallucinations in Q&A applications (Ji et al., 2023).

3 Experimental Setup

3.1 Data

We use the BoolQ dataset, a reading comprehension dataset consisting of 9,427 Yes/No questions derived from Wikipedia with human-annotated answers (Clark et al., 2019) to develop our model. However, we only utilize the questions and *do not use the passages* in our experiments, relying solely on the LLM’s internal knowledge to answer the question. In addition to its size, we find the wide-coverage of question types within the BoolQ a good proxy for assessing the ability of an LLM to cover a wider range of general knowledge topics. While we recognize additional LLM pre-training of a model may improve performance for domain specific questions, this is beyond the scope of this work. Furthermore, many commercial enterprises will not have such an option readily available.

Though the BoolQ dataset was originally created to only contain a response of “Yes” or “No”, we investigate the benefit of an additional response type “I don’t know” (see section 3.3). To validate the transferability of our model to a domain specific real-world scenario, we construct a small set of 25 questions from two government websites covering customs and import/export regulations³ and the electronic code of federal regulations⁴ with a specific focus on immigration regulations (Title 8). This allows us to 1) identify how accessible such publicly available data is in the model and 2) assess how the model performs on a more specific domain.

3.2 GPT4 Answer Probabilities

Assessing the accuracy and precision of responses requires the LLM to reliably answer in terms of only three categories {“Yes”, “No”, “I don’t

²See section 3.2 for indications that current LLMs still frequently do not fulfill this request even when explicitly prompted.

³<https://www.cbp.gov/>

⁴<https://www.ecfr.gov/>

Zero Shot Classification Pipeline

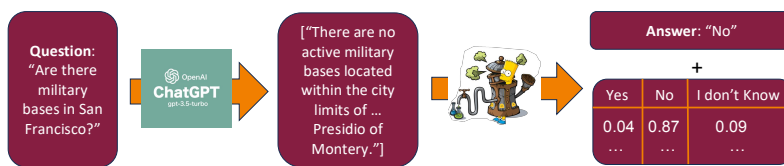


Figure 1: Our Zero-Shot Classification Pipeline with an example Yes/No question. BoolQ Train Dataset question 5124 is fed through GPT4 model, generating an imprecise output lacking probabilities. The response is fed to the facebook/bart-large-mnli Transformers model returning needed answer probabilities for analysis.

Evaluation Pipeline

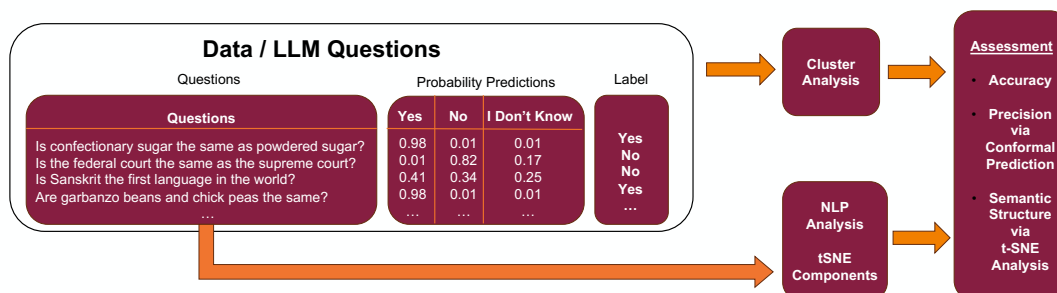


Figure 2: An illustration of the data analytic pipeline. Yes/No questions, labels, and probability predictions are fed into the cluster model. The Yes/No questions are embedded using sentence transformers and dimensionality reduction is performed with UMAP. Analysis examines measures of accuracy and precision within clusters and semantic structure related to highly uncertain answers.

know”}, while also providing the probability of each of these three answers, something many commercial and proprietary models do not readily distribute.

We first assessed the ability of both GPT4 and GPT4o to produce the probability predictions required for our clustering algorithm by adopting the prompting strategy of Zhou et al. (2023) and send the BoolQ question in addition to explicitly instructing the model to return a 0-1 confidence for its response.⁵ For GPT4 this took ≈ 16.47 seconds per API call (≈ 43.14 total hours), while for GPT4o took ≈ 16.79 per API call (≈ 43.95 total hours). Analyzing the responses allowed us assess the feasibility of automating the processing of analyzing responses by (i) examining a small collection of the outputs manually, and (ii) programmatically assessing rates at which the instructions were followed.

Among the responses we manually observed, the last lines included a single numeric response between 0 and 1, a stylistic string such as "““““",

⁵See Appendix A for prompt template and Appendix B for an example response.

prose, and one of "Yes", "No", or "I don't know", sometimes followed by a numeric score between 0 and 1. In total, the last line was numeric in 92.6% cases for the GPT4 model, and 75.5% of cases for the GPT4o model. It would thus take substantial follow-up intervention by a human to process answers suitably for aggregated analysis (or to further refine prompting strategy), making this strategy less scalable. Accurately extracting the confidence scores from these non-uniform responses would be even more difficult and likely prone to missing values.

3.3 Zero-Shot Classification Probabilities

To obtain probabilities for responses, we use a zero-shot LLM classification pipeline (depicted in Figure 1). We first send only the BoolQ questions without any context or prompt template to a GPT model, relying solely on the model's internal knowledge for its response to the question. For GPT4 this took ≈ 1.9 seconds per API call (≈ 20.2 total hours), half the amount of time than with our prompt template used in section 3.2, while GPT4o took ≈ 17.3

per API call (≈ 45.2 total hours).⁶

To obtain a probability of all potential responses, we pass each response to BART (Lewis et al., 2020a), specifically the Multi-Genre Natural Language Inference (MultiLNI) task (Williams et al., 2018) variant⁷ which took about 2 seconds per inference (≈ 5 hours total). This model enables zero-shot classification given a set of predetermined labels (in our case, “Yes”, “No”, and “I don’t know”) and provides a probability score that reflects BART’s confidence of each respective label. This approach allows us to both (1) classify responses into one of the desired categories, and (2) access a set of probability estimates and thus uncertainty of various responses.

4 Production and Assessment of Clusters

Figure 3 shows a ternary plot with three probability axes corresponding to “Yes”, “No”, and “I don’t know” outputs in three dimensions for GPT4. The goal is to characterize each of these clusters using the observed Q&A data \mathbf{y} . While several established clustering approaches exist, we have implemented a specific approach that obtains clusters of Q&A probabilities in their natural sum-to-one space. Our approach uses the EM algorithm (Dempster et al., 1977) for clustering with individual cluster densities that follow the Dirichlet distribution (Kotz et al., 2004), which automatically constrains the Q&A probabilities to sum to one.

4.1 Dirichlet Mixture Model Clustering via EM algorithm

We specify $K = 4$ clusters based on inspection of Figure 3. Each of these clusters has a shape governed by density function $f_k(\cdot)$ for $k = 1, \dots, 4$. The three-dimensional distribution $f(\mathbf{y})$ of the Q&A probabilities is a weighted average of the clusters according to the following mixture model:

$$f(\mathbf{y}) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}, \boldsymbol{\theta}_k). \quad (1)$$

The EM algorithm takes the observed data \mathbf{y} and user-specified K , then learns the values of the cluster sizes π_k interpreted as the proportion of points

⁶It is not known why GPT4o took longer to answer questions without a prompt than with one at this time. One potential reason may have been quota limits at the time of the API calls.

⁷facebook/bart-large-mnli available via HuggingFace API.

that belong in the k th cluster. The algorithm also estimates the Dirichlet shape parameters $\boldsymbol{\theta}_k$, which govern the shape of clusters as shown in the ternary plot in Figure 3.

While the EM algorithm is a well established, our contribution is its implementation making use of Dirichlet cluster densities $f_k(\cdot)$. Surprisingly, this is not readily available in other clustering-based implementations of the EM algorithm, e.g., (Benaglia et al., 2009; Wu, 2023).⁸ Upon convergence, this algorithm provides the user with cluster sizes and shapes, and assignments of each data point to the most appropriate cluster. We refer to the process of placing points in the most likely mixture model component as “clustering” as this is the common use of this term in the statistical literature (McLachlan and Peel, 2004). We have found our implementation of the EM algorithm to be robust to several different starting value specifications and only took ≈ 65 seconds per run.

4.2 Evaluation and Analysis

We report the accuracy rate and precision via conformal prediction (see section 4.3) both in the presence and absence of the cluster structure determined by the EM algorithm. We also report the weights and shape parameter estimates obtained by the EM algorithm. To assess accuracy rate, we determine how often the highest probability answer agrees with the true label for each question. We note that “I don’t know” is allowed as an answer, though this label does not appear in the BoolQ set. To avoid considering “I don’t know” as a wrong answer, our primary accuracy rate does not include questions for which the “I don’t know” answer has the highest probability. We assess accuracy on the full 9,427 question/answer pairings in the BoolQ training data set, and we provide 95% confidence intervals for these rates.

4.3 Conformal Prediction

The Q&A probability predictions for “Yes”, “No”, and “I don’t know” frequently indicate a reasonably high level of uncertainty. For example, one question shown in Figure 2 reads “is Sanskrit the first language of the world”. Zero-shot classification provides probabilities of 41% for “Yes”, 34% for “No”, and 25% for “I don’t know”. While the most probable answer is 41% for “Yes”, it is difficult to glean any clear course of action from this

⁸See Appendix D for more details on our approach.

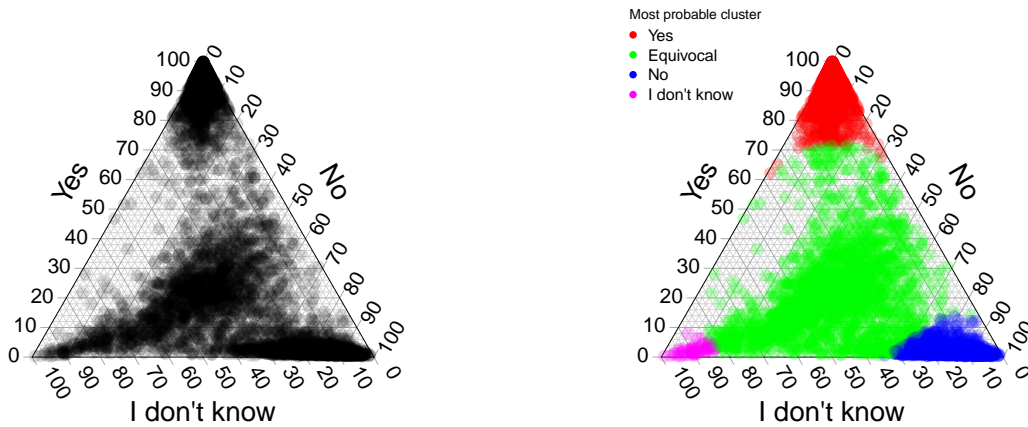


Figure 3: Ternary plots that show the probability predictions for “Yes,” “No,” and “I don’t know” categories for the 9,427 BoolQ questions using GPT4 shown in Figure 1. Points are semi-transparent to assist with visualizing concentration. The left panel shows the probability predictions and the right panel color codes those same predictions by the cluster obtained using the methods described in Section 4.

collection of uncertain probabilities, since none of the probabilities are close to 100%. In fact, the true label is “No” and thus the most probable answer is incorrect in this instance.

To better understand the extent to which the Q&A probabilities are indecisive, we subject our Q&A probability predictions to conformal prediction (Vovk et al., 2005) in order to obtain a set of answers that contains the truth with a user-specified high level of probability. Conformal prediction holds out a separate calibration set which is used to learn the threshold a probability prediction needs to be above in order to be included in the conformal set. Thus conformal prediction in the classification problem works by expanding the size of a prediction set until the probability that the true label is within the prediction set reaches the user-specified requirement, which we set to a standard value 90% following Angelopoulos and Bates (2023). Expanding the size of the answer set increases accuracy of the prediction set to 90% at the cost of reducing precision of the answer. In general, higher inclusion probability requirements lead to larger conformal prediction sets.

Conducting conformal prediction is accomplished by randomly selecting and holding out a calibration data set of 2,000 from the BoolQ training set, then using the calibration set to establish the probability threshold that an answer has to be above in order to be included in the conformal pre-

diction set. Then, the remaining 7,427 probability outputs are compared against the threshold to produce the conformal prediction set.⁹ This process is essentially instantaneous once the predictions are available, and we summarize the rates at which each answer appears in the conformal set, overall and within each cluster.

5 Results

Seen in Figure 3, the ternary plots visualize the probability predictions in terms of each answer: “Yes”, “No”, and “I don’t know”. The left panel shows that probabilities sum to one for each question, and there appear to be $K = 4$ clusters in the data. The right panel shows the result of our clustering approach. This analysis shows three specific virtues:

1. The size and shape of the clusters are determined automatically based on the three dimensional distribution of the data, obviating the need for a human to pre-specify decision thresholds.
2. Even though the clusters were determined automatically, they are readily interpretable and easy to visualize for humans. Responses that appear in clusters with higher accuracy than the overall analysis may be more trustworthy

⁹For an excellent tutorial for conformal prediction see Angelopoulos and Bates (2023).

| Description (color) | Accuracy rate (95% CI) | Cluster size $\hat{\pi}_k$ | Parameter estimates $\hat{\theta}_k$ |
|--------------------------------|------------------------|----------------------------|--------------------------------------|
| Probably Yes (Red) | 88.4% (87.5% - 89.3%) | 0.51 | (30.52, 0.99, 1.14) |
| Probably No (Blue) | 77.1% (75.3% - 78.9%) | 0.24 | (1.08, 35.53, 5.60) |
| Equivocal predictions (Green) | 59.4% (56.7% - 62.1%) | 0.25 | (1.76, 3.38, 3.23) |
| Probably I don't know (Purple) | - | 0.01 | (2.13, 7.34, 77.98) |
| No clustering (Black) | 80.8% (79.9% - 81.6%) | 1.00 | - |

Table 1: Accuracy rates, confidence intervals, and estimates for cluster size and shape parameters when GPT4 is used. Results are presented overall and for the clustering approach. Accuracy rate is based on the most probable answer to each question. Color corresponds to the clusters visualized in the right panel of Figure 3.

| Description (color) | One label | All labels | Yes | No | I don't know |
|--------------------------------|-----------|------------|--------|--------|--------------|
| Probably Yes (Red) | 51.5% | 15.8% | 100.0% | 28.7% | 35.6% |
| Probably No (Blue) | 2.1% | 10.6% | 10.8% | 100.0% | 97.7% |
| Equivocal predictions (Green) | 0.0% | 90.8% | 91.1% | 99.8% | 99.8% |
| Probably I don't know (Purple) | 6.5% | 17.7% | 17.7% | 93.5% | 100.0% |
| No clustering (Black) | 26.8% | 32.8% | 75.8% | 63.6% | 66.6% |

Table 2: Results of the conformal prediction exercise on the 7,427 available answers for GPT4. Percentages indicate how many questions included a single answer label, all three answer labels, and individual inclusion of “Yes”, “No”, and “I don't know” labels. Results are presented overall and by cluster.

than questions that land in low-accuracy clusters.

3. A by-cluster analysis of accuracy, precision, and semantic structure is more informative than an analysis which ignores clusters, and thus helps humans understand the conditions under which LLM answers can be trusted confidently.

5.1 Cluster Accuracy BoolQ

Table 1 provides overall and by-cluster accuracy rates and also maximum likelihood estimates of cluster size $\hat{\pi}_k$ and cluster-specific shape parameters $\hat{\theta}_k$. About half of the questions are in the “Probably Yes” cluster, with 24%, 25%, and 1% of questions in each of the “Probably No”, “Equivocal predictions”, and “Probably I don't know” clusters, respectively. This analysis shows that our approach has higher accuracy for questions in the Probably “Yes” cluster (88.4%) compared with an overall analysis that does not implement clustering (80.8%). Accuracy of the most probable answer is lower within the equivocal predictions cluster (59.4%), and accuracy in the Probably “No” cluster (77.1%) is statistically closer with the “Overall - No Clustering” strategy. A user of this analysis would thus know that they are able to make relatively more accurate decisions based on questions where the answer probabilities fall in the red cluster (“Probably Yes”) compared with other clusters or

when eschewing a cluster analysis altogether.

5.2 Conformal Prediction

Table 2 shows the results of the conformal prediction exercise on the remaining 7,427 available answers not used for calibration to assess precision. Since we used conformal prediction to obtain prediction sets that have a fixed 90% chance of containing the true label, we view conformal prediction sets with a smaller number of answers in them to be more precise than conformal sets that have a greater number of answers. Conformal prediction is thus useful since it indicates how decisive the most probable answer is. For example, the overall analysis indicates that the no clustering approach is highly indecisive for 32.8% of questions, as all three answers are included in the conformal set. In the no clustering approach only 26.8% of the questions have highly precise predictions, as these include a single answer in the conformal set. Within the clusters, however, the story is different as 51.5% of the questions in the “Probably Yes” cluster have a single label, while 15.8% contain all three labels.¹⁰

Table 2 indicates how precise the Q&A probability answers tend to be within each cluster, and how the cluster-level analyses differ substantially from an overall analysis that does not account for a clustering structure. This is useful since the user

¹⁰See Appendix C for GPT4o Results.

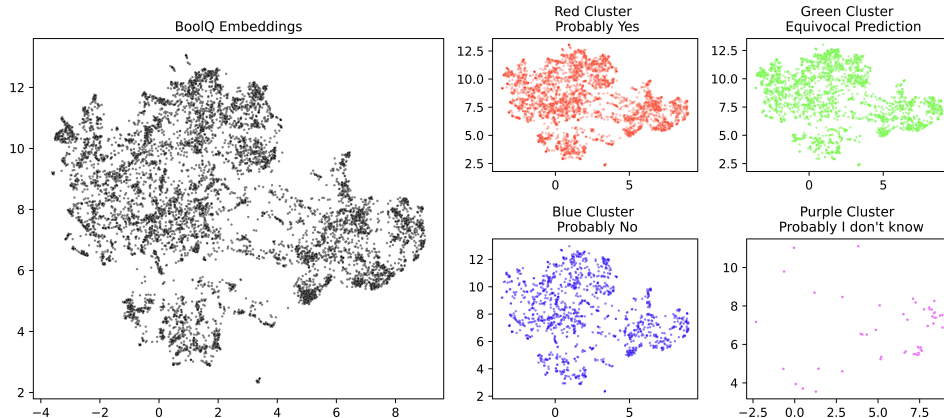


Figure 4: Plots of the two UMAP components for each question in the BoolQ analysis. Left panel shows overall distribution of components. Right panel shows the distribution of components within each of the four clusters identified by the Em algorithm. The probably “I don’t know” cluster (bottom right of right panel) appears to differ in distribution from the rest.

can, for the set of questions they are particularly interested in, determine which cluster the answers belong in, then assess how precise those answers are and note any improvement in precision they obtain over an analysis that does not involve clustering. Returning to our “is Sanskrit the first language of the world” example, the most probable answer of “Yes” at 41% is actually incorrect. The present analysis reveals that the conformal set for this question contains all three answers and thus is imprecise. That is, a user who wanted to assemble the smallest set that would have at least a 90% chance of including the truth would not be able to eliminate any answers from consideration.

5.3 Semantic Investigation

We generate embeddings for each question in the BoolQ using a sentence transformer (Reimers and Gurevych, 2019)¹¹ and use UMAP (McInnes et al., 2018) for dimensionality reduction to investigate any potential semantic patterns of interest. Figure 4 shows the results of the semantic analysis. While the distribution of components looks pretty similar in the overall analysis, the “I don’t know” cluster (purple) does show some potential differentiation. When looking at questions in this specific cluster, we see some commonalities such as questions dealing with media and entertainment especially wrt. future events (e.g., “Will there be a 13th season of Criminal Minds”) as well specific plot knowledge (“Did the Robinsons make it back to Earth”).

Potential reasoning could be the “futuristic” na-

¹¹Specifically we use sentence-transformers/all-mpnet-base-v2 based on Song et al. (2020) via HuggingFace API.

| Description (color) | Number of prompts | Accuracy rate (95% CI) |
|--------------------------------|-------------------|-------------------------|
| Probably Yes (Red) | 12 | 83.3% (51.6% - 97.9%) |
| Probably No (Blue) | 3 | 100.0% (29.2% - 100.0%) |
| Equivocal predictions (Green) | 5 | 60.0% (14.7% - 94.7%) |
| Probably I don’t know (Purple) | 0 | - |
| No clustering (Black) | 20 | 80% (56.3% - 94.3%) |

Table 3: Accuracy rates for the U.S. government websites using the GPT4 fitted model. Note that observations with “I don’t know” as the most probable answer are not included in this analysis.

ture in combination with information and answers BART was exposed to during training. While a commercial LLM’s response (in our case GPT4) may be able to be updated with newer information that might help discriminate contextual real-world knowledge and provide new information to resolve “futuristic” questions this may not directly be transferable in a zero-shot classification model that is restricted primarily to the model’s internal knowledge at the time of training.

6 Cluster Accuracy U.S. Government Websites

To assess the Dirichlet clustering model’s predictive capability beyond the BoolQ training set, we applied the learned clustering rule to a small set of 25 Q&A questions from U.S. government websites (see section 3.1). Questions were pre-appended with either “I am an immigration specialist” or “I am an import/export control specialist” respectively before being sent to GPT4.

Applying the GPT4 fitted BoolQ uncertainty model without any further clustering, the most probable answer was “Yes” 12 times, “No” 8

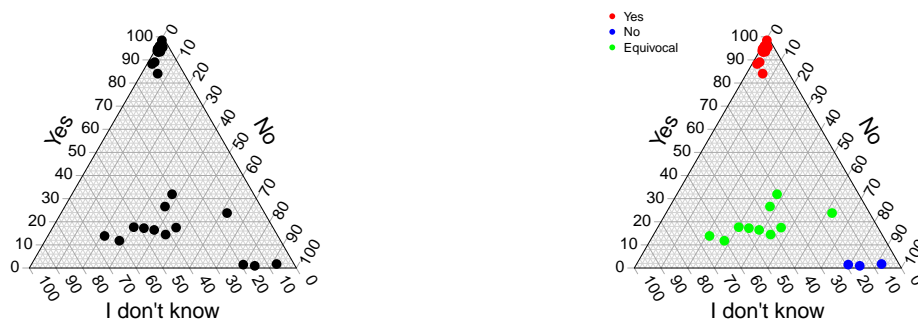


Figure 5: Ternary plots that show the probability predictions for “Yes,” “No,” and “I don’t know” categories for the 25 U.S. government websites Q&A prompts using the cluster rule learned from BoolQ analysis and GPT4. The left panel shows the probability predictions and the right panel color codes those same predictions by cluster assignment. No observations were observed in the “Probably I don’t know” cluster in the bottom left of the ternary plots. 4.

times, and “I don’t know” 5 times. Among the 20 “Yes” and “No” predictions, the accuracy rate is $16/20=80\%$. Using prediction based on the clustering approach, 15 out of 25 predictions are in either the “Probably Yes” or “Probably No” clusters, and 10 observations are in the “Equivocal” cluster. No observations appeared in the “Probably I don’t know” cluster. These predictions can be seen in the right panel of Figure 5. Table 3 shows the accuracy rate and confidence intervals for the most probable answer within clusters. While the small sample size precludes the ability to make definitive statements about statistical significance, the overall pattern of higher accuracy in the “Yes” and “No” clusters and lower accuracy in the “Equivocal” cluster is similar to what we observed with the BoolQ analysis.

When we look at some of the questions and both the response and the zero-shot probabilities, there are several instances in which GPT4 correctly answers in the text, but the zero-shot classification is not overly confident or ultimately wrong, even when the questions are on similar topics. For example, the question “Does an ESTA grant me entry to the US?” is correctly answered in the GPT4 response and while the zero-shot classification is also correct (“No”), it only achieves a 43% probability from the model (compared to 41% “I don’t know”). While the question “Is an an ESTA a visa?” is also correctly answered by GPT4, it receives much higher probability of “No” at 79% in its zero-shot classification. In another instance, the question “Are travelers checks considered money as defined by the Customs and Border Protection?”

is correctly answered by GPT4 (“Yes”), but the zero-shot classification is incorrect with “No” (37%), although all the probabilities are rather close indicating potential indecision.

These results however demonstrate that we can successfully optimize our uncertainty model on larger more general datasets of Q&A responses and effectively apply them to smaller, more domain specific datasets and achieve the same desired effect of identifying question responses where a user can make relatively more accurate decisions.

7 Conclusion

We developed a Dirichlet Mixture Model Clustering via EM algorithm framework for LLM Yes/No Q&A response certainty. Our approach zero-shot pipeline is particularly applicable for when the underlying probabilities are not available in the initial response from an LLM. Importantly, our approach is model independent, reusable, computationally efficient, and can be applied to any zero-shot pipeline where we have access to both the category labels and underlying probabilities. Our by cluster analyses reveal a more fine-grain analysis of accuracy, precision, and semantic similarities than without its implementation. A user is thus provided more information about if and under what conditions they can have more certainty in trusting responses for decision making, especially in domains in which they lack certainty.

While we limited ourselves to only Yes/No questions here, the framework can be extended to additional cases with a known, finite set of responses

(e.g., classification tasks or categorical responses) and has future potential integration with in-context learning (Brown et al., 2020) and to more effectively support retrieval-augmented generation (RAG) systems (Lewis et al., 2020b).

Limitations

We view using zero-shot classification probabilities from another LLM as a derivative of LLMs-as-a-Judge (Zheng et al., 2023), and assumes our approach is sufficient and reliant enough for scalability. Given that LLM-as-a-Judge has shown variable research (Shen et al., 2023; Hada et al., 2024) and factuality questions arise (Fu et al., 2023), there are still open questions and active research examining the reliability and effectiveness of various approaches using any LLM-as-a-Judge framework and any of its derivatives. Our developed method requires only class probabilities and labeled training data to be useful, and could be readily deployed on a future LLM that furnishes Q&A probabilities. But we recognize that our current approach for LLM uncertainty is affected by the BART model processing and probability generations and may show variable outcomes using different models.

Model creativity may potentially influence our framework’s stability. The framework would optimally work assuming that responses are static (i.e., have low or zero temperature settings) and are consistently classified by the zero-shot model. Additional experiments would need to be performed to determine how consistent the clustering approach is when dealing with higher temperatures and more volatility in classifications.

Ternary plots are ideal for visualizing cluster structures in three dimensions where the variables sum to a constant. In higher dimensions, i.e., tasks with more than three categorical outputs, our method still works since the EM algorithm extends trivially to higher dimensions. However, the visualization aspect will be more burdensome and assessing the effectiveness of the clustering structure visually might require examining multiple two and three dimensional plots.

While a GUI is not currently available, such a feature would be a worthwhile future endeavor that would enable a better UX in understanding whether and when to trust an LLM responses for Q&A tasks.

Ethics Statement

Using an LLM for zero-shot classification runs the risk of adding the model’s inherent bias when making classification decisions. We would advise attempting to ascertain data lineage and sources for training when selecting an LLM for zero-shot applications, as finding a neutral or domain relevant would help reduce these issues. However, given that many vendor LLMs are more black-box in nature with respect to ascertaining many of the training and implementation details, it is important to adequately examine and assess if the selected LLM is appropriate for the given data and task to reduce any negative impact such bias may have on a downstream application.

Given the use of U.S. government websites, it is important to take into consideration the ramifications of any incorrect answer generated at any step in the process, from a the initial question response from the black-box model, to the zero-shot model classification probabilities, to the uncertainty model. For this reason, it is imperative to also inform the user of the risks relying solely on any automatically generated answer on such important topics from such a system poses. An incorrect or misunderstood response runs the risk of a substantial negative real-world consequences on an individual, thus it is still important to provide individuals the relevant sources of information needed for any desired self-verification.

Acknowledgments

This document is export approved by Leidos for release under identification number **24-LEIDOS-1125-28579**.

References

- Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. [How many opinions does your LLM have? improving uncertainty estimation in NLG](#). In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Anastasios Angelopoulos and Stephen Bates. 2023. [Conformal Prediction: A Gentle Introduction](#).
- Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. 2009. [mixtools: An R package for analyzing finite mixture models](#). *Journal of Statistical Software*, 32(6):1–29.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

- Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. [A survey on evaluation of large language models](#). *ACM Trans. Intell. Syst. Technol.*, 15(3).
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. 2009. [Not a simple yes or no: Uncertainty in indirect answers](#). In *Proceedings of the SIGDIAL 2009 Conference*, pages 136–143, London, UK. Association for Computational Linguistics.
- Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.
- Mucahid Erdogan. 2024. [Evaluation of responses of the large language model GPT to the neurology question of the week](#). *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*.
- Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan Tn. 2023. [Are large language models reliable judges? a study on the factuality evaluation capabilities of LLMs](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 310–316, Singapore. Association for Computational Linguistics.
- Yarin Gal. 2016. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.
- Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. [Are large language model-based evaluators the solution to scaling up multilingual evaluation?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian’s, Malta. Association for Computational Linguistics.
- Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. [Quantifying uncertainty in natural language explanations of large language models](#). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1072–1080. PMLR.
- Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. [Towards more accurate uncertainty estimation in text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372, Online. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. [Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21258–21266.
- S. Kotz, N. Balakrishnan, and N.L. Johnson. 2004. *Continuous Multivariate Distributions, Volume 1: Models and Applications*. Continuous Multivariate Distributions. Wiley.
- Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, et al. 2023. [Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models](#). *PLOS Digital Health*, 2(2):e0000198.
- Gleb Kuzmin, Artem Vazhentsev, Artem Shelmanov, Xudong Han, Simon Suster, Maxim Panov, Alexander Panchenko, and Timothy Baldwin. 2023. [Uncertainty estimation for debiased models: Does fairness hurt reliability?](#) In *Proceedings of the 12th International Joint Conference on Natural Language Processing (IJCNLP-AACL)*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).
- G.J. McLachlan and D. Peel. 2004. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley.
- Nishant Mishra, Gaurav Sahu, Iacer Calixto, Ameen Abu-Hanna, and Issam Laradji. 2023. [LLM aided semi-supervision for efficient extractive dialog summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10002–10009, Singapore. Association for Computational Linguistics.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. [Getting closer to ai complete question answering: A set of prerequisite real tasks](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.
- Jad S. Samaan, Yoon-Kyo H. Yeo, Nikhil Rajeev, Logan Hawley, Samuel Abel, Wee-Hin Ng, Nivedhitha Srinivasan, Jessica Park, Maxine Burch, Russell Watson, Oryan Liran, and Kamran Samakar. 2023. [Assessing the accuracy of responses by the language model chatgpt to questions regarding bariatric surgery](#). *Obesity Surgery*, 33(6):1790–1796. Epub 2023 Apr 27.
- Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. [How certain is your Transformer?](#) In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. [Large language models are not yet human-level evaluators for abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.
- Elior Sulem, Jamaal Hay, and Dan Roth. 2022. [Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions](#). In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.
- Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020. [Reco: A large scale chinese reading comprehension dataset on opinion](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9146–9153.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Dongjie Wu. 2023. *em: Generic EM Algorithm*. R package version 1.0.0.
- Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. [The art of abstention: Selective prediction and error regularization for natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.
- 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](#). In *The Twelfth International Conference on Learning Representations*.
- Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. [Mitigating uncertainty in document classification](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*,

volume 36, pages 46595–46623. Curran Associates, Inc.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

A Prompt to examine if GPT can reliably produce probability predictions

You are evaluating a response that has been submitted for a particular task, using a specific set of standards. Below is the data:

[BEGIN DATA]

[Task]: The submission comes from a question answering exercise. Please assign confidence scores for 'Yes', 'No', and 'I don't know' to the submission such that all confidence scores are between zero and one, and the confidence scores sum to one.

[Submission]: <BoolQ prompt goes here>

[Criterion]: confidence:

"0": "Zero chance. It is impossible that the submission falls into the current category."

"0.5": "fifty-fifty. It is just as likely as not that the submission falls into the current category."

"1": "Certain – The submission definitely belongs in the current category."

[END DATA]

Assign confidence scores of 'Yes', 'No', and 'I don't know' to the submission. First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the choice only from between 0 and 1 (without quotes or punctuation) on its own line corresponding to the correct answer for each 'Yes', 'No', and 'I don't know' category.

Figure 6: Prompt adapted from [Zhou et al. \(2023\)](#) used to assess whether GPT4 and GPT4o models can reliably produce the probability outputs.

B Example Response from Template presented in Appendix A

Step by Step Reasoning:

1. **Contextual Understanding**: The question is asking whether Abraham Lincoln wrote the letter featured in the movie "Saving Private Ryan". Given the historical context, Abraham Lincoln was the 16th President of the United States and served from March 1861 until his assassination in April 1865. "Saving Private Ryan" is a movie set during World War II, specifically focusing on the D-Day landings and subsequent events.
2. **Historical Accuracy**: Given that Abraham Lincoln lived during the 19th century and "Saving Private Ryan" is set in the 20th century, it is historically impossible for Lincoln to have written any letter featured in the movie.
3. **Interpreting the Criterion**: Based on the given criterion for confidence scoring:
 - A '0' represents an impossibility,
 - A '0.5' indicates an event is just as probable as not,
 - A '1' represents certainty.
4. **Application of Criterion**: Knowing that Abraham Lincoln could not have written a letter in a 20th-century setting, the assertion that he did so falls into the realm of impossibility according to our criterion.
5. **Assigning Scores**:
 - The confidence score for 'Yes' (did Abraham Lincoln write the letter in Saving Private Ryan) is '0', based on historical impossibility.
 - The confidence score for 'No' must then be '1', as it is certain that he did not write the letter, given the historical context.
 - The confidence score for 'I don't know' would also be '0', as we can definitively state the impossibility based on historical facts.

Confidence Scores:

Yes:

0

No:

1

I don't know:

0

The confidence score for 'No' being the correct answer:

1

Figure 7: Example response for the question "Did Abraham Lincoln write the letter in Saving Private Ryan?" using our adapted Zhou et al. (2023) prompt template.

C Results from GPT4o

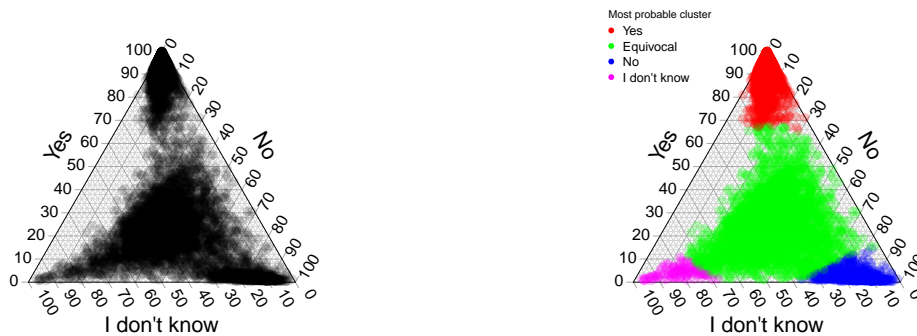


Figure 8: Ternary plots that show the probability predictions for "Yes," "No," and "I don't know" categories for the 9,427 BoolQ questions using GPT4o shown in Figure 1. Points are semi-transparent to assist with visualizing concentration. The left panel shows the probability predictions and the right panel color codes those same predictions by the cluster obtained using the methods described in Section 4.

| Description (color) | Accuracy rate (95% CI) | Cluster size $\hat{\pi}_k$ | Parameter estimates $\hat{\theta}_k$ |
|--------------------------------|------------------------|----------------------------|--------------------------------------|
| Probably Yes (Red) | 91.9% (91.0% - 92.7%) | 0.42 | (26.11, 1.13, 1.25) |
| Probably No (Blue) | 86.4% (84.0% - 88.5%) | 0.10 | (1.28, 36.53, 6.65) |
| Equivocal predictions (Green) | 56.2% (54.2% - 58.1%) | 0.47 | (3.45, 5.72, 5.10) |
| Probably I don't know (Purple) | | 0.01 | (2.17, 6.80, 39.21) |
| No clustering (Black) | 78.9% (77.9% - 79.8%) | 1.00 | - |

Table 4: Accuracy rates, confidence intervals, and estimates for cluster size and shape parameters when GPT4o is used. Results are presented overall and for the clustering approach. Accuracy rate is based on the most probable answer to each question. Color corresponds to the clusters visualized in the right panel of Figure 8.

| Description (color) | One label | All labels | Yes | No | I don't know |
|--------------------------------|-----------|------------|--------|--------|--------------|
| Probably Yes (Red) | 97.0% | 0.0% | 100.0% | 1.5% | 1.5% |
| Probably No (Blue) | 50.5% | 0.0% | 0.0% | 100.0% | 49.5% |
| Equivocal predictions (Green) | 0.2% | 75.7% | 80.5% | 99.8% | 95.1% |
| Probably I don't know (Purple) | 47.9% | 0.0% | 0.0% | 52.1% | 100.0% |
| No clustering (Black) | 45.9% | 35.9% | 79.5% | 58.5% | 51.9% |

Table 5: Results of the conformal prediction exercise on the 7,427 available answers for GPT4o. Percentages indicate how many questions included a single answer label, all three answer labels, and individual inclusion of "Yes", "No", and "I don't know" labels. Results are presented overall and by cluster.

| Description (color) | Number of prompts | Accuracy rate (95% CI) |
|--------------------------------|-------------------|-------------------------|
| Probably Yes (Red) | 13 | 84.6% (54.6% - 98.1%) |
| Probably No (Blue) | 2 | 100.0% (15.8% - 100.0%) |
| Equivocal predictions (Green) | 7 | 57.1% (18.4% - 90.1%) |
| Probably I don't know (Purple) | 0 | - |
| No clustering (Black) | 22 | 77.3% (54.6% - 92.2%) |

Table 6: Accuracy rates for the U.S. government websites using the GPT4 Turbo fitted model. Note that observations with "I don't know" as the most probable answer are not included in this analysis.

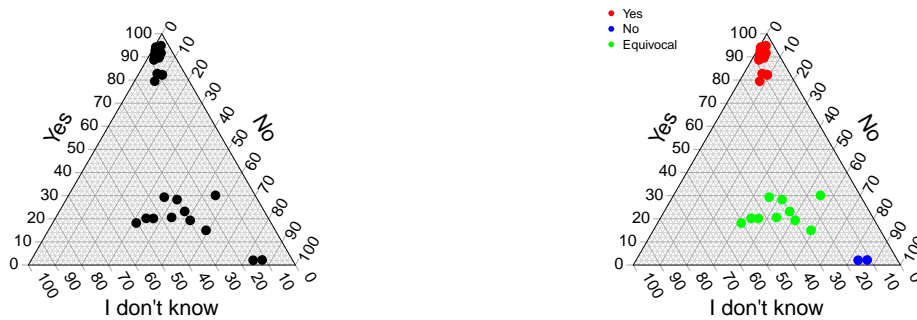


Figure 9: Ternary plots that show the probability predictions for “Yes,” “No,” and “I don’t know” categories for the 25 U.S. government website Q&A prompts using the cluster rule learned from BoolQ analysis and GPT4o. The left panel shows the probability predictions and the right panel color codes those same predictions by cluster assignment. No observations were observed in the “Probably I don’t know” cluster in the bottom left of the ternary plots.

For the customs and immigration and import and export control example using the GPT4o model, the most probable answer was “Yes” 13 times, “No” 9 times, and “I don’t know” 3 times. Among the 22 “Yes” and “No” predictions, the accuracy rate is $17/22=77\%$. Using prediction based on the clustering approach, 15 out of 25 predictions are in either the “Probably Yes” or “Probably No” clusters, and 10 observations are in the “Equivocal” cluster. No observations appeared in the “Probably I don’t know” cluster. These predictions can be seen in the right panel of Figure 9. Among the 15 non-equivocal predictions, the accuracy rate is $13/15=87\%$.

D Dirichlet Mixture Model Clustering via EM algorithm Details

The EM algorithm is a popular choice for clustering tasks in the context of finite mixture models shown in Equation (1). The unique aspect of our implementation is that we used Dirichlet cluster densities to enforce the sum-to-one constraint on the Q&A probabilities. Our approach to the EM algorithm follows the usual two step iterative process. First we take the expectation (i.e., the ‘‘E step’’), which replaces the unknown cluster membership labels with their expected value using current parameter estimates. Then we maximize (i.e., the ‘‘M step’’) the likelihood function to obtain estimates for the θ_k shape parameters for $k = 1, \dots, K$. The E and M steps are repeated until the likelihood value converges¹².

While many existing software implementations of the EM algorithm exist (Benaglia et al., 2009; Wu, 2023), we did not find any that implemented the Dirichlet distribution as a component density. For this reason, we implemented an EM algorithm that uses the Dirichlet distribution for component densities $f_k(\cdot)$.¹³

The functional form of the component densities is:

$$f_k(\mathbf{y}) = \frac{\Gamma(\sum_{l=1}^L \theta_{kl})}{\prod_{l=1}^L \Gamma(\theta_{kl})} \prod_{l=1}^L y_l^{\theta_{kl}}, \quad (2)$$

where $\Gamma(\cdot)$ is the gamma function, $l = 1, \dots, L$ indexes the possible answers ($L = 3$ corresponding to ‘‘Yes’’, ‘‘No’’, and ‘‘I don’t know’’). Thus, the complete log likelihood function is:

$$\log L_c(\Psi) = \sum_{k=1}^K \sum_{i=1}^n z_{ki} \{\log \pi_k + \log f_k(\mathbf{y}_i; \theta_k)\}. \quad (3)$$

where Ψ is a vector that contains all unknown parameters in the model, $i = 1, \dots, n$ indexes the number of observations in the analysis, $k = 1, \dots, K$ is the number of clusters in the model ($K = 4$ in our analysis), $z_{ki} = 1$ if observation i belongs in cluster k and $z_{ki} = 0$ otherwise, π_k is the weight for the k th component, \mathbf{y}_i is a length three vector of probability predictions corresponding to the i th question, and θ_k is a length three vector of shape parameters for the k th component density. Equation (3) is referred to as a complete log likelihood function because it presumes knowledge of the cluster memberships z_{ki} .

¹²See McLachlan and Peel (2004) for an overview on finite mixture models and details on the EM algorithm.

¹³Code will be made available upon publication.