A Appendix

A.1 Hyperparameters and Model configurations

Model hyper-parameters:

Our model used three GCN layers; typically, the embedding dimension was $256, 512, 1024$ for three GCN layers. For the training process, we used the binary cross-entropy loss with a decaying learning rate that reduced the learning rate by 0.9 if the validation loss did not improve 10 epochs (with an initial learning rate of $10^{-}4$ and a minimum learning rate of 10^{-7}). The optimizer was Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The batch size was 32. For the rephrased prompts, we set $k = 3$, $n = 30$, so for each rephrased question, we sampled ten answers. While calculating the ECE, we divide the confidence into $B = 10$ bins.

Evaluation Setup:

For each question, we evaluate the confidence prediction corresponding to the most likely answers from the LLM response. The setup is consistent with the baseline methods.

Graph construction:

For each question, we prompt the LLM to give 30 answers, and the temperature for LLM is set to be 0.6. For each answer, the SentenceBert model Reimers & Gurevych (2019) is used to get each answer's embedding. The cosine similarity between each answer's embedding is taken as the edge weight of the graph. We apply the K-Means clustering method to cluster similar semantic responses. The maximum cluster number is set as 3.

A.2 Computational cost

We performed all experiments on NVIDIA A100 GPUs with 80GB of memory. Generating 30 responses using the Llama3 and Vicuna models for 6000 questions from CoQA and TriviaQA data required up to 4 hours, with an average of approximately 2 seconds per question. The CoQA dataset demanded more processing time due to the longer contextual information in the input. The time can be shortened by parallel sampling.

A.3 Additional Cases

To better understand our method intuitively, we have collected a few examples to show the difference between our algorithm and APRICOT.

To summarize our observation here:

1. Multiple responses to the same question does reveal the LLM's confidence in its answers. 2. The LLM's self-evaluation of confidence is often much higher than it should be – the LLM is overconfident about its responses. 3. The chain-of-thought responses used by ApriCoT add some information to make each answer more complete and reasonable in the spirit of 1, but it mainly adds the information within one response, not as much information as the multiple responses used by ours.

Example 1:

Question: Who plays Captain Jack Sparrow's father Edward Teague in the Pirates of the Caribbean films?

True answer:: Keith Richards

LLM response: David Schofield

More responses from the LLM: Martin Klebba. Keith Richards, Geoffrey Rush, Martin Klebba. Keith Richards. Martin Klebba. David Schofield. (only list 7 responses here to save space)

GCC-estimated confidence: 0.23

CoT response: David Schofield,

Self-evaluation: 80

ApriCoT-estimated confidence: 0.79

Example 2:

Question: In which film will you find the Rodger Young?

True answer:: Starship Troopers

LLM response: The Bridge on the River Kwai.

More responses from the LLM: The Greatest Story Ever Told. The Best Years of Our Lives. The Bridge on the River Kwai. The Best Years of Our Lives (1946). 1949's Battleground. The Best Years of Our Lives.

GCC-estimated confidence: 0.22

CoT response: All the President's Men.

Self-evaluation: 95

ApriCoT-estimated confidence: 0.81

Example 3:

Question: BS is the international car registration of which country?

True answer:: Bahamas.

LLM response: Germany.

More responses from the LLM: Bahamas. Bahrain. Bangladesh. Bahamas. Belgium. Bahamas. Germany. Bhutan. Belgium.

GCC-estimated confidence: 0.34

CoT response: Belgium

Self-evaluation: 98

ApriCoT-estimated confidence: 0.61

A.4 Additional Visualizations

Besides the cases we show in the previous section. Here, we present several case examples and visualize the response patterns. We performed dimension reduction of LLM's responses to different questions and then plotted their embeddings to the 2-dimensional space. Fig 4 shows the responses generated by Llama3 as an example. From the figure, we observe that answers with higher confidence levels tend to cluster closely together, indicating consistency and reliability in these responses. In contrast, answers with lower confidence levels exhibit greater diversity, reflecting a broader range of possibilities. This behavior aligns well with our initial assumption, demonstrating that higher confidence responses are more consistent, while lower confidence responses capture a wider variety of potential answers.

A.5 Additional results

Additional reliability plots We showed all reliability diagrams for Llama3 for TriviaQA in Fig. 5 and CoQA dataset in Fig. 6. To summarize the trends, we observe that Platt scaling narrows the range to the middle value. Verbalized uncertainty cannot generate a wider range of confidence values. GraphSpecral with Platt tends to generate a wider range of confidence values, but the bias can not be improved across all cases, resulting in the bar height not following the diagonal line closely. Our model can predict a wider range of confidence values and achieve better calibration in all settings, with the auxiliary consistency graph and clustering features contributing to improved calibration overall.

Additional baseline results In Table 5, we showed the performance of the baseline method under varying training sizes. As the number of training data decreases, the ece will drop from 0.096 to 0.165.

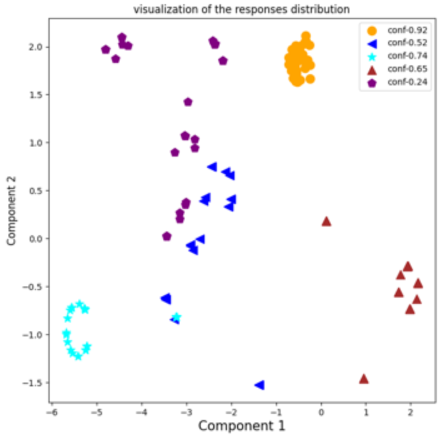
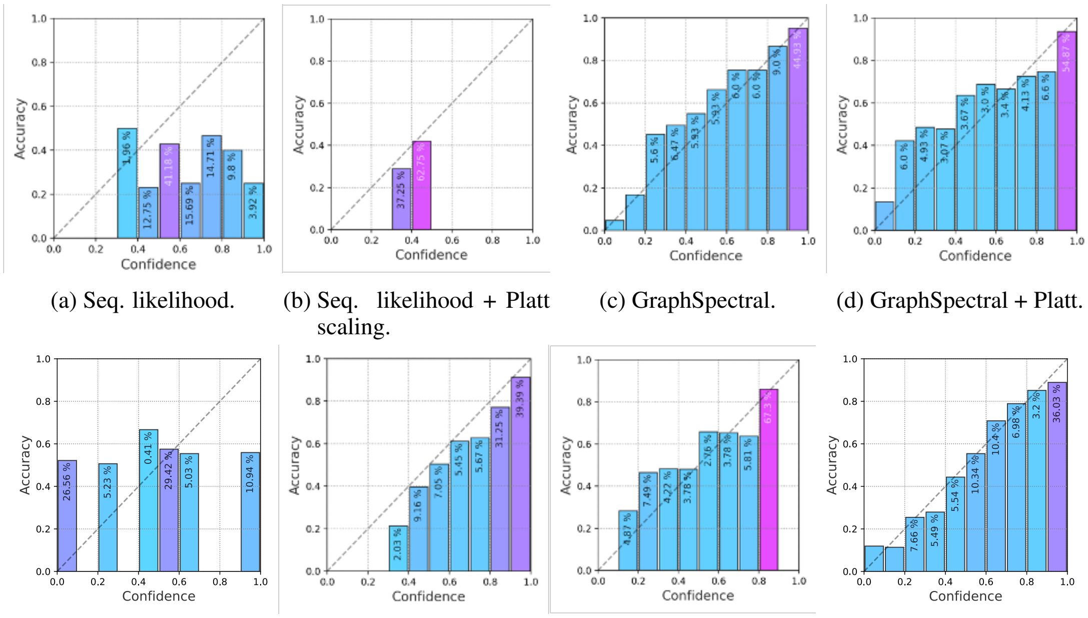


Figure 4: Visualization of the generated response patterns



(a) Seq. likelihood.

(b) Seq. likelihood + Platt scaling.

(c) GraphSpectral.

(d) GraphSpectral + Platt.

(e) Verbalized Qaul

(f) Apricot.

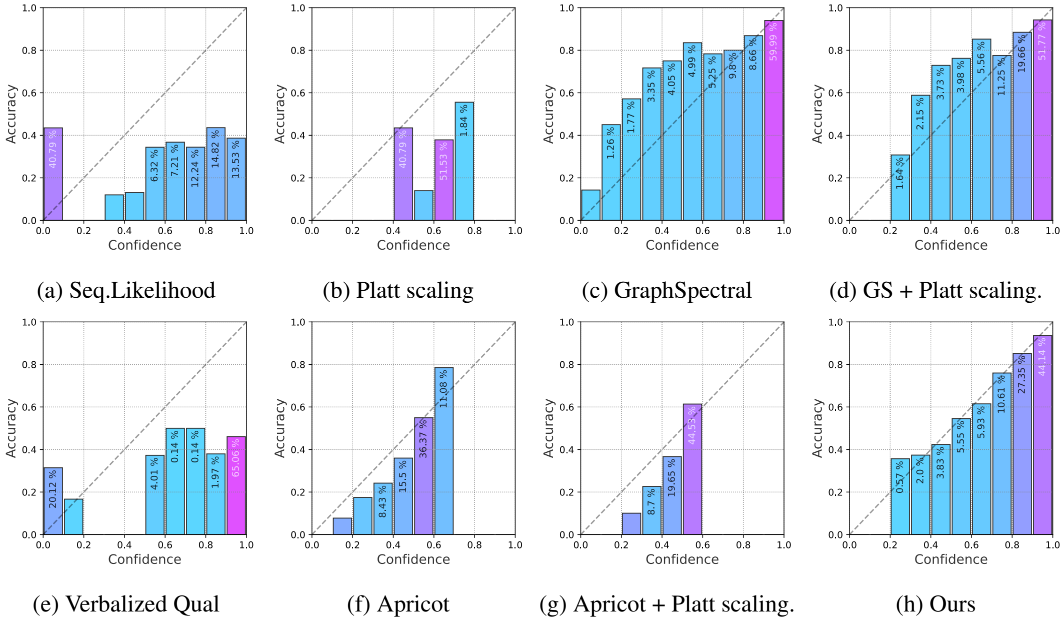
(g) Apricot + Platt scaling.

(h) Ours

Figure 5: Reliability diagrams for different methods using 10 bins each for TriviaQA from Llama3 model responses. The color and the percentage number within each bar indicate the ratio of responses contained in each bin. Larger values are represented by colors closer to purple.

A.6 Prompting strategy

In Fig. 7, we showed the prompts we used to generate the rephrasing questions.



(a) Seq.Likelihood

(b) Platt scaling

(c) GraphSpectral

(d) GS + Platt scaling.

(e) Verbalized Qual

(f) Apricot

(g) Apricot + Platt scaling.

(h) Ours

Figure 6: Reliability diagrams for different methods using 10 bins each for CoQA from Llama3 model responses. The color and the percentage number within each bar indicate the ratio of responses contained in each bin. Larger values are represented by colors closer to purple.

Table 5: Performance under varying Training Sample Sizes for the baseline methods(Apricot)

# of Training Samples	ECE	AUROC	Brier
100	0.165	0.611	0.229
300	0.133	0.634	0.211
500	0.112	0.695	0.204
1000	0.105	0.722	0.192
4000	0.096	0.743	0.187

Prompt for generating rephrasing questions:

You are a helpful assistant. I have a question that I would like to see rephrased in multiple ways. Please take the original question and generate several rephrased versions while maintaining the same meaning and the question can only have one direct answer. Here is the original question: {…}. Please provide four distinct rephrases of the question.

(a) Prompt strategy for rephrasing question

Figure 7: Prompt for generating rephrased prompts