Probing neural language models for understanding of words of estimative probability

Damien Sileo[1] and Marie-Francine Moens[2]

[1]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[2]Department of Computer Science, KU Leuven, Belgium
damien.sileo@inria.fr

Abstract

Words of Estimative Probability (WEP) are phrases used to express the plausibility of a statement. Examples include terms like *probably, maybe, likely, doubt, unlikely*, and *impossible*. Surveys have shown that human evaluators tend to agree when assigning numerical probability levels to these WEPs. For instance, the term *highly likely* equates to a median probability of 0.90 ± 0.08 according to a survey by Fagen-Ulmschneider (2015). In this study, our focus is to gauge the competency of neural language processing models in accurately capturing the consensual probability level associated with each WEP. Our first approach is utilizing the UNLI dataset (Chen et al., 2020), which links premises and hypotheses with their perceived joint probability p. From this, we craft prompts in the form: "[PREMISE]. [WEP], [HYPOTHESIS]." This allows us to evaluate whether language models can predict if the consensual probability level of a WEP aligns closely with p. In our second approach, we develop a dataset based on WEP-focused probabilistic reasoning to assess if language models can logically process WEP compositions. For example, given the prompt "[EVENTA] *is likely*. [EVENTB] *is impossible*.", a well-functioning language model should not conclude that [EVENTA&B] is likely. Through our study, we observe that both tasks present challenges to out-of-the-box English language models. However, we also demonstrate that fine-tuning these models can lead to significant and transferable improvements.

1 Introduction

Expression of uncertainty is an important part of communication. Formal statistics are the rigorous way to quantify uncertainty but do not fit all communication styles. Words of estimative probability (WEP) such as *maybe* and *believe* are adverbs or verbs that are informal alternatives. Kent (1964) noted the importance of clarifying WEP meaning for intelligence analysis in the Central Intelligence Agency, and provided guidelines for mapping WEP to numerical probabilities. Several studies then measured the human perceptions of probability words and discovered some agreement with Kent (1964)'s guidelines. In this work, we use the scale derived from a survey (Fagen-Ulmschneider, 2015), which is the largest and most recent WEP perception survey available. 123 participants were asked to label WEP with numerical probabilities. We use the median of the participant answers to assign a consensual value to each WEP. Associated probabilities for the 19 WEP we use are available in Appendix A, table 2.

Here, we assess whether neural language models learn the consensual probability judgment of WEP from language modeling pretraining. We develop datasets and a methodology to probe neural language model understanding of WEP. The first dataset leverages previously annotated probability scores between a premise and a hypothesis, in order to measure a language model's ability to capture the agreement between numerical probabilities and WEP-expressed probabilities. The second dataset is based on compositions of facts with WEP-expressed probabilities, and measures verbal probabilistic reasoning in language models.

Our contributions are as follows: (i) two datasets and methods to measure understanding of WEP; and (ii) evaluation of the ability of neural language models (GPT2, RoBERTa-trained on MNLI) to tackle WEP-related problems, showing that off-the-shelf models are very little influenced by them, even though fine-tuning on our constructed datasets quickly leads to high accuracies. The code and generated datasets are publicly available[1]

[1] /hf.co/.../probability_words_nli

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 469–476
July 13-14, 2023

2 Related work

Our work probes a particular aspect of language understanding. We do not analyze the inside of the models (Rogers et al., 2020). We focus on the models' ability to perform controlled tasks (Naik et al., 2018; Richardson et al., 2020) involving WEP. WEP were studied in the context of intelligence analysis and linguistics, our work is the first to look at them through natural language processing (NLP) models. Our study also pertains to NLP analyses of logical reasoning and probability problems, and to uncertainty in natural language inference tasks.

Linguistics study of WEP Kent (1964)'s seminal work was the first to link WEP and numerical probability estimates, with intelligence analysis motivations (Dhami and Mandel, 2021) and a prescriptivist approach. This inspired further quantifications of human perceptions of WEP, in the context of medical reports (O'Brien, 1989; Ott, 2021) and weather reports (Lenhardt et al., 2020). Fagen-Ulmschneider (2015) proposed the largest survey up to date with 123 participants about general-domain WEP perception.

Logical and probabilistic reasoning Another strand of work probes NLP text encoders capabilities, notably reasoning abilities. Weston et al. (2015) probed understanding of specific problems like negation, spatial and temporal reasoning with the bAbI dataset. Richardson et al. (2020) probe understanding of first-order logic reasoning, Sileo and Lernould (2023) probe epistemic logic reasoning. Our work is the first to address probabilistic logic, alongside Dries et al. (2017); Suster et al. (2021) who construct a dataset of natural language probability problems, e.g., *"A bag has 4 white and 8 blue marbles. You pull out one marble and it is blue. You pull out another marble, what is the probability of it being white?"*. They also rely on the ProbLog solver (De Raedt et al., 2007), but focus on numeric probability problems. By contrast, our work targets WEP, and textual probabilistic logical reasoning.

Natural language inference, uncertainty, modality, evidentiality Uncertainty was also studied in the context of natural language inference tasks. Zhou et al. (2022) study the disagreement across annotators when labeling entailment relationships. Zhang et al. (2017) annotate graded entailment with 5 probability levels, and the UNLI dataset (Chen et al., 2020) go further by annotating numerical probabilities. Our work also pertains to the study of modality (Palmer, 1992; Saurí et al., 2006) and more particularly evidentiality (Su et al., 2010), but where previous work focused on WEP.

3 Probing WEP understanding

3.1 Verbalization and distractor generation

Our goal is to measure the understanding of WEP. One requirement of WEP understanding is capturing the consensual probability level. To test that, we use contexts (PREMISE) paired with a conclusions (HYPOTHESIS). The likelihood of a conclusion, p, depends on the associated context. One example from UNLI (Chen et al., 2020), which annotates that, is (*A man in a white shirt taking a picture* , *A man takes a picture* , 1.0).

We convert a triplet (PREMISE, HYPOTHESIS, p) to the following verbalization:

$$\text{PREMISE}.\ T_p(\text{HYPOTHESIS}). \quad (1)$$

where T_p is a text template assigned to the probability p. To select a template, we find the WEP whose associated median probability (see table 2) is the closest to p. We then use handcrafted templates to construct a modal sentence from the selected WEP and the hypothesis, e.g., "*It is **certain** that **a man takes a picture***". Table 3 in appendix B displays the templates that we associate with each WEP.

We also generate an invalid verbalization by randomly selecting an incorrect WEP (a WEP whose consensual probability differs from p by at least 40%)[2], e.g., *It is unlikely that a man takes a picture*. We hypothesize that language models and entailment recognition models should give a higher score (respectively likelihood and entailment probability) to the correct valid verbalization than to the invalid verbalization of p.

3.2 WEP-UNLI: probability/WEP matching

The UNLI dataset annotates (PREMISE, HYPOTHESIS) pairs from the SNLI dataset (Bowman et al., 2015) with joint probability scores p, totaling 55k training examples, 3k/3k validation/test examples. We use these examples to generate WEP-understanding dataset with verbalization validity prediction as shown in the previous subsection.

[2]This threshold ensures sufficient distance, while also ensuring that each WEP has at least one possible distractor.

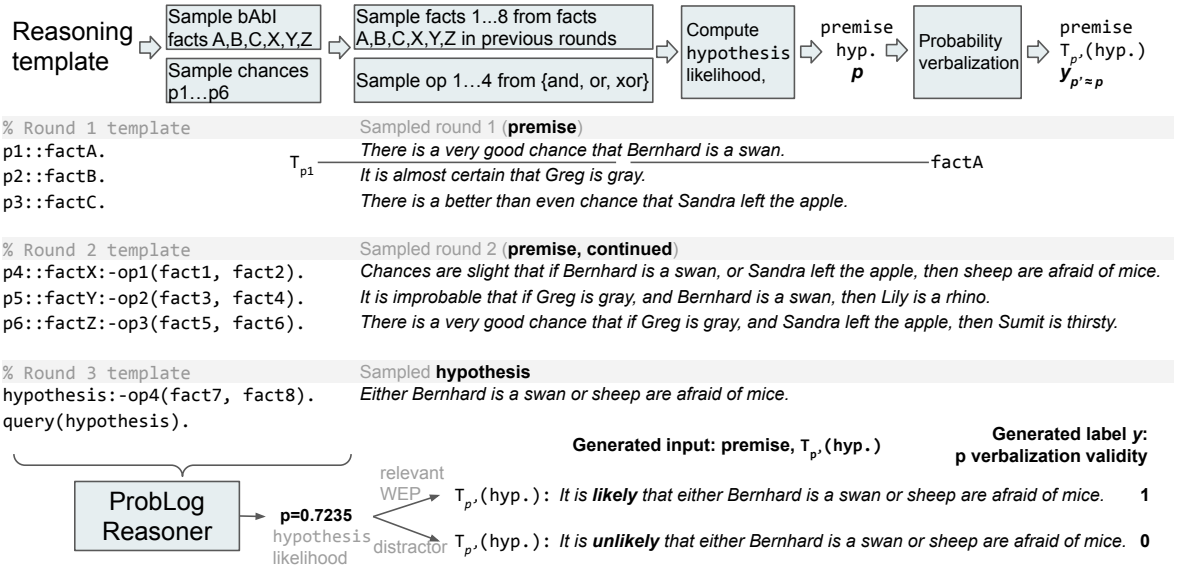

Figure 1: WEP-reasoning task constructions, with 2 hops. We sample randomly concrete facts $fact_i$ and probabilities p_i then build modal sentences with verbalization templates. We randomly sample logical operators to compose the modal sentences from the previous rounds to construct a premise, then a hypothesis, and we use a probabilistic soft logic solver to compute the hypothesis probability. We then correctly and incorrectly verbalize this probability. This process generates data for the task of probability verbalization validity. 1 hop reasoning skips the second round: `fact7` and `fact8` are sampled from {`factA`,`factB`,`factC`}

3.3 WEP-Reasoning: WEP compositions

Here, our goal is to assess models' ability to reason over combinations of probabilistic statements. We construct synthetic (PREMISE, HYPOTHESIS, p) examples from random factoids extracted from the bAbI dataset (Weston et al., 2015). Figure 1 illustrates the construction of WEP-reasoning examples:

We randomly sample initial facts and associated probability levels, and we verbalize them with the previously mentioned templates from Table 3 (Round 1). We further compose them with randomly sampled logical operators (and, or, xor). We then generate a hypothesis with logical combinations of the previous round. Finally, we feed the constructed premise and hypothesis to a probabilistic soft reasoning engine in order to derive the likelihood of the hypothesis given the premise. We rely on the ProbLog (De Raedt et al., 2007) reasoner which implements Dantsin (1992) semantics.

To evaluate different complexities of reasoning, we propose two variants: **2-hop reasoning**, where facts in Round 2 combine facts from Round 1, and the final hypothesis combines facts from Round 2. and **1-hop reasoning** where facts from the hypothesis combine Round 1 facts (Round 2 is skipped).

Since we want to sample more than two facts and we cannot a priori use text from the UNLI dataset, because UNLI only provides entailment likelihood for specific pairs. Combining several sentences could cause unaccounted interference. Therefore, we sample subject/verb/object factoids from the bAbI (Weston et al., 2015) datasets instead, which is built with handwritten arbitrary factoids such as *John went to the kitchen*. To sample multiple factoids, we prevent any overlap of concepts (verb, subject, object) between any pair of facts to make the facts independent of one another.

We sample probability levels from the list of medians of all WEP to prevent sampling the levels that too distant from a known WEP. When we assign a WEP to a probability level, we assume that the correct semantics is the consensual one, but humans differs slightly from this consensus. Still, when adding random perturbations of 20% to sampled $p_{1...6}$, the hypothesis probability is perturbed by less than 40% for 98% of examples.

We generate 5k examples using the template depicted in Figure 1, and use $10\%/10\%$ of the data for the validation/test splits. Appendix C shows the distribution of correct WEP for each dataset.

4 Experiments

We conduct verbalization validity prediction (binary classification task of WEP correctness detection between two candidates) under two settings.

	WEP-Reasoning (1 hop)	WEP-Reasoning (2 hops)	WEP-UNLI
Chance	50.0	50.0	50.0
Human baseline	97.0±1.0	**93.5±1.5**	**89.5±2.5**
GPT2 likelihood zero-shot	50.1±0.0	50.0±0.0	45.6±0.0
RoBERTa likelihood zero-shot	63.4±0.0	63.2±0.0	53.2±0.0
RoBERTa-MNLI zero-shot	49.2±5.4	41.7±4.2	54.6±3.7
RoBERTa+WEP-Reasoning (1 hop) fine-tuning	**97.8±0.4**	81.6±1.3	61.2±0.4
RoBERTa+WEP-Reasoning (2 hops) fine-tuning	85.0±1.6	**91.1±0.1**	62.3±1.7
RoBERTa+WEP-UNLI fine-tuning	62.4±0.4	64.3±0.1	**84.4±0.5**

Table 1: Test accuracy percentage of different models over the 3 WEP-understanding tasks. The last three rows display the accuracy when fine-tuning on each task, and transferability of the fine-tuned model outside the diagonal.

4.1 Zero-shot models

We use off-the-shelf language models to assign likelihood scores to a context and its conclusion. We evaluate the rate at which valid verbalization is scored higher than invalid verbalization. We refine the scores by also considering the average likelihood per token (Brown et al., 2020; Schick and Schütze, 2021) and calibrated scores (Brown et al., 2020; Zhao et al., 2021) where we divide the score of a PREMISE. T_p(HYPOTHESIS). by the score of T_p(HYPOTHESIS). We evaluate the normalized, length-normalized, and calibrated likelihood on the validation sets of each dataset and select the most accurate method for each dataset and model.

We also consider a pretrained natural language inference model, which is trained to predict entailment scores between a context and a conclusion.

GPT2 We use the pretrained GPT2 base version with 127M parameters (Radford et al., 2019), which is a causal language model trained to estimate text likelihood. We concatenate the premise and hypothesis and compute their likelihood as a plausibility score.

RoBERTa We also use the pretrained RoBERTa base model with 123M parameters (Liu et al., 2019) to score the masked language modeling likelihood of the premise/hypothesis pair.

RoBERTa-MNLI We fine-tune RoBERTa on the MNLI entailment detection dataset (Williams et al., 2018) with standard hyperparameters (see the following subsection).

Human baseline To establish human baseline performance on the constructed dataset, we had two NLP researchers annotate 100 examples randomly sampled from the test set of each dataset, with a multiple-choice question answering setting. Overall inter-annotator agreement is relatively high, with a Fleiss's κ of 0.70/0.68/0.71 for WEP Reasoning 1 hop, 2 hops and WEP-UNLI respectively.

4.2 Fine-tuning and transfer across probes

We fine-tune RoBERTa-base models on our datasets, using standard (Mosbach et al., 2021) hyperparameters[3] (3 epochs, sequence length of 256, learning rate of 2.10^{-5} batch size of 16. We use length-normalization with GPT2 likelihood and calibration with RoBERTa likelihood as they worked best on the validation sets.). We use a multiple-choice-question answering setup (we predict logit scores for the valid and invalid verbalization, combine their score with a softmax, then optimize the likelihood of the valid verbalization). The same format is applied to all tasks, so we can also study the transfer of capacities acquired during fine-tuning of each probe, for instance, between probability matching and compositional reasoning.

4.3 Results and discussion

Table 1 shows the results of our experiments. The very low accuracy of causal and masked language models (first two rows) demonstrates how challenging the WEP-understanding tasks are.

RoBERTa fine-tuned on MNLI dataset performs better than chance for WEP-UNLI. MNLI contains 814 instances of *probably* in the MNLI dataset, but we found little to no evidence of WEP compositions among them, which can explain the results.

Finally, fine-tuning on the dataset of a particular probe leads to high test accuracy on the associated test set. More surprisingly, fine-tuning on one dataset also causes substantial accuracy gain on other probes. This suggests that our datasets can

[3]Deviation from these hyperparameters did not yield significant improvement on the validation sets.

be incorporated in text encoder training in order to improve WEP handling.

5 Conclusion

We investigated WEP understanding in neural language models with new datasets and experiments, showing that WEP processing is challenging but helped by supervision which leads to transferable improvement. Future work could extract WEP probability scales from the UNLI dataset as an alternative to human perception surveys, but our work suggests that this requires language modeling progress.

6 Acknowledgements

This work is part of the CALCULUS project, which is funded by the ERC Advanced Grant H2020-ERC-2017 ADG 788506[4].

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Eugene Dantsin. 1992. Probabilistic logic programs and their semantics. In *Logic Programming*, pages 152–164, Berlin, Heidelberg. Springer Berlin Heidelberg.

Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. 2007. Problog: A probabilistic prolog and its application in link discovery. In *IJCAI*, volume 7, pages 2462–2467. Hyderabad.

Mandeep K Dhami and David R Mandel. 2021. Words or numbers? communicating probability in intelligence analysis. *American Psychologist*, 76(3):549.

Anton Dries, Angelika Kimmig, Jesse Davis, Vaishak Belle, and Luc de Raedt. 2017. Solving probability problems in natural language. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3981–3987.

Wade Fagen-Ulmschneider. 2015. Perception of probability words.

Sherman Kent. 1964. Words of estimative probability. *Studies in intelligence*, 8(4):49–65.

Emily D Lenhardt, Rachael N Cross, Makenzie J Krocak, Joseph T Ripberger, Sean R Ernst, Carol L Silva, and Hank C Jenkins-Smith. 2020. How likely is that chance of thunderstorms? a study of how national weather service forecast offices use words of estimative probability and what they mean to the public. *Journal of Operational Meteorology*, 8(5).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

B J O'Brien. 1989. Words or numbers? the evaluation of probability expressions in general practice. *The Journal of the Royal College of General Practitioners*, 39 320:98–100.

Douglas E Ott. 2021. Words representing numeric probabilities in medical writing are ambiguous and misinterpreted. *JSLS: Journal of the Society of Laparoscopic & Robotic Surgeons*, 25(3).

F.R. Palmer. 1992. Words and worlds; on the linguistic analysis of modality. (european university studies, series xiv, vol. 191): Richard matthews, frankfurt am main/bern/ new york/paris, peter lang, 1991. 310 pp. sfr 76.00 (pb.). *Lingua*, 88(1):87–90.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721.

[4] https://calculus-project.eu/

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Roser Saurí, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, Melbourne Beach, Florida, USA, May 11-13, 2006*, pages 333–339. AAAI Press.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Damien Sileo and Antoine Lernould. 2023. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Qi Su, Chu-Ren Huang, and Kai-yun Chen. 2010. Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 10–17, Uppsala, Sweden. Association for Computational Linguistics.

Simon Suster, Pieter Fivez, Pietro Totis, Angelika Kimmig, Jesse Davis, Luc de Raedt, and Walter Daelemans. 2021. Mapping probability word problems to executable representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3627–3640, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

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. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed nli: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

A Associated probabilities

WEP	Median probability judgment
certain	$100^{\dagger}$
almost certain	95.0 ± 10.9
highly likely	90.0 ± 8.4
very good chance	80.0 ± 10.8
we believe	75.0 ± 15.0
likely	70.0 ± 11.3
probably	70.0 ± 12.9
probable	70.0 ± 14.7
better than even	60.0 ± 9.1
about even	50.0 ± 4.9
probably not	25.0 ± 14.4
we doubt	20.0 ± 16.9
unlikely	20.0 ± 15.0
little chance	10.0 ± 12.2
chances are slight	10.0 ± 10.9
improbable	10.0 ± 17.5
highly unlikely	5.0 ± 17.3
almost no chance	2.0 ± 17.0
impossible	$0^{\dagger}$

Table 2: Median probability percentage associated to words of estimative probability according to (Fagen-Ulmschneider, 2015). First and last words (†) are taken from (Kent, 1964).

B WEP verbalization template

WEP	Verbalization template
about even	*chances are about even that* [FACT]
almost certain	*it is almost certain that* [FACT]
almost no chance	*there is almost no chance that* [FACT]
better than even	*there is a better than even chance that* [FACT]
certain	*it is certain that* [FACT]
chances are slight	*chances are slight that* [FACT]
highly likely	*it is highly likely that* [FACT]
highly unlikely	*it is highly unlikely that* [FACT]
impossible	*it is impossible that* [FACT]
improbable	*it is improbable that* [FACT]
likely	*it is likely that* [FACT]
little chance	*there is little chance that* [FACT]
probable	*it is probable that* [FACT]
probably	*it is probably the case that* [FACT]
probably not	*it is probably not the case that* [FACT]
unlikely	*it is unlikely that* [FACT]
very good chance	*there is a very good chance that* [FACT]
we believe	*we believe that* [FACT]
we doubt	*we doubt that* [FACT]

Table 3: Templates used to convert a fact and a WEP expressed uncertainty into a modal sentence.

C WEP frequencies on the generated datasets

WEP-reasoning	(1 hop)	WEP-Reasoning	(2 hops)	WEP-USNLI	
WEP	frequency	WEP	frequency	WEP	frequency
about even	11.1	*impossible*	13.2	*impossible*	25.6
probably not	9.7	*about even*	10.8	*better than even*	10.7
better than even	7.7	*probably not*	9.0	*certain*	7.2
we believe	7.1	*highly unlikely*	8.2	*about even*	6.9
highly likely	6.4	*almost no chance*	8.0	*almost certain*	6.7
certain	6.0	*better than even*	6.6	*highly likely*	6.0
highly unlikely	5.9	*we believe*	4.3	*very good chance*	5.9
almost no chance	5.8	*highly likely*	4.0	*almost no chance*	5.0
impossible	5.3	*very good chance*	4.0	*we believe*	4.1
almost certain	5.1	*we doubt*	4.0	*highly unlikely*	4.1
very good chance	4.7	*improbable*	3.9	*probably not*	3.4
chances are slight	3.6	*chances are slight*	3.9	*likely*	2.5
little chance	3.5	*unlikely*	3.6	*probable*	2.4
probable	3.2	*little chance*	3.5	*probably*	2.4
unlikely	3.1	*almost certain*	2.9	*unlikely*	1.5
likely	3.1	*certain*	2.7	*little chance*	1.5
probably	3.0	*likely*	2.5	*chances are slight*	1.5
we doubt	2.9	*probable*	2.4	*improbable*	1.4
improbable	2.9	*probably*	2.2	*we doubt*	1.4

Table 4: Validation set frequency of WEP in the correct answer of each dataset (percentages).