An Inside Look Into How LLMs Fail to Express Complete Certainty: Are LLMs Purposely Lying?

Stanford CS224N Custom Project

Joshua Fajardo
Department of Computer Science
Stanford University
joshdf@stanford.edu

Abstract

Zhou et al. (2023) demonstrate that LLMs perform poorly in question-answering tasks when prompted to begin with high certainty expressions, as opposed to low certainty expressions. Our core goal for this paper is to use SAPLMA, as introduced by Azaria and Mitchell (2023), to show that LLMs misunderstand what it means to have 100% certainty, conflating its meaning with that of 0% certainty. We note that our results weakly reject our original hypothesis—in other words, they show that there may exist a reasonable internal intuition within Mistral 7B regarding the true meaning of 100% certainty. Specifically, we find that when feeding sentences prefixed markers of 0%, 70%, 90%, and 100% certainty into our LLM and performing true-false classification using the corresponding LLM activations, our SAPLMA classifiers display similar performances between the final 3 prefixes. This finding lightly suggests that LLMs may know that they're lying when incorrectly answering questions after being prompted to respond with complete certainty. Furthermore, we make significant discoveries about SAPLMA: we demonstrate that 1) SAPLMA continues to perform well using Mistral 7B, 2) SAPLMA may be sensitive to the particular sentence structure used during testing, and 3) the necessary training duration for SAPLMA classifiers may correlate to the average training sentence length.

1 Key Information to include

- Mentor: Caleb Ziems

2 Introduction

Zhou et al. (2023) give us a closer look into how models perform in question-answering tasks when prompted to begin their sentences with epistemic markers of varying degrees of certainty (e.g. "I'm certain it's", "It could be") (p. 17). They find that at extreme levels of certainty, models perform extremely poorly (Zhou et al., 2023). We find this counter-intuitive result to be very interesting. To our best knowledge, no other research has tried to understand this problem of why models perform so poorly under complete certainty. Our motivation with this research is to help fill this gap in understanding. We aim to study *why* models lie, in order to help limit them from doing so.

When focusing on epistemic certainty markers with percentages (e.g. "I'm 70% sure it's"), Zhou et al. (2023) find that LLMs consistently have the worst answer accuracy when values of 0% and 100% were used. They observe from a widely used pretraining dataset, The Pile, that there is very frequent co-occurrence between usage of "100%" and uncertainty (Zhou et al., 2023). Furthermore, they hypothesize: "both the use of negation with "100%" and the general lack of use of "100%" with expressions of certainty contribute to the lowered performance of these prompts" (Zhou et al., 2023, p. 7). We find this hypothesis to be very convincing, and speculate that there may even be extralinguistic

compounding factors, such as the Dunning-Kruger effect, where folks with lower levels of familiarity with a topic tend to be overconfident in their knowledge.

We generalize the hypothesis made by Zhou et al. (2023), arguing that the model misunderstands what it means to have 100% certainty. Essentially, we hypothesize that the model roughly internally equates 100% certainty to mean 0% certainty.

To test our hypothesis, we leverage "Statement Accuracy Prediction, based on Language Model Activations (SAPLMA)" (p. 5), as introduced by Azaria and Mitchell (2023). As the name suggests, SAPLMA features a binary classifier which is fed as input the activations of one of the layers of the LLM (Azaria and Mitchell, 2023) as it processes the last token of a statement. Additionally, they provide a balanced true/false dataset which is organized by topic (Azaria and Mitchell, 2023).

Our approach features three experiments. First, we perform a preliminary experiment that successfully demonstrates that SAPLMA continues to surpass the baselines set by Azaria and Mitchell (2023) when using Mistral 7B. We show that our models consistently achieve their best accuracies when fed as input the activations from layers closer to the center: per-topic, we only find peak performances in layers 16 and 20, as opposed to 24, 28, and 32.

In our second experiment, we create and test on an augmented dataset, where the statements in the dataset are prefixed with various numerical epistemic markers of certainty: "I am X% certain that", using 0, 70, 90, and 100 for values of "X". Our expectations were that for the 0% and 100% markers, our classifiers would have accuracies at or below 40%, while for the 70% and 90% markers, our classifiers would reach or exceed 60% accuracy. Ultimately, all of our reported average accuracies for this experiment perform slightly better than 50%, which meets neither of our expectations. While Azaria and Mitchell (2023) demonstrate that SAPLMA performs well on out-of-distribution topics, we find that they do not perform well on out-of-distribution sentence structures.

For our final experiment, we only change our training method. Our intention with this experiment is to create a closer alignment between the distributions of the training and test data by extending our list of prefixes in our augmented dataset beyond numerical prefixes. We use these statements with non-numerical prefixes for training only. We now change the lens through which we interpret our hypothesis: instead of focusing on the certainty of the prefix, we focus on whether the prefix affirms or rejects the statement that follows. During training, we flip the labels for statements with denying prefixes. During testing, we expected to create a graph with similar qualitative properties to Figure 4 by Zhou et al. (2023), where the accuracies for 0% and 100% are significantly lower than for intermediate percentages. Our findings suggest that our hypothesis may be incorrect: that Mistral 7B does not conflate the meanings of 0% certainty and 100% certainty. This implies that in the experiments by Zhou et al. (2023), that their LLMs may internally understand that they're lying when incorrectly generating answers beginning with claims of high levels of certainty. Additionally, we find that it takes an order of magnitude longer in order to train our SAPLMA classifiers on these augmented statements (which are, on average, roughly double the size of the original statements).

3 Related Work

The experiments and results by Zhou et al. (2023) build a strong foundation for our work, showing a stark difference in performance on causal language modeling for question-answering tasks when prompting the LLM to begin with prefixes of varying levels of certainty. Our work varies from theirs in three different ways. First, we focus on LLM behavior when they *interpret* text, as opposed to when they generate it. This is closely tied to the second core difference, that our chosen task is true/false classification, instead of question-answering. Finally, while Zhou et al. (2023) employ a black box approach to studying LLMs, we directly access (but do not modify) the hidden states of our LLM.

Our approach is heavily inspired by that of Azaria and Mitchell (2023). We utilize their baselines, their proposed method SAPLMA, several of their experiment details, and their provided dataset. We aim to explore past the breadth of their work by testing on an augmented version of their dataset, which allows us to study how various epistemic markers affect the beliefs of LLMs.

4 Approach

4.1 Building on Existing Work

We use SAPLMA to build various classifiers for statement truthfulness that take in the activations of a Large Language Model (LLM) after having been fed a statement (Azaria and Mitchell, 2023).

Taking inspiration from Zhou et al. (2023), we use one format of numerical epistemic marker prefixes: "I am X% certain that". For the values of "X", we use 0, 70, 90, and 100. 70 and 90 were chosen because Zhou et al. (2023) report them to achieve the best accuracies in question-answering tasks.

Unlike in the previously mentioned experiments, we utilize Mistral 7B, as it is shown to have many performance benefits over other recent LLMs (Jiang et al., 2023).

4.2 Preliminary Experiments

Experiment 1. Prior to testing our hypothesis, we aim to strengthen the validity of our results by demonstrating that we can successfully exceed the baselines set by Azaria and Mitchell (2023), as shown in Table 1 of their paper, when using Mistral 7B.

At the time of writing, the official implementation[1] for the work by Zhou et al. (2023) is in progress. Fortunately, they show that their findings hold for various GPT3 and GPT4 models (Zhou et al., 2023). We expect Mistral 7B to have similar performance, and save this extra verification for future work.

4.3 Main Experiments

Experiment 2. Our second experiment features an augmented dataset used at test time, where each statement is prefixed with various numerical epistemic markers. More information regarding these markers is provided in the Experiments section below, as well as in Appendix A.1. We designate one topic within the augmented dataset at a time as the test topic. These are our own modifications. Similar to Azaria and Mitchell (2023), we use the original dataset for training, but leave out the test topic. We average our results over all topics and provide an accuracy for each pair of chosen prefixes and LLM hidden layers.

Experiment 3. For our final experiment, we differ our training method but keep the rest the same as in our second experiment. Our goal here is to create a better alignment between our training data and our test data in order to encourage our classifiers to account for prefixes. We extend our augmented dataset by adding 24 non-numerical prefixes, where each prefix either affirms or rejects the statement that follows. For these prefixes, we vary several linguistic categories (e.g. evidentiality, level of certainty, perspective, etc.), but keep an equal number of affirming and rejecting prefixes. These prefixes were generated with the help of Table 6 from Zhou et al. (2023) and the help of Gemini[2] by Google. We divide our augmented dataset, using only the numerical prefixes for testing, and the remaining for training. Our process for training and testing is similar to what we did previously: using *only* the augmented dataset, we iteratively select a topic as the test topic, and train on the remaining topics. Since we have a relatively small number of prefixes, we incorporate dropout in order to prevent our models from overfitting to the training prefixes. This is our own modification to the SAPLMA classifier.

4.4 Baselines

Since both of our datasets are balanced with true and false statements for each topic, we can set a baseline at 50% accuracy for random guessing on both datasets.

For demonstrating the effectiveness of SAPLMA with Mistral 7B, we utilize the same baselines as found in Table 1 of the work by Azaria and Mitchell (2023). This includes accuracy scores for the original dataset computed using BERT, 3-shot learning, 5-shot learning, and direct LLM prompting.

[1] https://github.com/katezhou/navigating_the_grey

[2] https://gemini.google.com

4.5 Implementation

While there exists a community-provided implementation of SAPLMA on GitHub[3], we find that we are able to much more easily accommodate the needs of our experiment by writing our own implementation. Though this resource has been helpful in understanding how SAPLMA works behind-the-scenes, we claim our implementation to be our own original work. All code for our research can be found at: `https://github.com/joshuafajardo/cs224n-proj`.

5 Experiments

5.1 Data

Azaria and Mitchell (2023) provide a true-false dataset of consisting of 6,084 statements across 6 different topics. We designate two different true-false datasets, used for true-false classification:

1. **Original Dataset:** This is the dataset provided by Azaria and Mitchell (2023), used as-is.

2. **Augmented Dataset:** To produce this dataset, each statement within the original dataset is prefixed with each of 28 epistemic markers described in our Approach. These 28 prefixes include 4 numerical prefixes used for testing and 24 non-numerical prefixes (12 affirming and 12 rejecting) used for training. The full list of prefixes can be found in Appendix A.1 For each augmented statement, we also include the augmented label. These augmented labels are only used for training. For statements with affirming prefixes, their augmented labels are equal to their original labels. For rejecting prefixes, we flip the original labels to get the augmented labels. During testing, we do not use these augmented labels, and instead expect accuracies worse than 50% for rejecting prefixes.

5.2 Evaluation Method

For the first experiment, we simply require that all of our accuracies surpass the baselines set by Azaria and Mitchell (2023).

For experiments 2 and 3, using SAPLMA, we must achieve at least 60% accuracy, averaged across all topics within the augmented dataset, for statements with the 70% and 90% prefixes. Conversely, we must achieve no greater than 40% accuracy, on average, on the augmented dataset for statements with the 0% and 100% prefixes.

5.3 Experiment Details

Our experiment details are modeled after Azaria and Mitchell (2023).

For all experiments, we use an A100 on Google Colab. It takes approximately 90 minutes to load Mistral 7B and compute the activations for both datasets.

Experiments 1 and 2. For each topic, we train 5 different models, which each take in the activations from a different layer of Mistral 7B—our chosen layers are 16, 20, 24, 28, and 32. We train each model for 5 epochs with the binary cross-entropy loss and a batch size of 32. We use the Adam optimizer with an initial learning rate of 0.001 for all models. For experiments 1 and 2 combined, it takes approximately 2 minutes to train and evaluate all 60 truth classifiers.

Experiment 3. Similar to experiment 2, we use the hidden states from layers 16, 20, 24, 28, and 32 of Mistral 7B. Since we've added new prefixes, our training data is now 24 times larger. Nonetheless, we show our training results after 5 epochs and 20 epochs, as we observe that our training accuracy does not significantly improve (beyond 50%) after only one epoch. We now train with a batch size of 4096 for better efficiency. For all fully connected layers except for the last, we use a dropout rate of 0.2. It takes approximately 26 minutes in order to train and evaluate all 60 truth classifiers.

Table 1: SAPLMA test accuracies on the original dataset with Mistral 7B. Baselines were generated by Azaria and Mitchell (2023).

Layer	Cities	Elements	Companies	Animals	Facts	Inventions	Average
16	0.5501	**0.6591**	**0.7758**	**0.7480**	**0.7210**	0.6073	**0.6695**
20	**0.5576**	0.6247	0.6908	0.7331	0.6215	**0.6096**	0.6371
24	0.5432	0.5710	0.6233	0.6974	0.6852	0.5719	0.6072
28	0.5267	0.5624	0.6692	0.6032	0.7015	0.5913	0.5998
32	0.5542	0.5796	0.5575	0.6310	0.6460	0.5890	0.5857
BERT	0.5357	0.5537	0.5645	0.5228	0.5533	0.5302	0.5434
3-shot	0.5410	0.4799	0.5685	0.5650	0.5538	0.5164	0.5374
5-shot	0.5416	0.4799	0.5676	0.5643	0.5540	0.5148	0.5370
It-is-true	0.523	0.5068	0.5688	0.4851	0.6883	0.584	0.5593

Table 2: Test accuracies for SAPLMA models trained on the original data and evaluated on statements augmented with numeric epistemic markers. Accuracies are averaged across all topics.

Layer	0%	70%	90%	100%
16	**0.5374**	**0.5413**	**0.5334**	**0.5484**
20	0.5343	0.5257	0.5318	0.5369
24	0.5172	0.5173	0.5210	0.5241
28	0.5172	0.5129	0.5180	0.5152
32	0.5172	0.5272	0.5242	0.5168

5.4 Results

Experiment 1. Table 1 shows that SAPLMA with Mistral 7B, averaged across all topics, outperforms all baselines set by Azaria and Mitchell (2023). Our gathered test accuracies reach our expectations; given that SAPLMA has been shown to succeed for LLAMA2-7b and OPT-6.7b, we naturally expect that it should also work other LLMs like Mistral 7B. We note that our models very consistently achieve their best accuracies in layers 16 and 20. This behavior is similar to what Azaria and Mitchell (2023) report: for OPT-6.7b, they report their best scores in layers 20 and 24, and for LLAMA2-7b, they report best scores in only layer 16.

Experiment 2. Table 2 shows that all trained models, on average, perform very slightly better than 50%, and that none of our accuracies exceed 60%. Additionally, we find very little difference in performance between the various numerical epistemic markers, meaning our current findings do not support our hypothesis. We believe that this result may be because during training, our models have overfit to the structure of sentences within the original dataset. This was unexpected, as we expected SAPLMA to find an internal representation of truth that would generalize well to other sentence structures.

Experiment 3. Table 3 displays our test results for experiment 3. After 5 epochs, we find that all of our accuracies are within only 3% of our baseline of 50%. While all of our accuracies for the "0%" prefix are below 50%, they only deviate from 50% by up to 0.0047%, which is not significant. We do find that the other prefixes are able to achieve slightly higher scores, correctly guessing the label for around 2.5% more test cases.

After 20 epochs, we get some very interesting results. We initially expected there to be very little change in the test accuracies between 5 epochs and 20 epochs. To our surprise, we see that the differences in performances across prefixes is much more pronounced after 20 epochs: the accuracies for the "0%" prefix all stay within roughly 1% of our baseline of 50%, while the highest accuracies for the remaining prefixes now hover around 55%. While this does not reach our anticipated threshold of 60%, we find these results to lightly suggest that our hypothesis may be false.

[3]https://github.com/balevinstein/Probes

Table 3: Test accuracies for SAPLMA models trained and tested on different parts of the augmented dataset. For training, statements without numerical prefixes were used, and the labels were flipped for denying prefixes. For testing, statements with numerical prefixes were used. No labels were flipped during testing.

Layer	0%	70%	90%	100%	Num Epochs
16	0.4983	0.5265	0.5280	0.5241	5
20	0.4953	0.5177	0.5193	0.5180	5
24	0.4978	0.5150	0.5136	0.5196	5
28	0.4994	0.5086	0.5073	0.5103	5
32	0.4984	0.5048	0.5055	0.5022	5
16	0.5063	0.5540	0.5518	0.5486	20
20	0.4917	0.5224	0.5205	0.5168	20
24	0.4991	0.5173	0.5118	0.5122	20
28	0.5104	0.5152	0.5205	0.5159	20
32	0.4960	0.5121	0.5147	0.5101	20

6 Analysis

6.1 Experiment 1

Isolating the best test accuracies for each topic in Table 1, we find that SAPLMA achieves the best performance for the "companies" and "animals" topics, and that it achieves the worse performance for the "cities" topic. In some ways, this is different from what is observed by Azaria and Mitchell (2023): they found highest accuracies in the "cities" and "companies" topics, and found performance on the "animals" topic to be on the lower end. We believe that this may reflect some differences in the contents of the datasets used to train the LLMs used by them and by us.

6.2 Experiment 2

We note that, across all prefixes and layers, our models consistently achieve achieve (albeit mildly) above-average accuracies. This can be easily explained by the fact that the entirety of each original statement can be found in our augmented statements. For example, an augmented version of "Burma is a name of a country" is "I'm 70% certain that Burma is a name of a country."

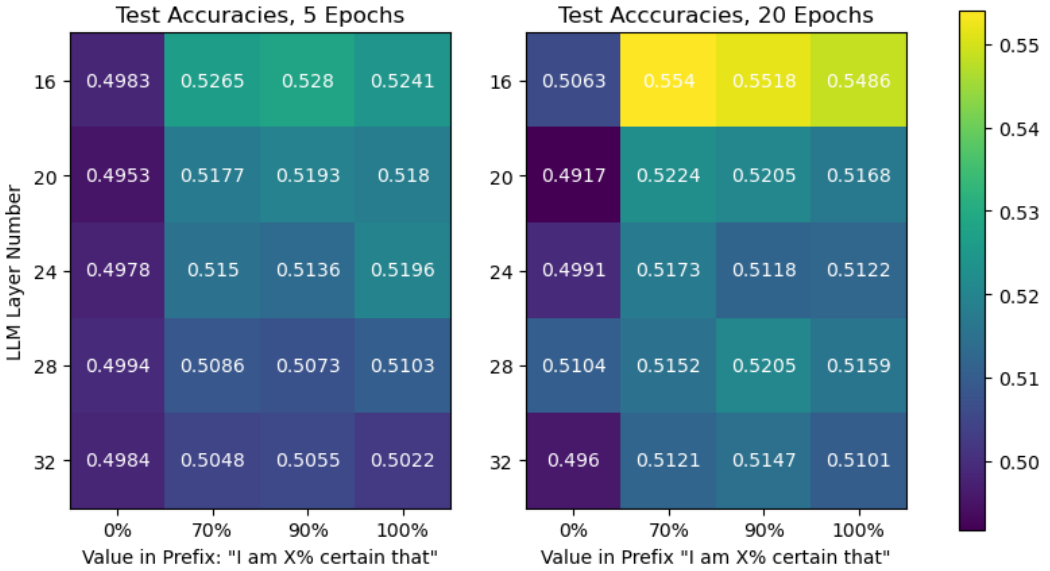
6.3 Experiment 3

Contrary to our expectations, we observe in Figure 1 that we do not achieve similar accuracies between the 0% and 100% prefixes, and instead find a grouping between the accuracies for the 70%, 90%, and 100% prefixes. We find there to be two likely causes for this. The first reason could be that Mistral 7B understands the true meaning of 100% certainty. The second reason could be that in our training dataset, since our augmented training data includes the correct augmented labels for statements prefixed with high certainty markers, we may be inherently helping our classifiers distinguish between uncertainty, high certainty, and complete certainty.

Training Time. While the training section of our augmented dataset is 24 times larger than the original dataset, there are a large amount of redundancies: each statement is replicated 24 times, and each prefix is prepended to several thousand statements. Therefore, we initially anticipated the models in our final experiment to train fairly quickly, in comparison to the models trained on the original dataset. To our surprise, it required a considerable number of epochs in order for our models to get close to local minima. In Table 4, we display our training accuracies after 5 epochs and after 20 epochs.

Test Accuracies After 5 and 20 Epochs. We show the average test accuracies after both 5 and 20 epochs in Table 3, and observe observe a considerable change in the test accuracies. This completely defied our expectations: we initially expected that between 5 and 20 epochs, our models would be overfitting to the training prefixes. Notably, we see that the characteristics of our results are much more pronounced after 20 epochs, as illustrated in Figure 1. One potential cause is that because the

Figure 1: Heat maps for the test results for experiment 3 after 5 epochs and 20 epochs, accordingly.



sentences from which our activations are gathered are now much longer, the information stored in these activations may be much more dense, thus making it harder for our classifier to find the LLM's representation of truth. For example, after augmentation, the original sentence, "Thimphu is a name of a city" can become "If I had to guess, I would deny that Thimphu is a name of a city"; in this case, the length of the sentence becomes more than twice as long. Thus, we claim that the average training sentence length may have an impact on the number of epochs required for SAPLMA classifiers to find the representation of truth.

Better Performance in Earlier Layers. Finally, we also observe that our models achieve more greater accuracies when trained on the activations of the earlier layers, and considerably worsen on the final layer. This may be explained by the intuition that, as mentioned by Azaria and Mitchell (2023), the final layers are mainly concerned with token generation.

Table 4: SAPLA training accuracies for experiment 3 after 5 epochs and 20 epochs. Models were trained using the augmented dataset. Topics in heading are the topics that were held out during training.

Layer	Facts	Animals	Cities	Companies	Elements	Inventions	Num Epochs
16	0.6157	0.6211	0.5880	0.6236	0.6075	0.5961	5
20	0.5821	0.5963	0.6039	0.6041	0.5889	0.5837	5
24	0.6081	0.5931	0.5769	0.5982	0.5845	0.5795	5
28	0.5848	0.6284	0.5708	0.6005	0.5802	0.5763	5
32	0.5092	0.5128	0.5151	0.5059	0.5135	0.5160	5
16	0.6889	0.6977	0.6365	0.7339	0.7664	0.6748	20
20	0.6818	0.6990	0.6367	0.7051	0.6825	0.6641	20
24	0.6847	0.6960	0.6419	0.7024	0.6789	0.6610	20
28	0.6811	0.6965	0.6380	0.7006	0.6796	0.6607	20
32	0.6389	0.6931	0.6241	0.6946	0.6685	0.6544	20

7 Limitations and Future Work

Similar to Azaria and Mitchell (2023), we perform minimal hyperparameter tuning for our experiments, keeping all hyperparameters the same for each experiment (except layer number). It may be possible that certain configurations would work better for certain topics and/or LLM layers. One

avenue for future work could be to find the maximum average test accuracies that can be achieved using SAPLMA on the original dataset.

We again note that we have not verified that the results by Zhou et al. (2023) can be replicated for Mistral 7B. We defer this verification for future work.

A natural extension to our findings regarding the dip in performance by SAPLMA on out-of-distribution sentence structures would be to train SAPLMA classifiers using activations from a wider variety of sentence structures; it would be interesting to see if SAPLMA would continue to succeed on in-distribution sentence structures and out-of-distribution sentence structures. This would give a stronger indication that SAPLMA can discover an LLM's understanding of truth.

One of the main limitations from our third experiment was 1) the limited number of prefixes that we had gathered for training, and 2) the fact that we only explored using our markers as prefixes. Further work may look into using controllable paraphrase generation in order to create a wider variety of markers, marker locations, and sentence structures.

8 Conclusion

In this work, we take a closer look at how LLMs interpret sentences where numerical epistemic markers of certainty are used, with the goal of understanding what LLMs believe about the meanings of these epistemic markers. We find that Mistral 7B may have a reasonable internal intuition regarding the true meaning of 100% certainty, indicating that LLMs may be encouraged to lie when prompted to begin their responses with statements of complete certainty. Furthermore, we find that SAPLMA continues to perform well when using the hidden states from Mistral 7B. Additionally, we demonstrate that while SAPLMA performs well on topics that were not seen during training, it does not seem to generalize well to sentence structures that were not seen during training. Finally, we provide possible topics of future work to build on top of our findings.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Prefixes

Table 5: Prefixes used to create the augmented dataset. Prefixes were gathered by the authors, with the help of Gemini. Values in the "Affirms/Rejects" column are based on the true meanings of the prefixes, not the hypothesized LLM interpretations.

Prefix	**Affirms/Rejects**	**Train/Test**
I'm 0% certain that	Rejects	Test
I'm 70% certain that	Affirms	Test
I'm 90% certain that	Affirms	Test
I'm 100% certain that	Affirms	Test
It is true that	Affirms	Train
It is false that	Rejects	Train
It must be true that	Affirms	Train
It must be false that	Rejects	Train
It should be the case that	Affirms	Train
It should not be the case that	Rejects	Train
It could be true that	Affirms	Train
It's probably not true that	Rejects	Train
I know it's true that	Affirms	Train
I know it's false that	Rejects	Train
I wouldn't doubt that	Affirms	Train
I highly doubt that	Rejects	Train
I firmly believe that	Affirms	Train
I do not believe that	Rejects	Train
I wouldn't be surprised to find that	Affirms	Train
I would be shocked to find that	Rejects	Train
I'm convinced that	Affirms	Train
I'm not convinced that	Rejects	Train
I think it's possible that	Affirms	Train
I think it's unlikely that	Rejects	Train
Not to be certain, but I think that	Affirms	Train
Not to be certain, but I doubt that	Rejects	Train
If I had to guess, I would say that	Affirms	Train
If I had to guess, I would deny that	Rejects	Train