Teaming LLMs to Detect and Mitigate Hallucinations

Demian Till[1]* **John Smeaton**[1] **Peter Haubrick**[1] **Gouse Saheb**[1]
Florian Graef[1] **David Berman**[2]
[1] Cambridge Consultants [2] Queen Mary University of London

Abstract

Recent work has demonstrated state-of-the-art results in large language model (LLM) hallucination detection and mitigation through consistency-based approaches which involve aggregating multiple responses sampled from a single LLM for a given prompt. These approaches help offset limitations stemming from the imperfect data on which LLMs are trained, which includes biases and under-representation of information required at deployment time among other limitations which can lead to hallucinations. We show that extending these single-model consistency methods to combine responses from multiple LLMs with different training data, training schemes and model architectures can result in substantial further improvements in hallucination detection and mitigation capabilities beyond their single-model consistency counterparts. We evaluate this *consortium consistency* approach across many model teams from a pool of 15 LLMs and explore under what conditions it is beneficial to team together different LLMs in this manner. Further, we show that these performance improvements often come with reduced inference costs, offsetting a significant drawback with single-model consistency methods.

1 Introduction

A well-known, major limitation of current LLMs is their propensity to hallucinate, producing plausible but factually-incorrect responses. Quality of pre-training data and instruction fine-tuning data plays a key role in hallucination behavior. When information relevant to deployment-time performance is under-represented or misrepresented in the pre-training corpus, the model is less likely to be able to provide accurate responses [1, 2]. Moreover, instruction fine-tuning can incentivize models to make educated guesses in the absence of reliable knowledge on a given topic [3]. During instruction fine-tuning it is relatively expensive to determine what a model genuinely does not know and to include fine-tuning examples which encourage it to admit when it does not know something. Such examples are therefore likely to be underrepresented in typical fine-tuning data, thereby providing insufficient counterbalance to the pressure to make educated guesses which often result in hallucinations [4].

Self-consistency [5] effectively mitigates a class of hallucinations by sampling multiple generations from an LLM in response to a given prompt and a final answer is selected by taking a majority vote over the responses. This approach, and follow up work [6] demonstrated improvement over alternative methods such as debate [7] and self-reflection [8], and that smaller models using this approach can match or surpass the accuracy of substantially stronger models. This method can be understood as mitigating a class of hallucinations where a model is able to produce correct answers in response to a given prompt more often than not, whether by means of imperfect recall or intelligent inference based on known/provided information.

*Corresponding author: demian.till@cambridgeconsultants.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Reliable ML from Unreliable Data.

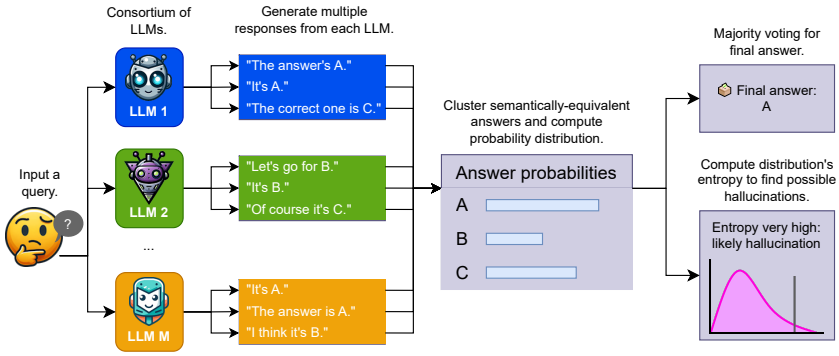


Figure 1: Illustration of consortium consistency. A given query is input to multiple LLMs, one or more responses are sampled from each model. Semantically-equivalent answers are clustered together, and the probability distribution of different answers is computed from these clustered samples. The distribution is used to calculate a final answer to the query, and an entropy score. Queries with higher entropy have less consistent responses and are hence more likely to contain hallucinations. Combining responses from multiple different LLMs reduces the likelihood of incorrectly assigning high confidence to hallucinated answers and allows consistently hallucinating models to be out-voted by other models.

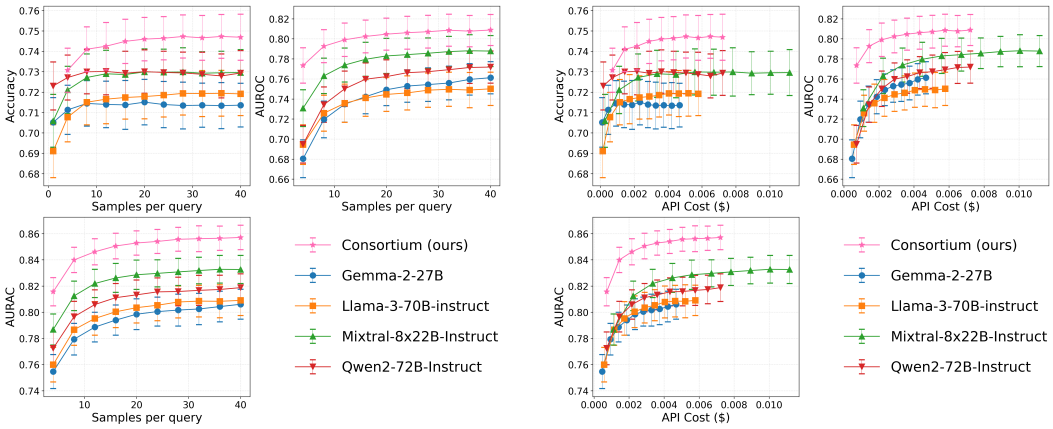
Semantic entropy [9, 10] uses a similar consistency-based approach to detect likely hallucinations by grouping together similar generations sampled in response to a given prompt and computing the entropy over the resulting clusters of responses. An LLM is deemed more likely to be hallucinating when its responses to a given prompt have higher semantic entropy, reflecting greater uncertainty and more guesswork. The authors showed that semantic entropy achieves greater hallucination detection accuracy than alternative methods, including ones requiring white-box model access and model training [10]. [11] extend this idea to compute consistency across multiple responses based on semantic information within internal model embeddings rather than output text, achieving state-of-the-art (SOTA) hallucination detection results.

However these single-model consistency approaches naturally fail in cases where models produce relatively consistent hallucinations in response to a given prompt. In these cases the wrong answer can win the majority vote (hallucination mitigation failure) and semantic entropy can be low [9] or internal embeddings can be semantically similar [11], indicating that the answer is unlikely to be a hallucination (hallucination detection failure). We hypothesize that heterogeneous models with different training data, training methods and model architectures are less likely to share the same shortcomings in their training data or to making the same educated guesses. Heterogeneous collections of models therefore ought to be less prone to the aforementioned failure modes which arise when using single-model consistency approaches.

This motivates extending single-model consistency methods to incorporate multiple different LLMs. We therefore propose *consortium voting* and *consortium entropy* as multi-model counterparts to self-consistency (single-model voting) and semantic entropy respectively. These multi-model formulations, which we refer to collectively as *consortium consistency*, work in tandem to select answers and estimate confidence in selected answers from a pool of candidate responses generated by two or more LLMs. Other consistency-based hallucination detection methods such as [11–14] could similarly be extended to use multiple different LLMs, however we leave this to future work, and in the case of [11] this would require aligning the embedding spaces of different LLMs.

We compare consortium consistency with *single-model consistency*, which analogously uses self-consistency and semantic entropy in tandem with a single model. We evaluate both approaches on a set of 11 tasks, testing for reasoning capabilities, general knowledge, and domain-specific knowledge across a variety of domains. We explore consortia formed using various combinations from a pool of 15 different LLMs, ranging from 6B to 141B parameters in size, and using a range of different architectures, training methods, and training datasets.

We find that for many combinations of models, consortium voting and consortium entropy substantially outperform their single-model consistency counterparts whilst simultaneously reducing



(a) Performance versus number of samples

(b) Performance versus API cost

Figure 2: (a) A representative example showing consortium consistency improving on average across 11 test sets over single-model consistency applied to each of the constituent models, across a range of sample budgets per-query. (b) Consortium consistency dominates single-model consistency on the cost-performance frontier, achieving both higher performance and lower cost simultaneously. X-axes show mean API cost in dollars per query, which grows with increasing number of sampled responses per query.

inference costs. However we also find that these performance gains are sensitive to consortium composition i.e. which LLMs are teamed together. We therefore investigate under what conditions consortium consistency tends to deliver the strongest results compared to single-model consistency.

In summary, our main contributions are[2]:

- We propose *consortium voting* and *consortium entropy*, collectively referred to as *consortium consistency*: black-box, post-training methods which further advance LLM hallucination mitigation and detection capabilities beyond their single-model consistency counterparts.
- We evaluate these methods using a wide variety of tasks, across a broad range of consortia composed of varying combinations of models from a pool of 15 diverse LLMs, finding that under reasonable constraints in consortium composition, consortium consistency outperforms single-model consistency when controlling for sample budget and model availability.
- We investigate which factors regarding consortium composition result in the most reliable improvements in performance compared to single-model consistency baselines, finding that performance gains tend to be greatest when all of the LLMs in a consortium are similarly capable and relatively strong (i.e. high-performing LLMs are better at complementing the capabilities of other high-performing LLMs).
- We additionally find that sometimes stronger models are able to benefit from being teamed with much weaker models, resulting in substantially reduced inference cost compared with single-model consistency, whilst simultaneously boosting hallucination detection and mitigation performance compared with single-model consistency with the entire response budget allocated to the stronger model.

2 Methodology

Our approach is illustrated in Figure 1. Given an input query x, a set of models $\mathcal{M} = \{m_1, m_2, \ldots, m_{|\mathcal{M}|}\}$, and a total sampling budget of N responses, we begin by sampling $N/|\mathcal{M}|$ responses from each model i.e. evenly distributing our sample budget over the M models. Each response is generated independently using nucleus (top-p) sampling with temperature scaling [15]. These responses are then clustered based on semantic equivalence as described below. We propose

[2]Code to follow.

two related methods: *consortium voting* and *consortium entropy*, for respectively generating a final answer and providing a confidence estimate in that answer being a hallucination. These methods are straightforward multi-model generalizations of the single-model consistency-based methods: *self-consistency* [5] and *semantic entropy* [9]. Since they work together in tandem to select and estimate confidence in answers, we refer to them collectively as *consortium consistency*, and we similarly refer to self-consistency and semantic entropy collectively as *single-model consistency*.

2.1 Semantic clustering of responses

Similar to [5, 9], consortium consistency requires first clustering the N responses into a set of semantically distinct equivalence classes $\mathcal{C} = \{C_1, C_2, \ldots, C_{|\mathcal{C}|}\}$, where all responses within a cluster are considered equivalent in meaning and $|\mathcal{C}|$ is determined automatically by the clustering algorithm. For example given the prompt "What is the capital of France", the responses "Great, question! Paris is the capital of France", and "The capital of France is Paris" would be considered semantically equivalent for our purposes.

To determine equivalence of responses when clustering, we follow [6] in using task-specific approaches. For multiple-choice tasks, responses are deemed equivalent if they select the same final option, regardless of their reasoning paths. For math tasks, responses are deemed equivalent if their final answers are mathematically equivalent, again regardless of reasoning paths. We also use these equivalence checks when comparing final answers against ground truth answers during evaluation. More general equivalence checking is possible, e.g by prompting another LLM to determine equivalence as in [9], but for convenience we restrict our focus to domains where equivalence can be computed algorithmically.

2.2 Multi-model response generation via consortium voting

Given a set of clustered responses, consortium voting determines the final answer via majority voting. That is, it determines which cluster has the most responses across all M models:

$$\text{answer} = \arg\max_{C_i \in \mathcal{C}} \sum_{m \in \mathcal{M}} \sum_{j=1}^{N/|\mathcal{M}|} \mathbf{1}[r_{m,j} \in C_i] \tag{1}$$

where $r_{m,j}$ denotes the j-th response sampled from model m, and $\mathbf{1}[\cdot]$ is the indicator function.

We compare this to single-model majority voting (referred to in the literature as *self-consistency* [5]), where all N responses are drawn from a single model:

$$\text{answer}_{\text{single}} = \arg\max_{C_i \in \mathcal{C}} \sum_{j=1}^{N} \mathbf{1}[r_j \in C_i] \tag{2}$$

2.3 Hallucination detection via consortium entropy

To estimate hallucination likelihood, we extend semantic entropy [10] from single-model settings to multi-model consortia. For input query x, we estimate the consortium's distribution over equivalence classes f as:

$$P(C_i \mid x) = \frac{1}{N} \sum_{m \in \mathcal{M}} \sum_{j=1}^{N/|\mathcal{M}|} \mathbf{1}[r_{m,j} \in C_i] \tag{3}$$

Then the consortium entropy is the semantic entropy over the clustered responses from all models in a consortium:

$$\text{SE}(x) = -\sum_{C_i \in \mathcal{C}} P(C_i \mid x) \log P(C_i \mid x) \tag{4}$$

As in [9], the semantic entropy for a given input query reflects the level of diversity of responses across distinct semantic equivalence classes. A value of zero indicates unanimous agreement, while higher values indicate greater uncertainty and therefore a greater likelihood of hallucination. Unlike token-level entropy, which may overstate uncertainty due to superficial differences such as different

ways of phrasing equivalent responses, semantic entropy captures uncertainty at the level of response *meaning*.

We compare consortium entropy to single-model semantic entropy [9, 10] where all N responses are drawn from a single model. In the single-model case the distribution over clusters simplifies to:

$$P_{\text{single}}(C_i \mid x) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}[r_j \in C_i] \tag{5}$$

3 Experimental setup

3.1 Evaluation metrics

We report evaluation accuracy following [5, 6], as well as AUROC and AURAC following [9, 10]. Accuracy simply measures the percentage of evaluation inputs answered correctly. We use this as a proxy for hallucination *mitigation*, with higher accuracy generally indicating fewer hallucinations. For hallucination *detection*, we use AUROC to evaluate how well consortium entropy and single-model semantic entropy are able to distinguish correct from incorrect final answers, aggregating over all classification thresholds.

We also report *area under rejection accuracy curve* (AURAC), introduced in [10]. *Rejection accuracy* is the accuracy when only considering a subset of questions on which semantic entropy scores are above a given threshold. Less confident answers are considered potential hallucinations, and rejection accuracy effectively measures the resulting accuracy if the approach were to abstain from answering those questions. AURAC aggregates rejection accuracy across all confidence thresholds.

3.2 Baselines

Given the impressive results achieved by the single-model consistency methods that we extend, we use these single-model consistency methods as baselines. We evaluate consortium consistency on many different selections of M models comprising different *consortia*. For each consortium we compare its performance on the metrics defined above with that of applying single-model consistency using individual models from the M models in the consortium, controlling for sample budget.

Specifically, when we evaluate a given consortium of M models, using a sample budget of N responses per-question, we also evaluate the result of applying the single-model consistency methods to each of the M models, in each case with the full sample budget of N responses all allocated to that one model. We use these single-model consistency scores to define three baselines against which to compare the consortium score:

- **Hard baseline:** the highest of the M single-model consistency scores on a given metric. This is the most difficult baseline to beat as it assumes that we know which of the M models will perform best on the test data using single-model consistency (which often would not be known in practice).
- **Standard baseline:** the median of the M single-model scores. This represents average performance of single-model consistency methods in the common case where we do not know *a priori* which of the M models is best suited to the target domain.
- **Worst-case baseline:** the lowest of the M single-model scores. This represents the worst-case performance of single-model consistency methods where the least suitable model for a given target domain is selected.

3.3 Sampling procedure

Unless otherwise specified, we generate $N = 40$ responses per input prompt, either (i) distributed evenly across the consortium of models $\mathcal{M}$, or (ii) drawn entirely from a single model (when evaluating single-model consistency). When $|\mathcal{M}|$ is not a factor of 40, we use the largest multiple of $|\mathcal{M}|$ less than 40 (e.g., $N = 39$ for $|\mathcal{M}| = 3$) and use the same N for the single-model baselines to ensure fair comparison. All responses are sampled independently using nucleus sampling with top-$p = 0.9$ and temperature $= 0.5$, and chain-of-thought prompting [16], unless otherwise specified.

Table 1: Benefits of consortium consistency over single-model consistency baselines when composing consortia using well-matched, strong models. Results are averaged over the 586 consortia which met the following criteria: standard deviation of constituent mock benchmark scores ≤ 5 and mean constituent mock benchmark score ≥ 70. Each row reports either the mean percentage change in score vs the corresponding baseline (± std) or the percentage of teams that outperform the corresponding baseline (i.e. where the change in score vs the baseline is positive).

Metric	**Baseline**	**Accuracy** ↑	**AUROC** ↑	**AURAC** ↑
Mean score Δ (%)	Hard	+1.33 $\pm$ 1.03	+1.84 $\pm$ 1.48	+2.75 $\pm$ 0.69
	Standard	+3.70 $\pm$ 1.20	+5.63 $\pm$ 1.46	+5.39 $\pm$ 1.09
	Worst-case	+9.67 $\pm$ 3.44	+18.80 $\pm$ 10.41	+16.20 $\pm$ 7.22
% of teams improved	Hard	92	92	100
	Standard	99	100	100
	Worst-case	100	100	100

3.4 Uncertainty estimation

Unless otherwise specified, each evaluation metric for each model/consortium is computed using 100 bootstrap samples over the input questions and model/consortium responses. Reported values are mean averages across these samples. Where shown, error bars indicate one standard deviation across the bootstrap samples.

3.5 Models

We evaluate consortium consistency using varying subsets of models from a pool of 15 LLMs ranging in size from 6B to 141B parameters. These include models from the LLaMA, Mistral, Qwen, and Gemma families (full list in Appendix C). We experiment with different strategies for selecting which models to team together into consortia.

3.6 Datasets

We evaluate consortia and baselines on 11 tasks covering reasoning and general/domain knowledge:

- **Reasoning:** GSM8K [17] (200 randomly sampled questions), GPQA-Diamond [18].
- **General and domain knowledge:** 8 MMLU [19] subsets covering virology, world religions, jurisprudence, astronomy, public relations, anatomy, college chemistry, and global facts. TruthfulQA [2], which probes for common misconceptions.

We report metrics averaged across all 11 tasks to approximate performance in mixed-domain, real-world deployment settings.

3.7 Separate tasks for model selection

The strategies we propose for selecting which models to team together into consortia consider the relative and absolute capability levels of the candidate models. Ideally public benchmark scores would be used for these purposes, however we could not find any public benchmarks covering all 15 models. We therefore compiled a separate set of tasks with which to estimate a *mock benchmark score* for each model (detailed in Appendix D).

3.8 Compute costs

Gathering and processing all of the LLM responses discussed in this paper cost approximately $1000. The majority of this cost resulted from the API costs required for sampling 40 responses per question across the 11 main datasets and the 15 models we used.

4 Results

4.1 Performance with well-matched, strong models

Figure 2a shows a representative example of the benefits of consortium consistency over single-model consistency when applied to a set of similarly capable, relatively strong models. Consortium consistency outperforms single-model consistency applied to any of the individual constituent models across all three metrics and across a wide range of response budgets. Table 1 summarizes results aggregated across many such consortia, identified using the mock benchmark scores of their constituent models (detailed in Section D). Specifically we select teams with (i) a standard deviation in mock benchmark scores below 5 points and (ii) a mean mock benchmark score above 70%.

Across these consortia, consortium consistency delivers significant improvements over all single-model consistency baselines. Of particular note, consortium consistency outperforms the hard baseline in the vast majority of cases ($\geq 92\%$ of teams on each metric; see Table 1). The performance gap widens further against the other baselines, with over 99% of consortia outperforming the standard baseline on each metric, and all consortia outperforming the worst-case baselines.

4.2 Impact of model strength

Figure 3b shows how the advantage of consortium consistency over the hard baseline is impacted by the mean strength of the models within a consortium, as measured by their mean mock benchmark scores. We observe that as mean strength increases, the advantage of consortium consistency over the hard baselines grows more reliable across all three metrics. It was not immediately obvious to us why this should happen, since as mean strength increases, so do the baseline scores against which consortia are evaluated.

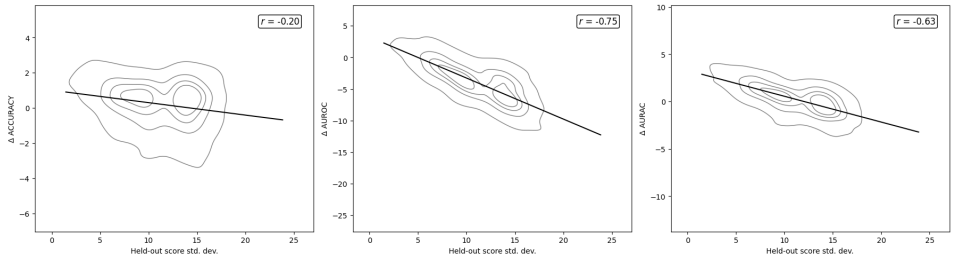
We hypothesize that this result is in part due to more capable models being more likely to make more intelligent (less random) guesses and mistakes, making them more likely to generate consistent (rather than random) hallucinations in response to some queries. Weaker models on the other hand tend to produce more varied responses when they hallucinate, corresponding to more random guesses. This can result in lower semantic entropies when stronger models hallucinate, making such hallucinations more difficult to detect using single-model semantic entropy. This leaves more room for consortium entropy to benefit from different models being less likely to all hallucinate in the same way, making low entropy on incorrect answers less likely.

Table 5 shows detailed results vs baselines for consortia selected based on high mean model strength. Compared with results from selecting based on both high mean model strength and low variance in model strength Section 4.1, AUROC scores are significantly lower, however 81% of these consortia still beat the hard baseline, with all of the 1580 consortia evaluated beating the standard and worst-case baselines across all three metrics.

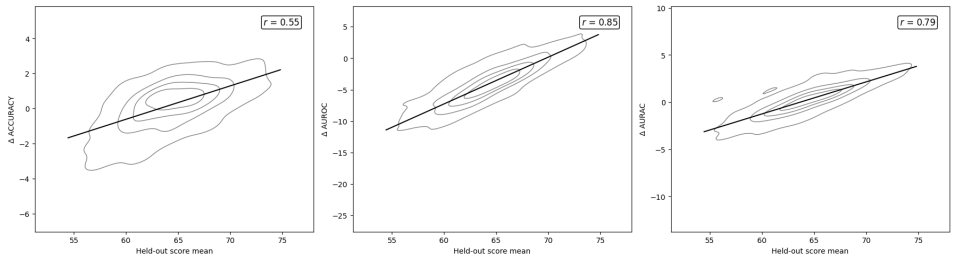
4.3 Impact of variance in model capability

Figure 3a shows how the advantage of consortium consistency over the hard baseline is impacted by diversity in model capability within a consortium, as measured by the standard deviation of the constituent models' mock benchmark scores. We observe that as variance in model capability decreases, the advantage of consortium consistency over the hard baselines grows more reliable across all three metrics. This aligns with intuition: a relatively strong model is less likely to benefit by sharing its response sample budget with substantially weaker models than with other similarly capable models. However, interestingly we see in figure 3a that in the case of accuracy, many consortia with high diversity in model capabilities still exhibit substantial improvements over the hard baseline.

Table 4 shows detailed results for consortia selected based on low-variance in model capability only. Compared to selecting based on mean model capability (Section 4.2), AUROC improvements over the hard baseline are less reliable, with only 68% of consortia beating the hard AUROC baseline. We investigate this in Section A, finding that at the lower entropy regions, consortium consistency maintains a strong advantage over single-model consistency, however when using weaker models, this is often outweighed by consortia being more prone to producing higher entropies on correct responses due to a significantly higher chance of "dissenting opinions" when weaker models are used.



(a) Impact of diversity of mock benchmark scores on consortium performance vs hard baselines



(b) Impact of mean individual model mock benchmark scores on consortium performance vs hard baselines

Figure 3: KDE plots showing performance of consortia vs hard baselines as a function of varying properties of constituent models. Plots are generated using 1000 consortia randomly selected from all $2^{15} - 15$ consortia that can be formed from the available models.

4.4 Cost-performance tradeoffs

Figure 2b compares performance against approximate API cost for consortium consistency and single-model consistency for a representative consortium and each of its constituent models. Across all evaluation metrics, the consortium dominates the single-model consistency baselines on the cost-performance frontier, achieving both higher performance and lower cost simultaneously. Note that this is an expected result because the strongest individual model is on average likely to be the most expensive. Therefore reallocating some of its sample budget across a consortium including cheaper models results in lower cost as well as the performance improvements which come with consortium consistency. See Section E for more examples of detailed plots for varied consortium compositions.

5 Related works

5.1 Hallucination detection

White-box methods have explored using token output probabilities [20–22] to calculate uncertainty scores, as well as training hallucination detection models using an LLM's internal embeddings [23, 24]. Black-box methods have explored prompting LLMs to provide confidence scores [25] and sampling multiple responses and evaluating consistency across responses [12, 13, 9, 10]. [26] uses a verifier model to check the answers of a target model, but is limited to teaming two models together in this manner. Other methods combine consistency across multiple samples with white-box model access [14], with [11] achieving SOTA results. Our method builds on [9, 10], inheriting the benefit of being black-box, and extends their approach to use an arbitrary number of different models, reducing the chance of unanimous agreement on hallucinated responses. However other consistency-based methods such as [11–14] could similarly be extended to use multiple models, and we believe they would likely see similar improvements as a result, but we leave this to be explored in future work.

5.2 Hallucination mitigation

While typically not explicitly framed as addressing hallucinations, certain consistency-based approaches [5, 27, 6] can be seen as mitigating a subset of hallucinations. Similarly, works combining the strengths of multiple models, both before generation of response(s) via model selection [28–31], during generation [32–34], and combining responses after generation [7, 35–37], can be understood as mitigating a different set of hallucinations arising from inherent limitations of individual models. Note that these works are concerned with improving the accuracy of generated answers rather than *detecting* hallucinations. Our work leverages the hallucination mitigation advantages of sampling multiple generations per-model and using multiple diverse models, whilst also estimating confidences in the selected responses which can be used for hallucination detection. Another line of work tackling hallucination mitigation involves retrieval-augmentation [38–40]. We see these approaches as largely orthogonal and potentially complementary to consistency-based approaches, and future work could look at combining them.

6 Conclusions and limitations

In this paper, we extended the single-model consistency methods: *self-consistency* and *semantic entropy*, proposing corresponding consortium consistency methods: *consortium voting* and *consortium entropy*. We demonstrated that consortium consistency improves over single-model consistency for mitigating and detecting hallucinations across a wide range of consortia, and does so whilst simultaneously reducing inference costs. We also analyzed the impacts of the capabilities of constituent models on consortium consistency performance relative to single-model consistency baselines, finding useful rules of thumb to help select models which are more likely to work well together in consortia. We hope that these results help motivate and guide future work in combining multiple LLMs for hallucination detection and mitigation.

Thus far, our evaluation has focused on average performance across 11 tasks. Further work could investigate how performance vs single-model consistency baselines varies with task diversity. Our hypothesis is that greater task diversity reduces the likelihood of any single model dominating, thereby enhancing the effectiveness of multi-model approaches. Conversely, in settings with more narrowly focused tasks, it may more often be preferable to rely on single-model consistency using the single strongest model in that domain when the best model is known.

While consistency-based approaches to hallucination detection and mitigation have achieved SOTA results, these come at the expense of substantially increased inference costs due to the need to sample multiple responses. This limits their applicability to situations where performance requirements outweighs inference cost concerns. We have seen that a multi-model approach can partially alleviate the increased costs due to the ability to combine models with varying inference-time demands, however this approach still incurs greater costs than more lightweight methods.

Recent work [37] indicates that stronger models can exhibit more similar failure modes, which could limit the benefits of multi-model consistency approaches. However, within the range of models explored thus far, we have observed the opposite, with stronger models benefiting more than weaker models from being teamed together. This could be the result of an interplay between two factors: the convergence of cross-model failure modes with increasing model strength indicated in [37] (which would harm consortium consistency), and the potentially even greater propensity for stronger single models to hallucinate more consistently (which would harm single-model consistency).

Another limitation of our approach arises when queries require knowledge of niche topics that few models within the consortium are experts in. Based on the current setup, a single expert model can be out-voted if some of the other models share a hallucination, perhaps based on some incorrect data points they share within their training data. To overcome this, further work may explore weighted aggregation based on known model strengths or per-model confidence estimates, to help ensure that authority can sometimes win out over consensus.

Acknowledgments

We are grateful to James Oldfield, Douglas O'Rourke, David Rimmer, Rupert Thomas, and Joe Corrigan for valuable discussions and feedback.

References

[1] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

[2] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/. arXiv:2109.07958 [cs].

[3] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

[4] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can AI assistants know what they don't know? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8184–8202. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cheng24i.html.

[5] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations*, May 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. arXiv:2203.11171 [cs].

[6] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More Agents Is All You Need. *Transactions on Machine Learning Research*, October 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bgzUSZ8aeg. arXiv:2402.05120 [cs].

[7] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of the 41st International Conference on Machine Learning*, pages 11733–11763. PMLR, July 2024. URL https://proceedings.mlr.press/v235/du24e.html. ISSN: 2640-3498, arXiv:2305.14325 [cs].

[8] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 8634–8652, Red Hook, NY, USA, December 2023. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3666122.3666499. arXiv:2303.11366 [cs].

[9] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *Proceedings of the Eleventh International Conference on Learning Representations*, May 2023. URL https://openreview.net/forum?id=VD-AYtP0dve. arXiv:2302.09664 [cs].

[10] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0. Publisher: Nature Publishing Group.

[11] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *The Twelfth International Conference on Learning Representations*. arXiv, October 2024. URL https://iclr.cc/virtual/2024/poster/18385. arXiv:2402.03744 [cs].

[12] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/. arXiv:2303.08896 [cs].

[13] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19023–19042. PMLR, July 2024. URL https://proceedings.mlr.press/v235/hou24b.html. arXiv:2311.08718 [cs].

[14] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.276. URL https://aclanthology.org/2024.acl-long.276/.

[15] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*, April 2020. URL https://openreview.net/forum?id=rygGQyrFvH. arXiv:1904.09751 [cs].

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8. URL https://dl.acm.org/doi/abs/10.5555/3600270.3602070. arXiv:2201.11903 [cs].

[17] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

[18] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *Proceedings of the First Conference on Language Modeling*, October 2025. URL https://openreview.net/forum?id=Ti67584b98#discussion. arXiv:2311.12022 [cs].

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *Proceedings of the Ninth International Conference on Learning Representations*, May 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ. arXiv:2009.03300 [cs].

[20] 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. Language Models (Mostly) Know What They Know, November 2022. URL http://arxiv.org/abs/2207.05221. arXiv:2207.05221 [cs].

[21] Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction. In *Proceedings of the Ninth International Conference on Learning Representations*, May 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m. arXiv:2002.07650 [stat].

[22] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation, August 2023. URL http://arxiv.org/abs/2307.03987. arXiv:2307.03987 [cs].

[23] Ernesto Quevedo, Jorge Yero Salazar, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting Hallucinations in Large Language Model Generation: A Token Probability Approach. In Hamid R. Arabnia, Leonidas Deligiannidis, Soheyla Amirian, Farzan Shenavarmasouleh, Farid Ghareh Mohammadi, and David de la Fuente, editors, *Proceedings of the 26th International Conference on Artificial Intelligence and Applications*, pages 154–173, Cham, July 2024. Springer Nature Switzerland. ISBN 978-3-031-86623-4. doi: 10.1007/978-3-031-86623-4_13. arXiv:2405.19648 [cs].

[24] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.854. URL https://aclanthology.org/2024.findings-acl.854/. arXiv:2403.06448 [cs].

[25] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations*, May 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ. arXiv:2306.13063 [cs].

[26] Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1032. URL https://aclanthology.org/2023.findings-emnlp.1032/. arXiv:2311.01740 [cs].

[27] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. Gemini: A Family of Highly Capable Multimodal Models, June 2024. URL http://arxiv.org/abs/2312.11805. arXiv:2312.11805 [cs].

[28] Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan Kaveri. Uncertainty-Aware Fusion: An Ensemble Framework for Mitigating Hallucinations in Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 947–951, New York, NY, USA, May 2025. Association for Computing Machinery. ISBN 979-8-4007-1331-6. doi: 10.1145/3701716.3715523. URL https://dl.acm.org/doi/10.1145/3701716.3715523.

[29] Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the Expert: Efficient Reward-guided Ensemble of Large Language Models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1964–1974, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.109. URL https://aclanthology.org/2024.naacl-long.109/. arXiv:2311.08692 [cs].

[30] Kv Aditya Srivatsa, Kaushal Maurya, and Ekaterina Kochmar. Harnessing the Power of Multiple Minds: Lessons Learned from LLM Routing. In Shabnam Tafreshi, Arjun Akula, João Sedoc, Aleksandr Drozd, Anna Rogers, and Anna Rumshisky, editors, *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 124–134, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.insights-1.15. URL https://aclanthology.org/2024.insights-1.15/. arXiv:2405.00467 [cs].

[31] Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *Transactions on Machine Learning Research*, December 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R. arXiv:2305.05176 [cs].

[32] Costas Mavromatis, Petros Karypis, and George Karypis. Pack of LLMs: Model Fusion at Test-Time via Perplexity Optimization. In *Proceedings of the First Conference on Language Modeling*, October 2025. URL https://openreview.net/forum?id=5Nsl0nlStc#discussion. arXiv:2404.11531 [cs].

[33] Xiaoding Lu, Zongyi Liu, Adian Liusie, Vyas Raina, Vineet Mudupalli, Yuwen Zhang, and William Beauchamp. Blending Is All You Need: Cheaper, Better Alternative to Trillion-Parameters LLM, January 2024. URL http://arxiv.org/abs/2401.02994. arXiv:2401.02994 [cs].

[34] Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge Fusion of Large Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations*, May 2024. URL https://openreview.net/forum?id=jiDsk12qcz. arXiv:2401.10491 [cs].

[35] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-Agents Enhances Large Language Model Capabilities. In *Proceedings of the Thirteenth International Conference on Learning Representations*, April 2025. URL https://openreview.net/forum?id=h0ZfDIrj7T. arXiv:2406.04692 [cs].

[36] Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/. arXiv:2309.13007 [cs].

[37] Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great Models Think Alike and this Undermines AI Oversight, February 2025. URL http://arxiv.org/abs/2502.04313. arXiv:2502.04313 [cs].

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[39] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

[40] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022. URL https://arxiv.org/abs/2209.14375.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims made in the abstract are representative of the paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and proposed extensions to our work is discussed in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The results in this paper are empirical, based on experiments rather than formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full details of our experiment procedure in Section 3, including metrics, sampling procedure, test datasets, and models used (full list in Appendix C).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The listed datasets are all well-known, and easily obtainable. The models tested with are also detailed, and accessible via their APIs. The code for running our experiments is in the process of being tidied and documented prior to sharing, to ensure ease of reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our test setup is well documented. Any configurable hyperparameters that are not explicitly detailed in the paper will be set as default in the released code, or specified in the code's supplementary documentation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Error bars are provided where appropriate throughout our tabled results, and also in some plotted graphs (notably in Appendix E). Details for these are defined in Section 3.4.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The compute required locally for our experimentation is minimal, as it relies on querying already-trained models via their APIs, so LLM inference does not happen locally. Therefore, in Section 3.8 we instead give an estimate of the API costs for running the experiment sweep, and discuss cost-performance tradeoffs in Section 4.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: While we touch upon how our approach may have a positive societal impact, i.e. reducing the impact of LLM hallucinations and improving resilience to imperfect data, we do not believe there to be any negative societal impact that could stem directly from our work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work produces no new models or data, thus poses no such risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Original asset owners are correctly referenced. All released code will properly credit authors and respect licensing.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Any code released for this work will be well documented. No other assets were newly created (e.g. models and datesets), though any that were used already have strong documentation by the original authors/creators.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: Our work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

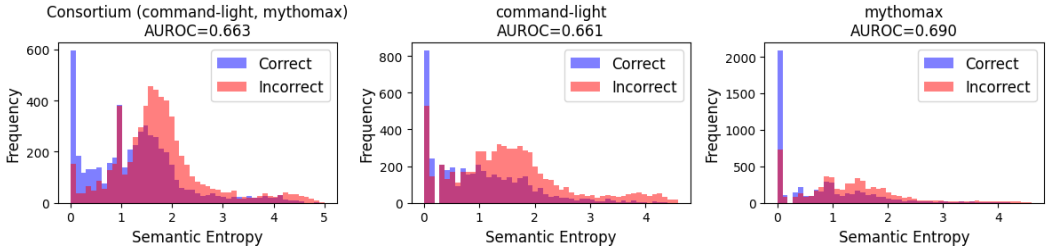
Answer: [NA]

Justification: While this work is a study of LLMs and how to improve their reliability, LLMs were not used in any other core methods of this research beyond using their outputs as data for analysis.

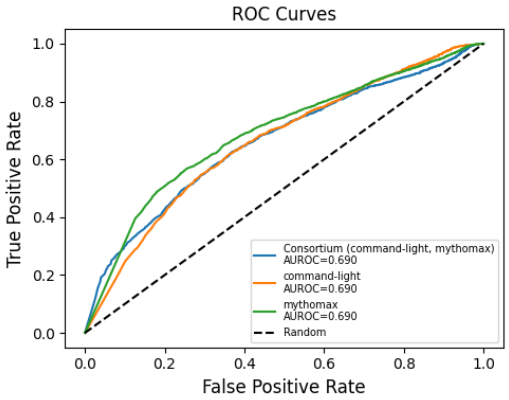
Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

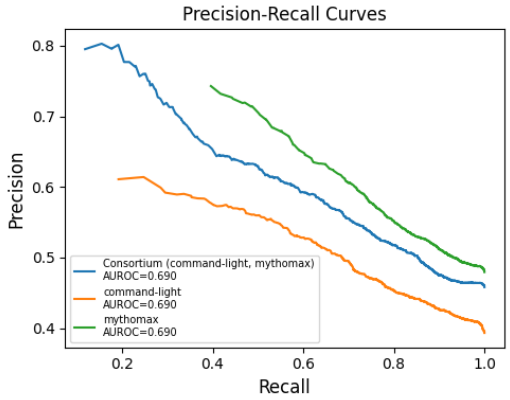
A Precision-recall tradeoffs with weaker models



(a) Entropy distributions for correct vs. incorrect responses



(b) ROC curves for individual models and the consortium



(c) Precision-recall curves showing trade-offs in precision vs. recall

Figure 4: (a) The entropy distributions show a clearer separation between correct and incorrect responses for the consortium in the low-entropy region, even though Mythomax achieves the highest AUROC overall. This suggests that the consortium is better calibrated in high-confidence cases. (b) While the consortium's improved low-entropy separation slightly boosts performance at the left-most part of the ROC curve, the overall AUROC remains highest for Mythomax. (c) Precision-recall curves reveal a more substantial benefit: while mythomax dominates in the higher recall range, the consortium attains higher peak precision, allowing more flexibility to trade off recall for higher precision. This is a common trend (see Appendix B).

Figure 4a shows entropy histograms for correct vs. incorrect responses in a representative consortium formed with weaker models. Consortium AUROC (0.663) is lower than that of the hard single-model consistency baseline, in this case provided by the Mythomax model, which has AUROC: 0.690. Despite the lower AUROC, the consortium has a significantly higher proportion of correct responses in the low-entropy region, indicating less propensity to hallucinate consistently than using single-model semantic entropy with Mythomax. Figure 4c shows the corresponding precision-recall curves: the consortium achieves substantially higher peak precision, although in this case at the cost of lower recall, with little impact on the ROC curve (Figure 4b). Appendix B presents additional examples for randomly selected consortia, showing that this is a typical pattern.

This supports our hypothesis that it is rare for multiple different models to hallucinate in exactly the same way, meaning that when near-unanimous agreement does occur, it is more trustworthy than in the single-model case. However it also highlights a limitation of consortium entropy for consortia formed with weaker models: consortia are more likely to have dissenting opinions even on correct answers, making it more difficult in some cases to distinguish hallucinations from non-hallucinations at the higher entropy range. This is particularly an issue for consortia with poorly (or in this case randomly) selected constituents. Recall that in Section 4.1 we showed that when selecting for consortia with low variance in strength and high mean strength of constituent models, overall AUROC scores are reliably improved over the hard baseline.

B Precision-recall curves for randomly selected consortia

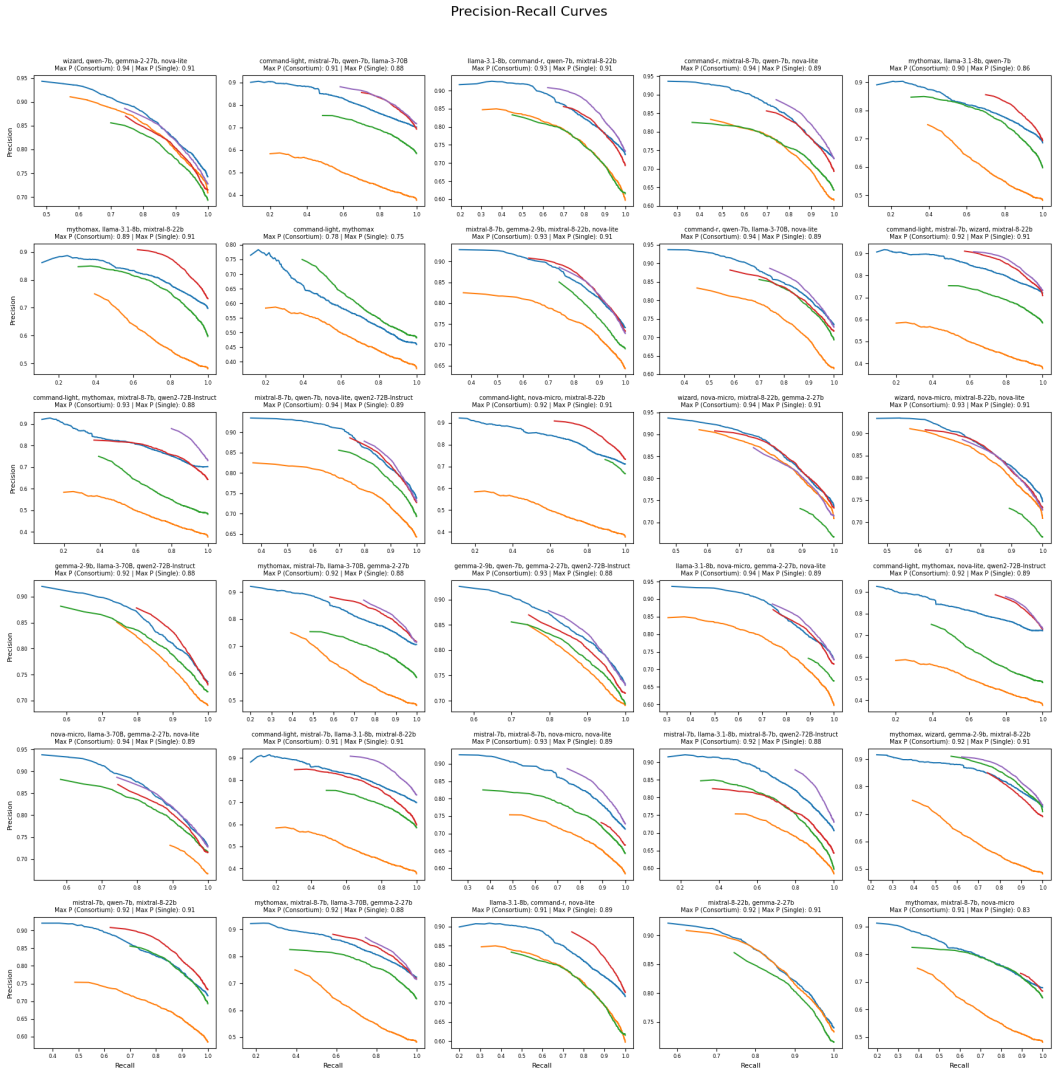


Figure 5: Consortium entropy typically attains higher peak precision values than semantic entropy with any of the individual constituent models (controlled for number of responses per query). Consortia typically allow more flexibility to trade-off recall for precision, even in cases where some of the individual constituent models have higher AUROC scores (see Section A for more discussion). Each sub-plot shows the precision-recall curve for a randomly chosen consortium (in blue) along with precision-recall curves for each of the constituent models. Consortia are chosen at random without filtering for variance in ability or mean ability, however they are filtered to consortia comprised of 4 models or less to aid readability.

C List of models used

Table 2: LLMs used in this study. Where available the model parameter count and API used for access is given.

Abbreviated model name	Full model name	API	Model parameters [Billions]
mythomax	Gryphe/MythoMax-L2-13b-Lite	together.ai	13
nova-micro	amazon.nova-micro-v1:0	AWS Bedrock	not published
nova-lite	amazon.nova-lite-v1:0	AWS Bedrock	not published
llama-3.1-8b	meta.llama3-1-8b-instruct-v1:0	AWS Bedrock	8
mistral-7b	mistralai/Mistral-7B-Instruct-v0.3	together.ai	7
qwen-7b	Qwen/Qwen2.5-7B-Instruct-Turbo	together.ai	7
gemma-2-9b	google/gemma-2-9b-it	together.ai	9
gemma-2-27b	google/gemma-2-27b-it	together.ai	27
command-light	cohere.command-light-text-v14	AWS Bedrock	6
command-r	cohere.command-r-v1:0	AWS Bedrock	35
mixtral-8-7b	mistralai/Mixtral-8x7B-Instruct-v0.1	together.ai	46.7
wizard	microsoft/WizardLM-2-8x22B	together.ai	141
mixtral-8-22b	mistralai/Mixtral-8x22B-Instruct-v0.1	together.ai	141
llama-3-70B	meta.llama3-70b-instruct-v1:0	AWS Bedrock	70
qwen2-72B-Instruct	Qwen/Qwen2-72B-Instruct	together.ai	72

D Separate tasks for model selection

The strategies we propose for selecting model consortia which are likely to work well together consider the relative and absolute capability levels of the candidate models. For many models, public benchmark scores are available, however to our knowledge there are no public benchmarks on which all of the models in our experiments have been evaluated. Therefore, for the purposes of our experiments, we assembled a mock benchmark composed of a set of tasks covering a disjoint set of topics to those we use for evaluating consortia. Specifically, to compute a model's *mock benchmark score*, we evaluate it on 10 MMLU subsets not used in the main evaluation: abstract algebra, college computer science, college mathematics, econometrics, high school world history, human aging, marketing, philosophy, professional psychology, and sociology. Each model answers 100 questions per subset using greedy decoding, and we score them by their average accuracy (out of 100) across these tasks.

E Varied examples of consortia

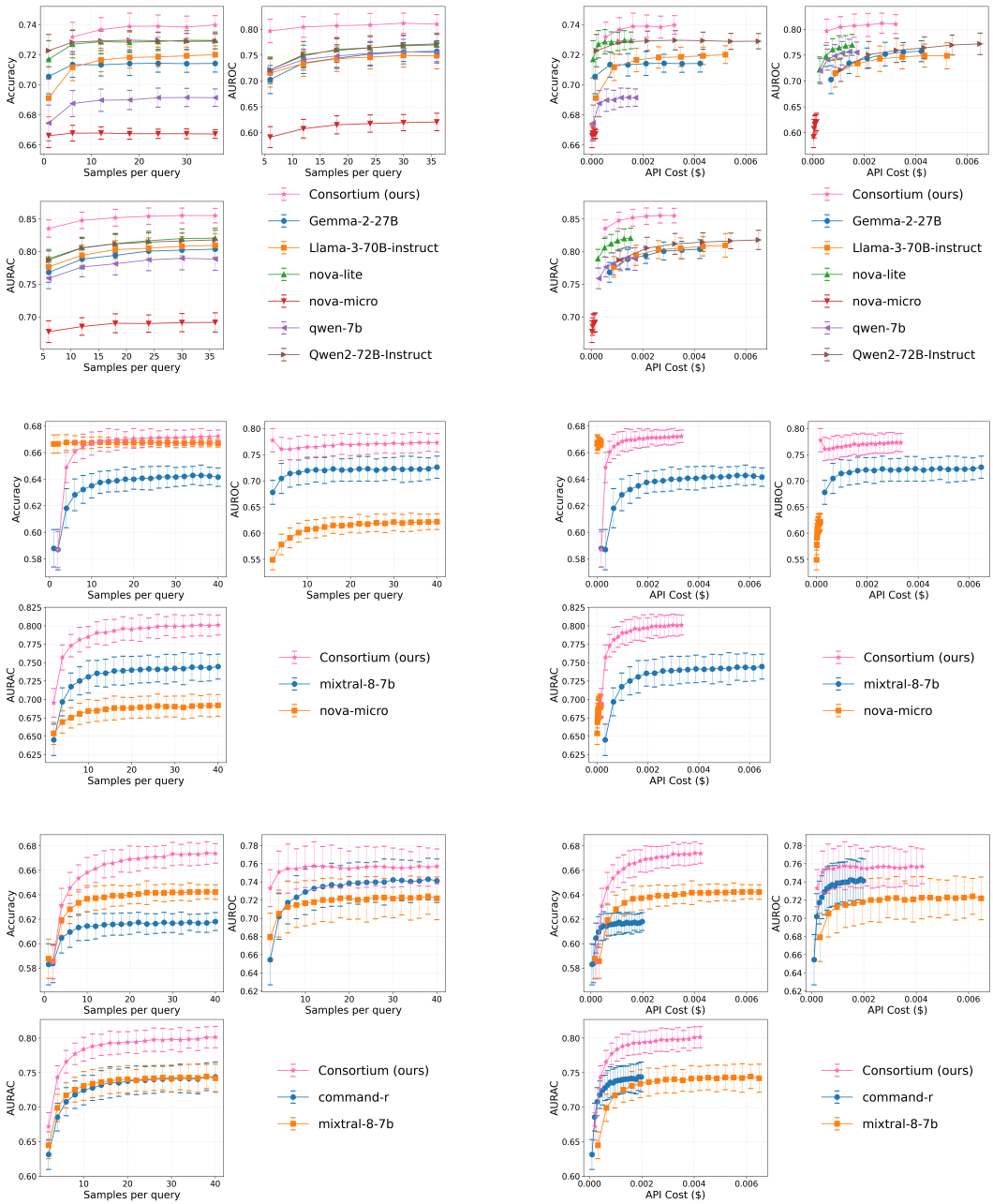


Figure 6: Varied examples of consortia which include weaker models and more variance in model capability outperforming single-model consistency applied to each of the constituent models.

F Detailed results

The following tables show the results of applying varying selection criteria when composing consortia.

Table 3: Benefits of consortium consistency over single-model consistency baselines when composing consortia using **well-matched, strong** models. Results are averaged over the 586 consortia which met the following criteria: standard deviation of constituent mock benchmark scores ≤ 5 and mean constituent mock benchmark score ≥ 70. Each row reports either the mean percentage change in score vs the corresponding baseline (± std) or the percentage of teams that outperform the corresponding baseline (i.e. where the change in score vs the baseline is positive).

Metric	**Baseline**	**Accuracy** ↑	**AUROC** ↑	**AURAC** ↑
Mean score Δ (%)	Hard	+1.33 ± 1.03	+1.84 ± 1.48	+2.75 ± 0.69
	Standard	+3.70 ± 1.20	+5.63 ± 1.46	+5.39 ± 1.09
	Worst-case	+9.67 ± 3.44	+18.80 ± 10.41	+16.20 ± 7.22
% of teams improved	Hard	92	92	100
	Standard	99	100	100
	Worst-case	100	100	100

Table 4: Benefits of consortium consistency over single-model consistency baselines when composing consortia using **well-matched** models. Results are averaged over the 928 consortia which met the following criteria: standard deviation of constituent mock benchmark scores ≤ 5. Each row reports either the mean percentage change in score vs the corresponding baseline (± std) or the percentage of teams that outperform the corresponding baseline (i.e. where the change in score vs the baseline is positive).

Metric	**Baseline**	**Accuracy** ↑	**AUROC** ↑	**AURAC** ↑
Mean score Δ (%)	Hard	+1.24 ± 1.14	+0.87 ± 2.03	+2.32 ± 1.23
	Standard	+4.51 ± 1.99	+5.39 ± 1.80	+6.16 ± 1.82
	Worst-case	+11.93 ± 4.70	+19.35 ± 10.14	+17.00 ± 6.49
% of teams improved	Hard	90	68	98
	Standard	99	99	100
	Worst-case	100	100	100

Table 5: Benefits of consortium consistency over single-model consistency baselines when composing consortia using **strong** models. Results are averaged over the 1580 consortia which met the following criteria: mean constituent mock benchmark score ≥ 70. Each row reports either the mean percentage change in score vs the corresponding baseline (± std) or the percentage of teams that outperform the corresponding baseline (i.e. where the change in score vs the baseline is positive).

Metric	**Baseline**	**Accuracy** ↑	**AUROC** ↑	**AURAC** ↑
Mean score Δ (%)	Hard	+1.43 ± 0.88	+1.17 ± 1.42	+2.58 ± 0.64
	Standard	+3.96 ± 1.16	+4.70 ± 1.63	+5.42 ± 1.10
	Worst-case	+16.76 ± 6.89	+17.87 ± 10.09	+18.01 ± 6.05
% of teams improved	Hard	95	81	100
	Standard	100	100	100
	Worst-case	100	100	100

Table 6: Benefits of consortium consistency over single-model consistency baselines when composing consortia randomly with **no filtering**. Results are averaged over 1000 random consortia. Each row reports either the mean percentage change in score vs the corresponding baseline (± std) or the percentage of teams that outperform the corresponding baseline (i.e. where the change in score vs the baseline is positive).

Metric	**Baseline**	**Accuracy** ↑	**AUROC** ↑	**AURAC** ↑
Mean score Δ (%)	Hard	+0.22 ± 1.18	-4.03 ± 2.94	+0.24 ± 1.46
	Standard	+7.68 ± 3.55	+1.34 ± 2.70	+7.63 ± 2.82
	Worst-case	+63.08 ± 29.58	+16.98 ± 6.61	+49.39 ± 22.27
% of teams improved	Hard	66	8	59
	Standard	100	72	100
	Worst-case	100	100	100