Claim-level Uncertainty Estimation through Graph

Stanford CS224N Custom Project

Mingjian Jiang
Department of Computer Science
Stanford University
jiangm@stanford.edu

Abstract

In the realm of natural language processing, Large Language Models (LLMs) like GPT-4 have set benchmarks for generating coherent responses across a diverse range of user queries. Yet, the propensity of these models to fabricate information or "hallucinate" poses a significant challenge, undermining the reliability of their outputs. This paper introduces a novel approach to uncertainty estimation tailored to claims within long-form text generations without assumptions of any resource retrieval or model internal access, aiming to fortify trust in LLM outputs. Unlike traditional methods that apply uncertainty estimation at a broader claim level, our methodology utilize more information through graph structure. Through comparative analysis against standard baselines, our approach demonstrates superior performance in identifying hallucinated content, with marked improvements in handling obscure or "long-tail" knowledge domains. Furthermore, we pointed out a prototype of uncertainty-aware decoding that effectively diminish the incidence of hallucinations. This advancement not only contributes to the enhancement of LLM reliability but also paves the way for future research in the domain of trustworthy AI.

1 Key Information to include

- Mentor: Tatsu Hashimoto (Instructor), Tianyi Zhang (CA).
- External Collaborators (if you have any): Yangjun Ruan, Prasanna Sattigeri, Salim Roukos.
- Sharing project: None.

2 Introduction

In the rapidly evolving landscape of artificial intelligence, Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text (OpenAI et al., 2024; Team et al., 2023; Touvron et al., 2023) , which has profound implications for a myriad of applications ranging from automated content creation (Agossah et al., 2023) to real-time decision support systems (Umerenkov et al., 2023). However, LLM is not trusted in many application areas like medical since it often generates ungrounded or hallucinated output (Bang et al., 2023; Guerreiro et al., 2023). Instances of hallucination and unreliability in model outputs not only compromise the integrity of the generated content but also pose substantial risks in high-stakes scenarios such as medical diagnostics, legal advice, and safety-critical systems. Thus, improving the mechanisms for estimating and communicating the uncertainty of LLM generations becomes not just an academic pursuit but a crucial step towards mitigating the risks associated with their deployment.

The scholarly exploration into methods of uncertainty estimation unveils promising avenues to tackle these challenges, advocating for a shift towards more accountable and reliable artificial intelligence applications. Traditional uncertainty quantification, predominantly focused on classification tasks, has recently garnered interest in the context of Natural Language Generation (NLG), highlighting the necessity for innovative approaches. Prevailing methods (Kuhn et al., 2023; Tian et al., 2023; Lin et al., 2023) often provide a singular uncertainty score for an entire text output, which proves uninformative for longer generations and lacks the granularity needed for effective manipulation and enhancement of the decoded text. This highlights a pressing need for more refined uncertainty estimation techniques, such as claim-wise uncertainty, which offer a more detailed understanding and improvement scope for generated text.

In this study, we introduce a novel approach to uncertainty estimation and further improve decoding in Large Language Models (LLMs) that offers significant advancements over traditional methods. Our contributions include:

- We present a novel graph-based framework that constructs a bipartite graph from LLM outputs to analyze the relationships between various outputs and their claims. This approach leverages closeness centrality to assess the credibility of claims, providing a comprehensive tool for hallucination detection in NLG.
- Alongside our methodology, we pointed out a prototype of uncertainty-aware decoding that explicitly demonstrates the effectiveness of integrating claims from multiple generations alongside our graph-based uncertainty estimation method.

3 Related Work

3.1 Uncertainty Estimation of LLM

The exploration of uncertainty quantification has established itself as a pivotal field of inquiry within various machine learning disciplines, including natural language processing (NLP). Prior studies have predominantly been classified into three methodologies: likelihood-based approaches (Kadavath et al., 2022; Kuhn et al., 2023), consistency-based approaches (Xiong et al., 2023), and verbalization-based strategies (Lin et al., 2022; Tian et al., 2023). Notably, consistency-based methods are often regarded as a form of Monte Carlo estimator for likelihood-based approaches, typically operating under a black-box assumption.

Nonetheless, a significant portion of the extant literature focuses on quantifying the uncertainty of entire generative outputs, which inherently restricts these analyses to relatively brief and unidimensional narratives. Aiming for a more refined analysis, Manakul et al. (2023) advances the concept of self-consistency (Wang et al., 2023) to assess uncertainty at the sentence level within extended textual outputs, presupposing a black-box large language model (LLM) framework. Building upon this, Mohri and Hashimoto (2024) further elaborates this approach to the level of individual claims, with a particular emphasis on conformal prediction techniques.

4 Approach

Task. Given an prompt x and its output y from an Language Model, we want to break the output y into a set of claims C that are included in y. Meanwhile, for each claim $c \in C$, it is associated with a score s_c such that the score is positively correlated to the correctness of c. It is crucial to highlight that our approach treats the language model as a black-box entity, meaning that we operate without access to the model's internal details or requiring any additional resources.

The high level idea of our method is: given a prompt input x, we can sample several generations from our LLM from using the same input, and the relationship between generations and all claims within them could be represented as a bipartite graph. We propose an method using some graph-based metrics like closeness centrality as a good uncertainty indicator, and show it is highly correlated to its correctness. In the subsequent sections, we will detail the construction of the consistency graph, the derivation of the uncertainty score, and an prototype of uncertainty-aware decoding.

4.1 Graph Construction

This section outlines our methodology for constructing a bipartite graph $G = ((N_1, N_2), E)$ from a given input x, using a large language model (LLM). The constructed graph G captures the relationships between the LLM's outputs and the claims contained within these outputs. Specifically,

N_1 denotes the set of outputs generated from the input x, N_2 represents the set of claims identified within those outputs, and an edge $e \in E$ indicates the association between an output $o \in N_1$ and a claim $c \in N_2$. We detail the construction process for the output nodes (N_1), claim nodes (N_2), and the edges (E) below.

Output Nodes: Upon receiving an input prompt x, such as "Tell me a bio of Billy Snedden," we generate an initial output answer g_0 using an LLM with the temperature parameter set to $t = 0$. Subsequently, we generate $P - 1$ alternative outputs with the temperature parameter adjusted to $t = 1$, resulting in a series of generations $g_1, \ldots, g_{P-1}$. This process produces a set of generations $N_1 = \{g_0, \ldots, g_{P-1}\}$, where the size of N_1 is equal to P.

Claim Nodes: With the set N_1 constructed, we employ a method analogous to those described in Mohri and Hashimoto (2024); Min et al. (2023), prompting the same LLM to decompose its long-form output into discrete claims for each $g_i \in N_1$, denoted by $BD(g_i) = C_i$, where C_i is a set of claims contained within g_i. The prompt is detailed further in appendix.

To amalgamate all distinct claims from C_i based on semantic similarity, we prompt the LLM to merge C_i into a comprehensive set of claims. Formally, we define a comprehensive union of all unique, semantically distinct claims as $\mathcal{C}$, with the power set of $\mathcal{C}$ represented by $\mathcal{P}(\mathcal{C})$.

We introduce a union function $\mathcal{M} : \mathcal{P}(\mathcal{C}) \times \mathcal{P}(\mathcal{C}) \to \mathcal{P}(\mathcal{C})$, where $s \in \mathcal{M}(S_1, S_2) \iff s \in S_1$ or $s \in S_2$. This function is approximated by sequentially prompting the LLM to merge two sets of claims, formally, $H_0 = C_0, H_i = \mathcal{M}(H_{i-1}, C_i), \forall 1 \geq i \geq n$. This will result in a cumulative set H_n that encompasses all claims across all generations, thus forming our set of claim nodes N_2.

Edge Construction: The bipartite graph is constructed by linking output generations in N_1 to the claims in N_2, where an edge between a generation g and a claim c is established if g directly mentions c. The methodology for determining the existence of an edge is aligned with practices outlined in previous studies, leveraging LLM prompts for accurate determination. We adopt the same prompt from Manakul et al. (2023).

4.2 Uncertainty Estimation from Graph

Drawing on the premise that claims enjoying broader support tend to be more proximate to all nodes within a graph, we leverage the principle of closeness centrality for a claim within such a graph as a metric to gauge uncertainty. Specifically, the uncertainty $U(v)$ associated with a fact v is quantified by its closeness centrality, mathematically expressed as:

$$U(v) = \frac{N-1}{\sum_u d(u,v)}, \tag{1}$$

where N represents the aggregate count of nodes in the graph, and $d(u, v)$ denotes the distance between nodes u and v. Closeness centrality thus serves to mirror a claim's capacity for dissemination throughout the network, indicative of its potential for broad recognition and corroboration across generations.

To further refine this model, we introduce three heuristic distance measures: the shortest path length distance (d_{vanilla}), the verbalized confidence distance (d_{vc}), which cumulates the LLM's verbalized confidence deficits along a given path, and the combined distance (d_{combined}) that synthesizes both elements. For any two nodes u, v within a graph, let the shortest path between them be denoted as $p = (p_1, p_2, ..., p_n)$ with $p_1 = u$ and $p_n = v$. To adjust the indentation of the following list, we use the 'enumitem' package:

- Shortest path length distance: $d_{\text{vanilla}}(u, v) = \text{Length}(p) = n$
- Verbalized confidence distance: $d_{\text{vc}}(u, v) = \sum_{i=1}^{n}(1 - vc(p_i))$
- Combined distance: $d_{\text{combined}}(u, v) = d_{\text{vanilla}}(u, v) + d_{\text{vc}}(u, v)$

Notably, the conceptualization of distance is predicated on the shortest path between two nodes. While our exposition presumes a singular shortest path for simplicity, in instances of multiple shortest paths, an average of the distances as defined by each path is computed.

These metrics are designed not only to elucidate the structural attributes of the graph but also to integrate the LLM's confidence levels, thus offering a nuanced perspective on uncertainty. Contrary to the principle of self-consistency, exemplified in prior works such as Manakul et al. (2023) and Wang et al. (2023), which is limited to data from immediate neighbors (with a distance of 1), our approach exploits the graph's extensive architecture to consider interactions with multi-hop neighbors. This enables a more holistic exploration of uncertainty by leveraging the graph's wider connections.

4.3 Decoding with Uncertainty Aware

In our preceding exploration, we unveiled a technique for meticulously estimating uncertainty on an individual claim basis. The current section unfolds a comprehensive framework designed for uncertainty-aware decoding, which is instrumental in refining the generation of coherent long-form text. This refinement is achieved by judiciously selecting claims that exhibit lower uncertainty levels from a wide-ranging candidate pool. Here, we elucidate the operational mechanics of this framework by integrating the introduction of its four fundamental components with their formal definitions.

The proposed framework for uncertainty-aware decoding seeks to optimize the generation of coherent long-form content by prioritizing claims with lower uncertainty from a diverse array of candidates. The decoding process within this framework is structured around four pivotal components: the uncertainty estimation method, the claim selection pool, the threshold criteria for claim selection, and the algorithm for integrating the selected claims into a cohesive narrative output.

Formally, we denote the uncertainty estimation function as $U : \mathcal{C} \to \mathbb{R}$, where $\mathcal{C}$ symbolizes the entire set of potential claims. The subset of these claims considered for selection is represented by $P \subset \mathcal{C}$, with δ serving as the threshold for determining claim selection. The integration function, $M : \mathcal{P}(\mathcal{C}) \to \Sigma^*$, then maps the chosen subset of claims into a seamless textual output. The operational subset of claims, $P^o = \{c \in P | U(c) < \delta\}$, forms the basis from which the Language Model constructs its final output, denoted as $M(P^o)$.

Within this framework, the approach presented in Mohri and Hashimoto (2024) can be interpreted as follows: the uncertainty estimation method, $U(\cdot)$, is implemented via the claim-level SelfCheckGPT technique; the claim pool, P, is derived from BD(g_0); and the merging function, M, prompts the LM to amalgamate the claims. We propose an innovative claim-wise uncertainty quantification approach as a viable alternative for $U(\cdot)$. Additionally, we expand the claim pool, P, to incorporate claims sourced from multiple generation cycles, denoted as H_n in Section 4.1, thereby enhancing the model's capacity to generate nuanced and contextually rich outputs. We engage the same Language Model to reconstitute H_δ into a novel output, employing the methodology delineated in Mohri and Hashimoto (2024) as the merging function $M(\cdot)$. This process underscores the flexibility and adaptability of our approach in generating content that is both relevant and contextually comprehensive.

5 Experiments

5.1 Data

Data This study harnesses subsets from two distinct datasets, FactScore and PopQA, to examine the effectiveness of uncertainty estimation.

- **FactScore Dataset** Our research utilizes a subset from FactScore (Min et al., 2023), which includes 183 entities linked to Wikidata and Wikipedia, focusing on a subset of 40 entities with around 1000 claims each, annotated as True, False, or Subjective. The annotation leverages GPT-4-turbo, chosen for its low error rate, as outlined in the FactScore methodology (Min et al., 2023), ensuring accurate claim classification.

- **PopQA Dataset** Additionally, we incorporate the PopQA dataset (Mallen et al., 2023), containing 14,000 questions on a diverse range of subjects. We also focus on a subset of 40 entities, convert the data to input prompt like 'Provide me with a paragraph detailing some facts related to {subject}'.

5.2 Uncertainty Quantification Experiments

Baseline methods Current literature lacks works specifically focused on uncertainty estimation at the claim level, presenting a challenge in identifying directly comparable baselines. We plan to modify existing approaches for our purpose of claim-level uncertainty estimation. We will consider two methods for adaptation:

- SelfCheckGPT (Manakul et al., 2023), which utilizes a method of generating multiple outputs from the same prompt and selecting the most frequent response for each sentence, focuses on sentence-level analysis. We propose adapting this technique to assess uncertainty at the claim level by tailoring it to the specific needs of evaluating individual claims.
- The approach of verbalized confidence, as introduced by Lin et al. (2022), involves the model explicitly stating its confidence in its assertions.

Evaluation methods In evaluating our model, we utilize the Area Under the Receiver Operating Characteristic (AUROC) curve and the Area Under the Precision-Recall Curve (AUPRC) as our primary metrics. The AUROC serves as a fundamental measure of binary classification, indicating the model's proficiency in distinguishing between two classes. A higher AUROC score suggests a stronger ability to correlate uncertainty with hallucination rates. Similarly, the AUPRC is crucial for assessing performance in imbalanced class distributions, focusing on the precision-recall balance and the model's effectiveness in identifying positive instances amidst numerous negatives. A higher AUPRC signifies better precision and recall performance, complementing the AUROC in evaluating classification accuracy comprehensively.

Experimental details The LLM that we are using for paragraph generation are GPT-3.5-turbo and GPT-4. To construct the set of claims N_2, we use a greedy decoded generation (temperature $t = 0$) and $N = 4$ generations with temperature $t = 1$. As for the set of generations G, we are using $M = 5$ or $M = 10$ generations where 5 of the generations are those obtaining the claims, and the others are also generated with temperature $t = 1$.

We collect all the claims in the data we used, label them using the method provided in Min et al. (2023). Then, we filter out those claims annotated as 'subjective', thus, all the others can be determined as True or False. This will results in a subset of claims $C^o \subset C$. We calculate their uncertainty and compute AUROC and AUPRC, where the results are shown in Table 1.

Results and Analysis Table 1 we find that our proposed method consistently higher than the baseline methods, even a near 10% gain in GPT-3.5-turbo case, and the gain is relatively robust for different sizes of generation set, M. This suggests that our method excels in identifying correctness-correlated uncertainty within these datasets, thereby enhancing the detection of hallucinations.

	Setup	GPT-3.5, $M = 5$		GPT-3.5, $M = 10$		GPT-4, $M = 5$		GPT-4, $M = 10$	
	Metric	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
FactScore	SelfCheckGPT	0.831	0.81	0.852	0.836	0.811	0.821	0.823	0.852
	Verbalized	0.781	0.71	0.781	0.7	0.723	0.727	0.711	0.731
	CC (d_{vanilla})	0.902	0.89	0.904	0.894	0.85	0.851	0.858	0.873
	CC (d_{vc})	0.884	0.866	0.882	0.857	0.8	0.827	0.792	0.83
	CC (d_{combined})	**0.92**	**0.907**	**0.922**	**0.911**	**0.862**	**0.873**	**0.863**	**0.882**
PopQA	SelfCheckGPT	0.677	0.519	0.704	0.577	0.693	0.577	0.698	0.58
	Verbalized	0.578	0.455	0.614	0.495	0.517	0.486	0.515	0.484
	CC (d_{vanilla})	**0.717**	**0.631**	0.74	0.662	**0.755**	**0.725**	**0.751**	**0.713**
	CC (d_{vc})	0.621	0.493	0.684	0.633	0.505	0.53	0.511	0.534
	CC (d_{combined})	0.704	0.608	**0.753**	**0.687**	0.601	0.594	0.613	0.607

Table 1: AUROC obtained from different methods using GPT-series models with number of generations $M \in \{5, 10\}$. CC stands for closeness centrality we proposed using different distance metric d. Results are presented separately for two different datasets, with a vertical column indicating the dataset.

5.3 Experiments on Uncertainty-Aware Decoding

Dataset For our experiments, we utilized the same dataset as described in Section 5.2, ensuring consistency across our analyses.

Baseline Methods Our study benchmarks the performance of uncertainty-aware decoding against zero-resource decoding methods. We delineate the decoding configurations as follows:

1. **Greedy Decoding:** For a given input prompt x, this method generates a response with a temperature setting of $t = 0$. This approach is widely acknowledged for producing outputs with high likelihood and serves as a fundamental baseline.
2. **Conformal Factuality Decoding (Mohri and Hashimoto, 2024)**: This method, referred to as 'SelfCheckGPT + Greedy Generation', utilizes the self-consistency uncertainty estimation method (SelfCheckGPT, as detailed in Section 5.2) to exclude claims of high uncertainty from the output generated through greedy decoding.
3. **SelfCheckGPT + Multiple Generations:** Implements the SelfCheckGPT method for uncertainty estimation ($U(\cdot)$) across multiple generations, aggregating claims from $P = \bigcup_{i=0}^{n} BD(g_i)$.
4. **CC + Multiple Generations:** Applies our proposed closeness centrality (CC) method with d_{combined} for uncertainty estimation ($U(\cdot)$), utilizing a claim pool aggregated from multiple generations, $P = \bigcup_{i=0}^{n} BD(g_i)$.

Evaluation Metrics To assess the efficacy of long-form text generation, we report on two critical dimensions: the accuracy of the generated content (measured by FactScore as introduced in Min et al. (2023)) and the quantity of true claims within the output. These metrics are averaged across the dataset to provide a comprehensive view of performance.

Experiment Details Our findings are presented in a scatter plot (Figure 1), with accuracy on the y-axis and the quantity of true claims on the x-axis, highlighting the preference for methods positioned towards the upper right. Uncertainty-aware decoding methods delineate a trajectory within the plot when varying the uncertainty estimation threshold, represented by points (x_i, y_i) corresponding to specific threshold settings. In contrast, the greedy decoding method is depicted as a single point due to its lack of threshold variation. The experiments utilize GPT-3.5-turbo on the FactScore dataset, aiming to illustrate the superiority of uncertainty-aware decoding in bolstering the informativeness and reliability of generated content. This visual comparison elucidates the trade-offs between information density and accuracy, shedding light on the inherent strengths and limitations of each method.

Results and Analysis The analysis of Figure 1 reveals several key insights:

- The comparison between greedy decoding and conformal factuality decoding underscores the accuracy benefits derived from incorporating uncertainty estimation, albeit at the cost of reducing the volume of useful information.
- The evaluation of conformal factuality decoding against SelfCheckGPT + Multiple Generations indicates that utilizing a claim pool from multiple generations outperforms mere greedy decoding in terms of accuracy for a given quantity of useful information, and conversely, provides more useful information for a given level of accuracy. This suggests that incorporating multiple generations into the claim pool significantly enhances generation quality.
- The comparison between SelfCheckGPT + Multiple Generations and CC + Multiple Generations highlights the efficacy of our CC uncertainty method in improving uncertainty-aware decoding, emphasizing the value of a superior uncertainty estimation method in elevating generation quality.

6 Conclusion

In conclusion, our study introduces an practical approach to uncertainty estimation in text generated by LLMs with zero-resource required. This innovative graph-based methodology, alongside the

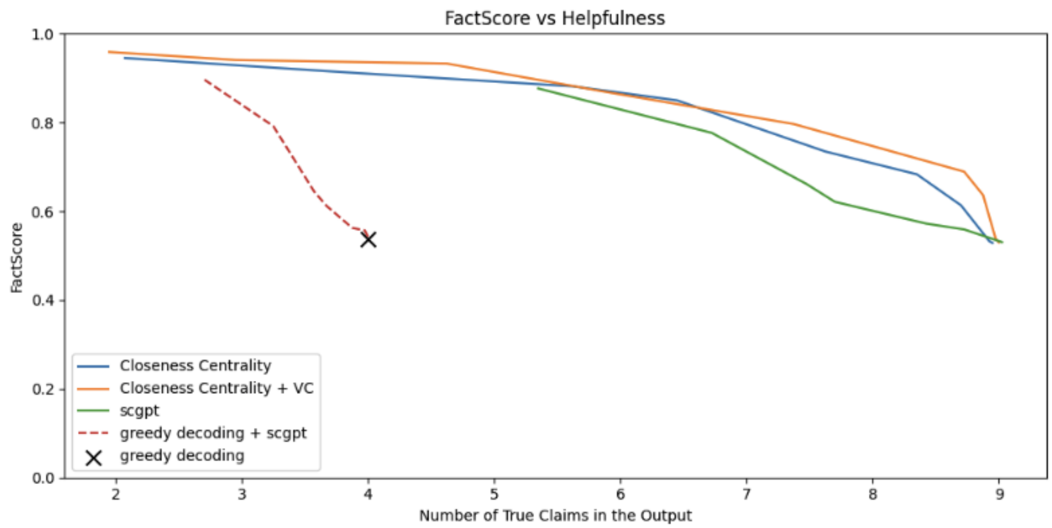

Figure 1: This plot shows the tradeoff of output accuracy (FActScore) and the number of true claims included in the output in the Factscore dataset, as the threshold varing.

development of an uncertainty-aware decoding prototype, marks a significant advance over traditional methods. By leveraging a bipartite graph to intricately map the relationships between outputs and claims and employing closeness centrality for assessing claim credibility, we offer a robust tool for hallucination detection in NLG. Furthermore, the introduction of an uncertainty-aware decoding prototype underscores the practicality of our approach in real-world applications, demonstrating significant advancements over traditional uncertainty quantification methods.

Our methodology's zero-resource nature signifies that it does not rely on extensive additional datasets or external computational resources beyond what is already required for LLM operation. This aspect not only enhances the accessibility of our approach but also underscores its practicality for a wide range of applications.

However, it's crucial to note the limitations associated with the intensive computation demanded by the method's requirement for self-prompting multiple times. While this ensures detailed uncertainty estimation, it may lead to impractical computational costs for some applications. Future work will need to focus on optimizing the computational efficiency of our approach, balancing the comprehensive nature of our uncertainty estimation with the need for computational pragmatism.

By addressing these computational challenges and further refining our methodology, we aim to make our uncertainty estimation approach even more versatile and applicable across various domains. This will enhance the reliability, interpretability, and overall utility of LLM-generated text, paving the way for more accountable and trustworthy AI applications in the future.

References

Alexandre Agossah, Frédérique Krupa, Matthieu Perreira Da Silva, and Patrick Le Callet. 2023. Llm-based interaction for content generation: A case study on the perception of employees in an it department.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer 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, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted

Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor

Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey

Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

D. Umerenkov, G. Zubkova, and A. Nesterov. 2023. Deciphering diagnoses: How large language models explanations influence clinical decision making.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

A Appendix (optional)

If you wish, you can include an appendix, which should be part of the main PDF, and does not count towards the 6-8 page limit. Appendices can be useful to supply extra details, examples, figures, results, visualizations, etc. that you couldn't fit into the main paper. However, your grader *does not* have to read your appendix, and you should assume that you will be graded based on the content of the main part of your paper only.