Does It Know?: Probing and Benchmarking Uncertainty in Language Model Latent Beliefs

Brian R.Y. Huang
MIT CSAIL
branhung@alum.mit.edu

Joe Kwon
MIT
joekwon@mit.edu

Abstract

Understanding a language model's beliefs about its truthfulness is crucial for building more trustworthy, factually accurate large language models. The recent method of Contrast-Consistent Search (CCS) measures this "latent belief" via a linear probe on intermediate activations of a language model, trained in an unsupervised manner to classify inputs as true or false. As an extension of CCS, we propose *Uncertainty-detecting CCS (UCCS)*, which encapsulates finer-grained notions of truth, such as uncertainty or ambiguity. Concretely, UCCS teaches a probe, using only unlabeled data, to classify a model's latent belief on input text as true, false, or uncertain. We find that UCCS is an effective unsupervised-trained selective classifier, using its uncertainty class to filter out low-confidence truth predictions, leading to improved accuracy across a diverse set of models and tasks. To properly evaluate UCCS predictions of truth and uncertainty, we introduce a toy dataset, named *Temporally Measured Events (TYMES)*, which comprises true or falsified facts, paired with timestamps, extracted from recent news articles from the past several years. TYMES can be combined with any language model's training cutoff date to systematically produce a subset of data beyond (literally, occurring after) the knowledge limitations of the model. TYMES serves as a valuable proof-of-concept for how we can benchmark uncertainty or time-sensitive world knowledge in language models, a setting which includes but extends beyond our UCCS evaluations.

1 Introduction

Large language models (LLMs) form the backbone of many deployed AI systems. However, a significant challenge LLMs face is their tendency to produce factually incorrect, misleading, or even dishonest outputs–a phenomenon dubbed "hallucination" by the general public. In particular, many users of LLM-based systems interact with the model through use cases such as question-answering or information search, where the factual accuracy of model outputs is crucial for usefulness and safety. Ensuring that these models produce truthful outputs can therefore lead to the creation of more trustworthy AI systems.

Recent research into LLMs have delved into their capabilities to comprehend and relay real-world knowledge, pinpointing strengths and limitations. A noteworthy contribution in this arena is the Contrast-Consistent Search (CCS) method, as introduced by [CHD+23]. CCS elucidates a model's understanding of truth by examining its latent activations, training a linear probe on intermediate model layers to discern whether the model believes an input text is true or false, all without the need for labeled data. This unsupervised approach not only provides insight into the internal knowledge of the model, but also potentially identifies discrepancies between what a model knows and what it communicates. Given the scalability promised by the unsupervised nature of CCS, our research seeks to enhance its capability; we aim to incorporate a dimension of uncertainty in its truth assessments.

ATTRIB: Workshop on Attributing Model Behavior at Scale at NeurIPS 2023.

Our contributions. The core functionality of CCS lies in uncovering the latent knowledge embedded within language models. We extend this understanding by examining not just what the model knows but also its awareness of the limits of its own knowledge. Specifically, rather than merely classifying inputs as true or false, our objective is to assess situations where a language model is appropriately uncertain. Leveraging the unsupervised methodology inherent to CCS, we introduce a modified loss function combined with data augmentation techniques. This adaptation enables the CCS probe to execute a *three-way classification*–categorizing input texts as true, false, or uncertain. We refer to this enhanced method as Uncertainty-Detecting CCS, or UCCS. To validate its performance, we benchmark our method as a *selective binary classifier* against a substantial portion of the datasets used in [CHD+23].

To fully evaluate the uncertainty detection capabilities of UCCS, we require data samples where the model ought to be uncertain about the truth value of the sample. For a model with a known training cutoff date, one simple approach is to feed the model data samples about current and recent news events, some of which transpired before the cutoff date, some after. We manually assemble a dataset of 241 global news factoids from the years 2018-2023, about half of them randomly falsified, all of them accompanied with the date of occurrence. We use this dataset, named **T**emporall**Y** **M**easured **E**vent**S** (**TYMES**), to evaluate trained UCCS and CCS probes. On all recent news factoids past a model's training cutoff date, uncertainty becomes the correct ground truth for a trained UCCS probe on that model. Beyond the selective binary classification for previous benchmarks that UCCS does, we can examine the performance of UCCS on TYMES as a proper three-way classifier.

2 Investigating Uncertainty

To create more trustworthy, reliable language models, we want a more granular understanding of latent beliefs in language models; we seek to understand not only how language models conceptualize truthfulness and falsehood, but also how they deal with uncertainty, ambiguity, or knowledge limitations in their internal computations. To make progress in this direction, we extend the Contrast-Consistent Search (CCS) method of [CHD+23] by making significant modifications to their data preprocessing and training objective. We generalize their binary classification to a three-way classification between the true, false, and uncertain categories, producing the Uncertainty-Detecting CCS (UCCS) method.

2.1 Uncertainty-Detecting CCS (UCCS)

To generalize the unsupervised training procedure of CCS for our UCCS probe, we focus on two components of the original procedure: the data augmentation converting input text into "contrast pairs," and the custom loss function, which essentially trains the probe on the law of total probability and the law of excluded middle, both logic constraints for truth values.

For some input text x, such as a factual statement, a contrast pair (x^+, x^-) is created by casting x into a claim that x is true, corresponding to x^+, and a claim that x is false, corresponding to x^-. Concretely, a statement x is rephrased into a question, and contrast samples are written as $x^+ =$ "[question]? Yes" and $x^- =$ "[question]? No." For our uncertainty-detection experiments, we add a third contrast sample; for the same input text x, we generate $x^\varnothing$ as a claim that x is uncertain or ambiguous. For example, in our question format above, we write $x^\varnothing =$ "[question]? Uncertain" or "[question]? I don't know" for the third contrast sample. We thereby form a *contrast triplet* $(x^+, x^-, x^\varnothing)$ for any text sample x.

With contrast triplets in hand, we generalize the loss functions of [CHD+23] to accommodate $x^\varnothing$. Let θ denote the parameters of our linear probe, and let $p_\theta(x)$ denote the sigmoid output of probe θ for some input x. Our new consistency loss, which enforces the law of total probability, is

$$L_{\text{consistency}}(\theta; x) := \left[p_\theta\left(x^+\right) + p_\theta\left(x^-\right) + p_\theta\left(x^\varnothing\right) - 1\right]^2 \tag{1}$$

which encourages probabilities across a contrast triplet to sum to 1. Our new confidence loss, which enforces mutual exclusivity of the True/False/uncertain outcomes, is

$$L_{\text{confidence}}(\theta; x) := \min\left\{1 - p_\theta\left(x^+\right), 1 - p_\theta\left(x^-\right), 1 - p_\theta\left(x^\varnothing\right)\right\}^2 \tag{2}$$

which pushes the probability of exactly one contrast sample as close to 1 as possible. By constructing our loss functions as extensions of the [CHD+23] loss functions, we maintain some nice properties of

Figure 1: We plot a grid of 49 distinct MLP-based UCCS probes (7 models, 7 datasets), measuring the accuracy@coverage that arises from the unsupervised method. For each model-dataset combination, we run CCS as a baseline, and we use plot hue to display UCCS accuracy minus CCS accuracy for a given setting. UCCS improves upon CCS by several accuracy percentage points in the vast majority of model/dataset combinations.

the [CHD+23] loss functions such as convexity. Finally, we also formulate our overall loss function as the sum of consistency loss and confidence loss, i.e. given dataset $(x_i, y_i) \sim D$, $1 \leq i \leq n$, the unsupervised loss is $L_{\text{UCCS}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} [L_{\text{consistency}}(\theta; x_i) + L_{\text{confidence}}(\theta; x_i)]$.

To convert probe outputs $p_\theta(x^+)$, $p_\theta(x^-)$, and $p_\theta(x^\varnothing)$ into a truth value on x, we first take the greatest of the three outputs as the UCCS output for a text sample. Similarly to CCS, the symmetry inherent to the UCCS loss function may lead $p_\theta(x^+)$, $p_\theta(x^-)$, and $p_\theta(x^\varnothing)$ to map to probabilites of truth, falsehood, and uncertainty in any permutation. As such, at inference time, we try all six permutations of UCCS output to truth label, and report the maximum accuracy across the permutations.

For evaluation datasets with only True/False labels, i.e. all the external benchmark datasets, we cast UCCS as a *selective classifier*; we reject all samples for which UCCS predicts "uncertain", and we report accuracy of True/False predictions only within the accepted region of the dataset. Hence, we follow the convention of the selective classification literature, which, in addition to accuracy, typically reports *coverage* as the percentage of test set samples which are not rejected by the selective classifier [YR17]. For datasets with "Uncertain" labels, we evaluate as a standard classification task. Although none of our external benchmark datasets have uncertainty as an explicit ground-truth label, we later hand-design a dataset where ground-truth labels for uncertainty are baked into the evaluation. (We discuss this custom dataset in detail in Section 3.)

2.2 Experimental Setup

In our experimentation, we aimed to closely mirror the implementation and experimental configurations of the original CCS, as described in [CHD+23]. We note that given limited compute, we were not able to use larger encoder-decoder and encoder only models, such as the 11-billion parameter T5 or UnifiedQA models. We also trained on the generated hidden states from a single prompt for a

given dataset, and did not aggregate or select from multiple prompts on the same dataset. In future work, we hope to experiment with both settings as ablation studies.

Models: We experimented using a diverse array of models, encompassing decoder-only, encoder-only, and encoder-decoder architectures. Namely, we included GPT-J [BA21], GPT-2-large [AJR+19], DeBERTa-XXL [PXJ+21], T5-3b [CNA+20], T0-3b, UnifiedQA-3b [DST+20], and UnifiedQA-v2-3b in all main experiments.

Datasets: We trained and evaluated across various datasets, namely imdb [ARP+11], Amazon Polarity [JJ13], COPA [MCA11], RTE [Ada20], BoolQ [CKM+19], QNLI [AAJ+19], and PIQA [YRR+20].

Probes: Our baseline linear probes incorporated a linear projection succeeded by a sigmoid function. The original CCS employed linear probes in order to extract a single direction in latent space corresponding to latent belief; however, in our work, the relationship between truth, falsehood, and uncertainty/ambiguity may be a complex nonlinear interaction not encapsulated by a single "truth vector." As such, we also employ two-layer ReLU MLPs for our probes, and report MLP probe results as the main results.

For encoder-decoder models like UnifiedQA-3b, our probes were attached to the last layer of the encoder component. In the case of autoregressive decoder-only models, the probes were appended to various intermediate layers, including the terminal layer. We save hidden states for models after ingesting the entire text samples, so that probes operate on the hidden state of the final token of any input text. For optimization, we utilized the AdamW optimizer with a learning rate set to 0.01.

2.3 Results

Our primary results, for 49 trained MLP-based UCCS probes on a grid of 7 models and 7 datasets, are shown in Figure 1. Because all of our evaluation datasets in Figure 1 are binary classification tasks, we use the true/false/uncertain three-way classification of UCCS as a selective classification system, in which all samples with an "uncertain" prediction are rejected, and accuracy is measured only on the "coverage region" where "true" or "false" are predicted. We see that UCCS successfully extends the binary classifier of CCS into a selective binary classifier, attaining a higher accuracy on a limited coverage region of exclusively higher-confidence predictions. Of the 49 UCCS runs in the experiment, 38 attain a higher accuracy than the baseline CCS trained on the same respective model-dataset combination, and 24 runs, constituting half of the overall grid, attain 2.5 percentage points (pp) higher accuracy than the respective CCS. The highest improvement is 12pp for GPT-J on BOOLQ. Meanwhile, the lowest-performing UCCS occurs for UnifiedQA-3B on QNLI, which attains -4.7pp compared to CCS; this is one of only 2 UCCS runs in the grid which fall more than 2pp in accuracy vs. the respective CCS. Interestingly, there is very high variance in the coverage percentages of UCCS probes. On the sentiment classification tasks IMDB and Amazon-Polarity, UnifiedQA-3B attains 96.8% and 99% coverage, respectively, with its UCCS avoiding an "uncertain" prediction for nearly all samples; on the other hand, for entailment task RTE, T5-3B and GPT-J only have 12.4% and 16% coverage, respectively.

3 Temporal Uncertainty Dataset (TYMES)

Label	Statement	Date
True	NASA's James Webb Space Telescope discovered its first exoplanet, which it named LHS 475 b.	January 12, 2023
False	California was the first American state to hit 1 million COVID-19 vaccinations.	January 15, 2021
True	Amazon led a $700 million investment in Rivian.	February 15, 2019
False	Jean-Sebastien Jacques, former CEO of mining corporation Rio Tinto, retired amicably from his position.	September 10, 2020
False	Qatari officials announced that beer would be allowed in limited quantities at the 2022 World Cup.	November 18, 2022

Table 1: Example data from TYMES

All evaluation benchmarks in the previous section were done as selective binary classification tasks, since "uncertain" is not a native label in any benchmark we used or found. Because UCCS is originally trained as a three-way classification including an explicit uncertainty label, which encodes not only low-confidence truth predictions but also related concepts such as knowledge limitation ("I know that I don't know"), we require a dataset tailored for uncertainty in order to comprehensively evaluate the UCCS method. To address this requirement, we introduce the TYMES dataset, and provide details on dataset gathering and evaluation results in this section.

3.1 Dataset Details

TYMES uses whether some piece of real-world information occurred before or after a model's training cutoff as a clean signal for whether a model should be uncertain about the information. For example, a model trained on data up until 2019 can consider a true news statement from 2021 as false or uncertain, while a model trained on data until 2022 should consider the statement true. We constructed a dataset containing true or false statements from news articles (using https://www.random.org/ as our RNG to determine any given sample's label), with a roughly uniform distribution of statements about news ranging from 2018 to 2023. To falsify a sample, we took a valid factoid from a given news source and "bit-switched" a single detail to ensure the statement is still localizable to a specific real news occurrence, but is now false. (For example, a true factoid `"the 'Kiki Challenge' was a social media viral trend in which people danced to the Drake song 'In My Feelings'"` was altered to produce false ground truth sample `"the 'Kiki Challenge' was a social media viral trend in which people danced to the Travis Scott and Drake song 'Sicko Mode'"`.) One delicate detail of the data is that, if data samples are chosen or falsified carelessly, we may have samples where "false" is a reasonable prediction for a model with training cutoff before the sample's date; we crafted our samples to avoid this ambiguity as much as possible, occasionally adding temporal phrases into the samples (i.e. "in late 2021").

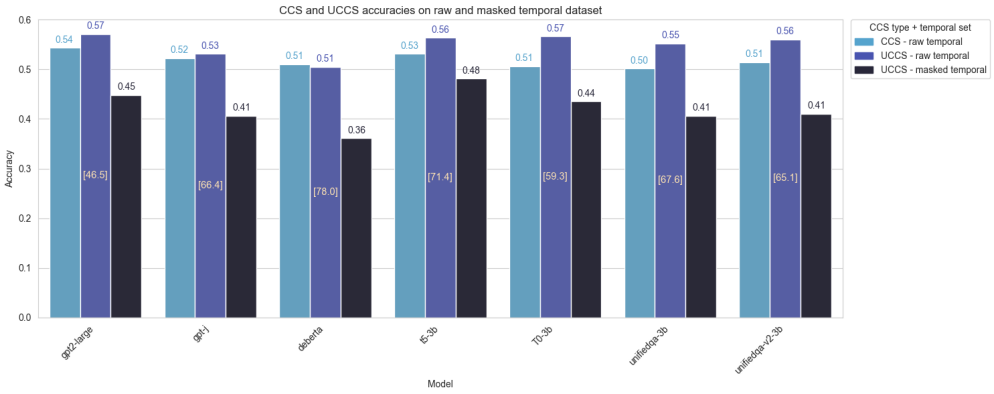

Figure 2: For each model, we train a CCS and UCCS on all of the 7 benchmark sets of Figure 1, and evaluate on TYMES with the original ground truth labels ("raw") or with samples after the model training cutoff relabeled to uncertain ("masked"). The bracketed number in the center bar of each group reports the coverage of UCCS with the raw TYMES set as a selective classification task.

3.2 Results

Our preliminary results on the TYMES dataset are shown in Figure 2. We trained linear CCS and UCCS probes on a grid of our original 7 models and 7 datasets, using the same true/false/uncertain three-way classification of UCCS as a selective classification. Then, we test the trained CCS and UCCS probes on TYMES, and the same UCCS probe on a masked version of TYMES, using a conservative estimate of the model's training cut-off date (often using the last updated model weight timestamp). We see that similar to our previous experiment, UCCS successfully extends the binary classifier of CCS into a selective binary classifier and attains a higher accuracy for most of our models, with the exception of DeBERTa.

We observed significant drops in accuracy in the masked temporal evaluations, from as low as 36% for the probe trained on DeBERTa to 48% for the probe trained on T5-3b. The lower performance may be due to the UCCS probe learning some useful notion of uncertainty with closer resemblance to confidence of the True/False binary value, which differs from what is important for the TYMES dataset – a notion of uncertainty regarding the language model's own epistemic calibration and especially some kind of awareness of its own limitations; namely, the training-cutoff date. A manual inspection on predictions from the temporally masked set reveals that "uncertain" label predictions are distributed roughly evenly across the samples' time range, indicating no awareness of the model's training cutoff. See Appendix 5.1.2 for detailed plans and thoughts on evaluating on TYMES in different directions, including: scale and variety of language models and increased signal from training samples.

4 Related Work

Latent knowledge. A few recent works have investigated extensions of or further experimental qualifications of CCS. [Fab23] reveals potential limitations of the CCS approach: there exist many orthogonal probes that achieve comparable accuracy, indicating the method likely misses important information and does not reliably recover a unique truth-like direction. Additionally, the author finds that CCS probes often overfit and demonstrate high loss on test data, rather than consistently identifying truth-like features. Their work provides useful analysis of the strengths of unsupervised techniques like CCS for belief extraction, while also delineating limitations in the method's ability to locate all relevant knowledge representations and reliably identify truth-like features. On the other hand, [NCC23] explores several directions for improving and analyzing CCS. In particular, they investigate why CCS fails on autoregressive models, finding that factors like sentence length and lack of context are not the main causes; rather, they improve CCS performance on autoregressive models via a new regularization term in the CCS objective that minimizes variance of outputs across paraphrasings of input text.

Uncertainty in language model outputs. [SJO22] proposes training language models to express calibrated uncertainty about their own answers using natural language, which they term "verbalized probability", rather than relying solely on model logits. Through experiments on their CalibratedMath benchmark, they demonstrate that GPT-3 can learn to output verbalized probabilities, like "90% confident", that are reasonably calibrated both in-distribution and out-of-distribution after finetuning. Verbalized probability outperforms baselines including model logits and simple heuristics, with analysis providing evidence that GPT-3 leverages pre-trained representations correlating with uncertainty. [KDT23] examines how natural language expressions of uncertainty impact the behavior of large pre-trained language models, studying the effects both when uncertainty cues are injected into model prompts and when models are trained to generate their own uncertainty expressions. Through prompts spanning multiple QA datasets, they demonstrate that uncertainty expressions significantly alter model accuracy, with surprising gains when weakening language is used over strengthening expressions like factive verbs which consistently hurt performance.

Model calibration. [TES+21] shows that few-shot learning with large language models like GPT-3 can be highly unstable, with accuracy varying dramatically based on small changes to the prompt format, training examples, or ordering of examples. They identify three biases that cause this instability: majority label bias, recency bias, and common token bias. To address this, they propose a simple calibration method called contextual calibration, which estimates the model's biases using a dummy input and adjusts the output probabilities accordingly. [STA+22] investigates methods for improving honesty in language models, defined broadly as truthfulness, calibration, self-knowledge, explainability and non-deceptiveness. They demonstrate that LLMs can be calibrated when predicting answers to multiple choice questions. The models are able to self-evaluate the validity of their open-ended text samples by predicting P(True), though this remains challenging. [KEA+23] evaluates methods for extracting calibrated confidence scores from language models fine-tuned with human feedback. They find that prompting RLHF-LMs to directly verbalize confidence scores produces better calibration than using the model's probabilities. Generating multiple hypotheses before assigning confidence further improves calibration and using linguistic expressions of uncertainty also works well.

References

[AAJ+19] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Sam Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *ICLR* (2019).

[Ada20] Adam Poliak. "A Survey on Recognizing Textual Entailment as an NLP Evaluation". In: *Eval4NLP Workshop at EMNLP* (2020).

[AJR+19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners". In: (2019).

[ARP+11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. "Learning Word Vectors for Sentiment Analysis". In: *ACL* (2011).

[BA21] Ben Wang and Aran Komatsuzaki. *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. 2021.

[CHD+23] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. "Discovering Latent Knowledge in Language Models Without Supervision". In: *ICLR* (2023).

[CKM+19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions". In: *NAACL* (2019).

[CNA+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* (2020).

[DST+20] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. "UnifiedQA: Crossing Format Boundaries with a Single QA System". In: *EMNLP (Findings)* (2020).

[Fab23] Fabien Roger. "What Discovering Latent Knowledge Did and Did Not Find". In: *LessWrong* (2023).

[JJ13] Julian McAuley and Jure Leskovec. "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text". In: *RecSys* (2013).

[KDT23] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B. Hashimoto. "Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models". In: *arXiv preprint arXiv:2302.13439* (2023).

[KEA+23] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. "Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback". In: *EMNLP* (2023).

[MCA11] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning". In: *AAAI* (2011).

[NCC23] Naomi Bashkansky, Chloe Loughridge, and Chuyue Tang. "Surely You're Lying, Mr. Model: Improving and Analyzing CCS". In: *In ICML Workshop on Challenges in Deployable Generative AI* (2023).

[PXJ+21] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". In: *ICLR* (2021).

[SJO22] Stephanie Lin, Jacob Hilton, and Owain Evans. "Teaching Models to Express Their Uncertainty in Words". In: *Transactions on Machine Learning Research (TMLR)* (2022).

[STA+22] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheel El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. "Language Models (Mostly) Know What They Know". In: *arXiv preprint arXiv:2207.05221* (2022).

[TES+21] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. "Calibrate Before Use: Improving Few-Shot Performance of Language Models". In: *ICML* (2021).

[YR17] Yonatan Geifman and Ran El-Yaniv. "Selective Classification for Deep Neural Networks". In: *NIPS* (2017).

[YRR+20] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. “PIQA: Reasoning about Physical Commonsense in Natural Language”. In: *AAAI* (2020).

5 Appendix

5.1 Future Directions

5.1.1 Aligning selective classifiers with ROC curves.

In our main experiments, the three-way classification of UCCS could be cast naturally into a selective binary classification by using the "uncertain" class as a rejection region. One further avenue is to explore other ways of extending the original CCS setup of [CHD+23] to selective classification. CCS includes one step at inference time where a manual threshold is used: namely, after obtaining predictions $p_\theta(x^+)$ and $p_\theta(x^-)$ from the contrast pair, CCS calculates

$$\tilde{p}_\theta(x) := \frac{1}{2}(p_\theta(x^+) + (1 - p_\theta(x^-))$$

and outputs prediction "true" if $\tilde{p}\,(x) \geq 0.5$, "false" otherwise. By generalizing beyond the manual threshold of 0.5 and incorporating stricter thresholds > 0.5, we can possibly bake uncertainty into the original binary classification of CCS. Namely, suppose we have two thresholds $1 > t_{\texttt{True}} > t_{\texttt{False}} > 0$, and we classify an input x as True if $\tilde{p}_\theta\,(x) > t_{\texttt{True}}$; False if $\tilde{p}_\theta\,(x) < t_{\texttt{False}}$; and uncertain otherwise, i.e. if $t_{\texttt{True}} > \tilde{p}_\theta\,(x) > t_{\texttt{False}}$. Indeed, if $\tilde{p}_\theta\,(x)$ is an accurate proxy for the level of confidence in the language model's believed truth value, then this two-threshold approach may produce a selective classifier.

The ROC curve of CCS, produced by varying the manual threshold for $\tilde{p}\,(x)$, hints at how a two-threshold selective classifier may be produced in line with the unsupervised setup of CCS. One potentially fruitful strategy is a "sliding window" approach applied to confidence scores from the binary classifier. For each data sample, the binary classifier outputs a confidence score, which conveys the model's certainty regarding that sample being true. These scores are then organized in ascending order. By sliding a fixed-size window across this sorted list, corresponding to a fixed coverage percentage, we can evaluate all possible decision thresholds for a given coverage. The two boundaries of the window are the two thresholds of our selective classifier; samples inside the window are labeled uncertain, above are labeled true, and below are labeled false. For any window, the true positive rate and false positive rate are easily obtained from the ROC curve, and these two metrics combined with coverage can be benchmarked against our UCCS.

Crucially, our UCCS and proposed two-threshold CCS distinguish themselves from classical selective classification paradigms. In traditional selective classification, the goal is to set aside or reject a subset of samples (determined by desired coverage) with the aim to maximize accuracy on the remaining, confidently classified samples. Essentially, "coverage" in this context relates to how much of the dataset the model is willing to make confident predictions on. In contrast, our strategy does not primarily seek to optimize accuracy within a specified coverage. Instead, our focus is on pinpointing and correctly labeling those samples for which an "uncertain" prediction from the model genuinely reflects the inherent ambiguity of the data point. In other words, we want to align the CCS outputs of the model with more nuanced real-world concepts of truth, from uncertain statements that mix truth and falsehood, to ambiguous statements borne from incomplete information.

5.1.2 Comprehensive training of CCS and UCCS.

In our initial experiments, we worked with a restricted scale of models which we could evaluate with CCS and UCCS. To more thoroughly assess these methods, it would be necessary to conduct experiments on larger models, such as the large 11-billion-parameter variants of T5 and UnifiedQA. Larger models may have greater capacity to learn coherent latent representations related to truth, falsehood, and uncertainty, so it is worth investigating if UCCS (and CCS) see returns in performance with increasing model scale.

Additionally, we trained our probes using the generated hidden states from only a single prompt per dataset. [NCC23] find that "paraphrase invariance," or constraining CCS to be consistent across multiple prompts of the same data, leads to significant improvements in CCS performance. Likewise, it's possible that UCCS performance may see similar returns if UCCS is trained on a variety of prompts at once using the same data. In particular, because uncertainty conflates multiple related concepts in the gray area between truth and falsehood, using multiple prompts for the contrast triplet can help capture all the different flavors of uncertainty. For illustration, using "This is uncertain,"

"This is ambiguous," or "I don't know" are all distinct concepts that UCCS may be capturing, so ablation tests between these different prompts can eludicate what notion(s) of uncertainty UCCS operates on. Varied prompting will also reduce the risk of probes learning spuriously correlated directions rather than meaningful signals about truth and uncertainty, an issue highlighted in prior work [Fab23].

For our experiments using the TYMES dataset, newer, larger models may better capture relevant knowledge about the strict timeline of global events. It's possible that only with a stronger "world model" will a model understand how its own training cutoff impacts its own internal knowledge about past or future factual occurrences. This world model may only emerge from larger model scale, more training FLOPs, or more refined pretraining processes. We hypothesize that the most advanced models will be best equipped to express appropriate uncertainty about facts after their training cutoff dates. We're also interested in ablation tests between models that have gone through instruction tuning or RLHF vs. models that haven't, as non-finetuned models may not have incentivized to learn any features related to temporally conditional uncertainty or veracity.

5.1.3 Expanding the TYMES dataset.

While promising, our initial TYMES dataset for evaluating temporal uncertainty is a very small size, containing around 240 examples, limiting it only to toy dataset or proof-of-concept settings. One potential future goal is to curate a larger collection of timestamped true/falsified news statements, expanding the TYMES dataset to contain thousands of factual claims in the same structure as our existing samples. Scaling up the dataset size will provide more robust conclusions about uncertainty and limitations in language model knowledge and beliefs. A larger-scale model would also be viable as a proper evaluation benchmark.

5.2 Additional Figures

Figure 3: The linear probe results for the same 49 model-dataset configurations as in Figure 1.

ROC plots for DeBERTa

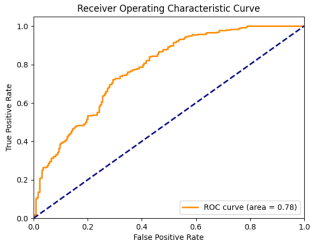

Figure 4: Amazon Polarity

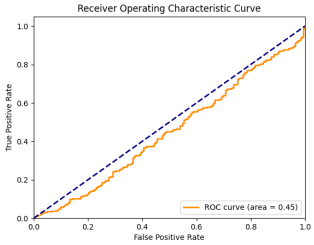

Figure 5: BoolQ

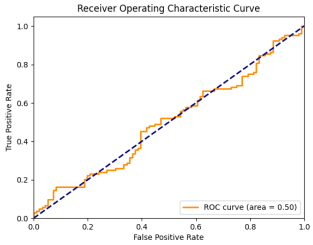

Figure 6: COPA

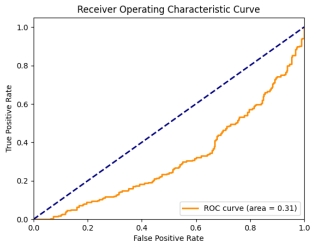

Figure 7: IMDb

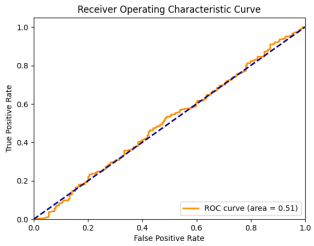

Figure 8: PIQA

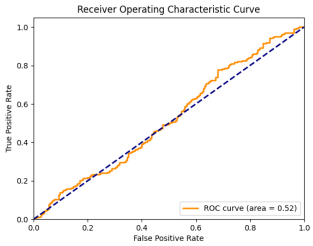

Figure 9: QNLI

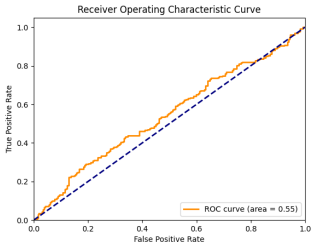

Figure 10: RTE

ROC plots for GPT-J

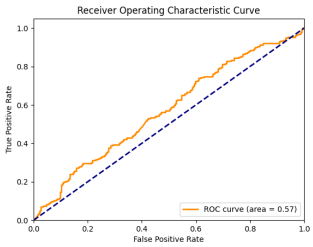

Figure 11: Amazon Polarity

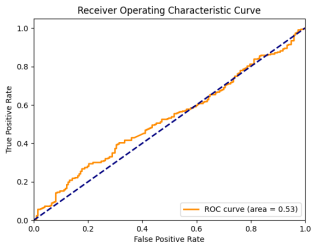

Figure 12: BoolQ

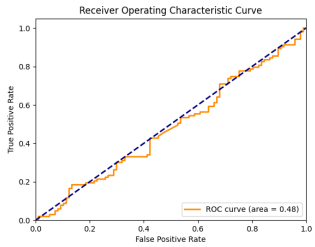

Figure 13: COPA

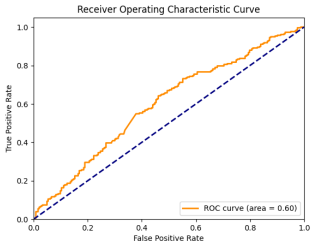

Figure 14: IMDb

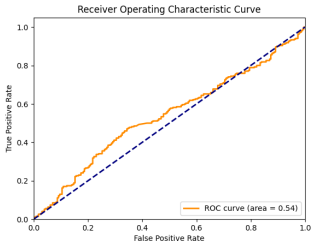

Figure 15: PIQA

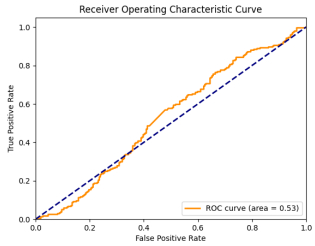

Figure 16: QNLI

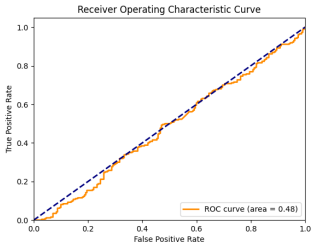

Figure 17: RTE

ROC plots for GPT-2-Large

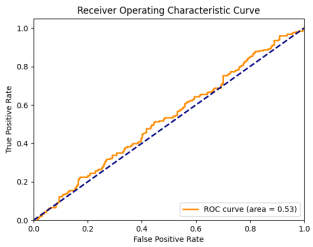

Figure 18: Amazon Polarity

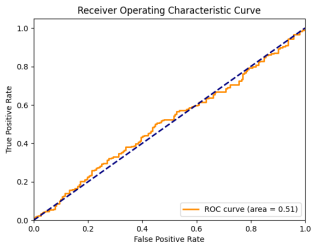

Figure 19: BoolQ

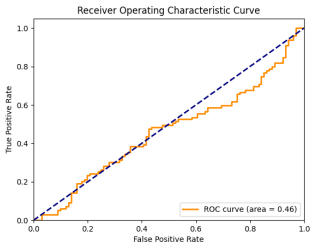

Figure 20: COPA

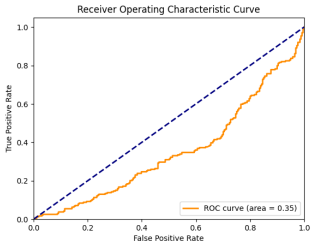

Figure 21: IMDb

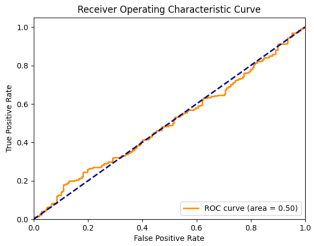

Figure 22: PIQA

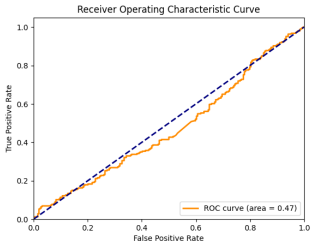

Figure 23: QNLI

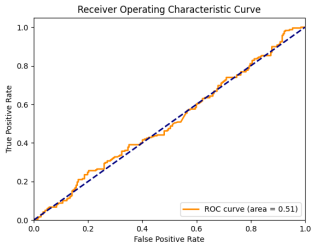

Figure 24: RTE

Linear CCS Results

Model	IMDb			Amazon Polarity			COPA		
	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.
GPT-J	0.506	**0.537**	57.4%	0.516	0.516	55.0%	0.510	**0.602**	64.0%
GPT-2-Large	**0.622**	0.622	46.6%	0.520	**0.549**	51.0%	0.520	**0.524**	10.5%
DeBERTa	**0.934**	0.847	45.8%	**0.778**	0.512	59.4%	0.545	**0.593**	40.5%
T5-3b	0.936	**0.941**	54.2%	0.950	**0.968**	49.4%	0.560	**0.569**	90.5%
T0-3b	**0.564**	0.559	54.0%	0.548	**0.556**	55.4%	**0.535**	0.529	60.5%
UnifiedQA-3b	0.946	**0.959**	97.0%	0.920	**0.943**	49.2%	0.510	**0.524**	63.0%
UnifiedQA-v2-3b	0.934	**0.938**	96.0%	0.938	**0.974**	46.8%	**0.700**	0.513	57.5%

Model	RTE			BoolQ			QNLI		
	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.
GPT-J	0.506	**0.514**	58.0%	0.512	**0.626**	42.8%	0.520	**0.543**	55.2%
GPT-2-Large	0.530	**0.566**	48.4%	0.552	**0.638**	42.6%	0.556	**0.563**	54.0%
DeBERTa	0.518	**0.531**	35.4%	0.510	**0.620**	55.2%	0.534	**0.556**	50.4%
T5-3b	0.514	**0.545**	64.6%	0.506	**0.601**	85.8%	0.542	**0.553**	24.6%
T0-3b	0.510	**0.525**	36.6%	0.502	**0.515**	64.8%	**0.514**	0.514	50.6%
UnifiedQA-3b	0.656	**0.746**	24.4%	0.530	**0.611**	62.8%	0.506	**0.549**	57.2%
UnifiedQA-v2-3b	0.590	**0.682**	13.2%	0.516	**0.575**	45.2%	0.512	**0.527**	48.6%

Model	PIQA		
	CCS Acc.	UCCS Acc.	UCCS Cov.
GPT-J	0.510	**0.543**	14.0%
GPT-2-Large	**0.550**	0.548	60.2%
DeBERTa	0.508	**0.546**	43.6%
T5-3b	0.520	**0.525**	12.2%
T0-3b	**0.530**	0.516	93.8%
UnifiedQA-3b	0.506	**0.562**	60.8%
UnifiedQA-v2-3b	**0.616**	0.584	91.4%

MLP CCS Results

Model	IMDb			Amazon Polarity			COPA		
	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.
GPT-J	0.514	**0.532**	60.2%	**0.546**	0.536	98.8%	0.525	**0.527**	45.5%
GPT-2-Large	0.602	**0.629**	93.2%	0.514	**0.57**	48.4%	0.505	**0.552**	71.5%
DeBERTa	**0.916**	0.88	91.8%	**0.712**	0.7	44.6%	0.525	0.508	65.0%
T5-3b	0.934	**0.951**	49.0%	0.914	**0.94**	47.0%	0.52	**0.62**	89.5%
T0-3b	**0.54**	0.521	53.4%	0.586	**0.606**	55.4%	0.54	**0.575**	60.0%
UnifiedQA-3b	0.948	**0.959**	96.8%	0.948	0.945	99.0%	0.525	**0.524**	52.5%
UnifiedQA-v2-3b	0.936	**0.958**	47.2%	0.928	**0.947**	45.0%	0.525	**0.629**	52.5%

Model	RTE			BoolQ			QNLI		
	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.	CCS Acc.	UCCS Acc.	UCCS Cov.
GPT-J	0.51	**0.55**	16.0%	0.536	**0.656**	58.2%	0.532	**0.559**	49.0%
GPT-2-Large	0.508	**0.527**	51.6%	0.516	**0.614**	68.4%	0.514	**0.554**	53.4%
DeBERTa	0.522	**0.552**	40.6%	0.508	**0.565**	46.0%	0.534	**0.583**	84.4%
T5-3b	0.506	**0.597**	12.4%	0.502	**0.594**	94.0%	0.512	**0.565**	26.2%
T0-3b	0.53	**0.535**	56.4%	0.518	0.517	65.4%	0.512	**0.525**	51.0%
UnifiedQA-3b	0.634	**0.747**	47.4%	0.54	**0.608**	52.0%	**0.54**	0.522	91.6%
UnifiedQA-v2-3b	0.562	**0.608**	52.0%	0.534	**0.625**	60.8%	**0.602**	0.555	49.0%

Model	PIQA		
	CCS Acc.	UCCS Acc.	UCCS Cov.
GPT-J	0.522	**0.56**	53.6%
GPT-2-Large	0.55	**0.585**	43.4%
DeBERTa	0.538	**0.541**	44.4%
T5-3b	0.502	**0.523**	30.6%
T0-3b	0.518	**0.538**	54.6%
UnifiedQA-3b	0.516	**0.535**	45.6%
UnifiedQA-v2-3b	**0.602**	0.589	53.0%