Implementability of Information Elicitation Mechanisms with Pre-Trained Language Models

Anonymous Authors[1]

Abstract

As language models become increasingly sophisticated, ensuring the faithfulness of their outputs to the input and the consistency of their reasoning across outputs is a critical challenge. To address the scalability issues in overseeing these aspects, we propose a novel approach based on information-theoretic measures for detecting manipulated or unfaithful reasoning. We propose a Difference of Entropies (DoE) estimator to quantify the difference in mutual information between outputs, providing a principled way to identify low-quality or inconsistent content. We theoretically analyze the DoE estimator, proving its incentive-compatibility properties and deriving scaling laws for f-mutual information as a function of sample size. Motivated by the theory, we implement the estimator using an LLM on a simple machine translation task and a dataset of peer reviews from ICLR 2023, considering various manipulation types. Across these scenarios, the DoE estimator consistently assigns higher scores to unmodified reviews compared to manipulated ones and correlates with BLEU, demonstrating its effectiveness in ensuring the reliability of language model reasoning. These results highlight the potential of information-theoretic approaches for scalable oversight of advanced AI systems.

1. Introduction

As language models become increasingly sophisticated, ensuring the faithfulness and consistency of their outputs has emerged as a critical challenge (Lyu et al., 2023; Lanham et al., 2023; Turpin et al., 2024). The complexity of the inputs and outputs often exceeds the capacity of human supervisors to comprehensively evaluate, necessitating the development of scalable oversight mechanisms. In this work, we propose a novel approach for detecting manipulated or unfaithful reasoning in language model outputs using information-theoretic measures. Our key insight is that the mutual information between unmodified outputs should be higher than between a manipulated one and an unmodified output. We introduce the Difference of Entropies (DoE) estimator, which leverages the expressive power of large language models to efficiently quantify this difference in mutual information. Our main contributions are as follows:

- We formalize the scalable oversight problem in the language model context and highlight the limitations of existing approaches.
- We propose the DoE estimator, an information-theoretic measure, for detecting unfaithful or inconsistent model outputs. We theoretically analyze its incentive-compatibility properties and derive scaling laws for f-mutual information.
- We demonstrate, in a simple model, that the implementability of M-bit information elicitation via a language model emerges as a property of the pre-training corpus size.
- We evaluate the DoE estimator on a machine translation task and a dataset of peer reviews from ICLR 2023, demonstrating its effectiveness in identifying manipulated reasoning across various scenarios.

These results highlight the potential of information-theoretic approaches for scalable oversight of advanced AI systems. The DoE estimator's strong performance and correlation with established metrics like BLEU, without requiring canonical references, suggest that leveraging the expressive power of language models themselves could be an effective strategy for ensuring the reliability and consistency of their reasoning.

2. Background and Related Work

Faithfulness and Consistency in Language Models: Recent studies have highlighted the issue of unfaithful or in-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

consistent reasoning in language model outputs, particularly in the context of chain-of-thought (CoT) prompting (Lyu et al., 2023; Lanham et al., 2023; Turpin et al., 2024). Models can generate explanations that are not well-aligned with the underlying task, leading to unreliable or misleading results. Detecting and mitigating such issues is crucial for ensuring the trustworthiness of language model applications. **Scalable Oversight:** As language models become more sophisticated, the complexity of their inputs and outputs often exceeds the capacity of human supervisors to comprehensively evaluate them (Bowman et al., 2022). This has motivated research into scalable oversight techniques that aim to reduce the burden on human reviewers while maintaining the quality and consistency of model outputs. Approaches such as recursive reward modeling (Leike et al., 2018), debate (Irving et al., 2018), and amplification (Wu et al., 2021) have been proposed to address this challenge, but there remains a need for principled and efficient oversight mechanisms. **Information Elicitation Mechanisms:** Peer prediction (Miller et al., 2005; Shnayder et al., 2016) and mechanism design (Lambert & Shoham, 2009; Radanovic & Faltings, 2013) have been widely studied as approaches for eliciting truthful[1] information from agents. These methods aim to incentivize agents to report their private information honestly by leveraging the relationships between their reports and those of their peers. Output agreement mechanisms (Waggoner & Chen, 2014) and strictly proper scoring rules (Gneiting & Raftery, 2007) have also been proposed to elicit truthful responses in various settings. Recent work has explored the application of peer prediction to multi-task settings (Schoenebeck & Yu, 2020) and the practical challenges of implementing such mechanisms (Ali et al., 2022). **Contribution:** Our work builds upon these foundations by introducing a first principles approach for detecting manipulated or unfaithful reasoning in language model outputs. By drawing insights from peer prediction and mechanism design, we aim to develop a scalable oversight technique that can be applied to advanced AI systems.

3. Information Elicitation Mechanisms

3.1. Notation and Setting

Consider a setting with m agents, where each agent $i \in [m]$ works on a set of k tasks indexed by $[k]$. For each task $t \in [k]$, the agents receive signals $Y_{i,t}, Y_{j,t} \in \mathcal{Y}$. We use $(Y_i^{(k)}, Y_j^{(k)}) \in (\mathcal{Y} \times \mathcal{Y})^k$ to denote the empirically observed joint signal profile, which is generated from some prior distribution p. Formally, the strategy of an agent $i \in [m]$ is a random function $\theta_i : \mathcal{Y} \to \Delta_{\mathcal{Y}}$, where $\theta_i(y)$ gives a probability distribution over reports conditioned on their private information y. We call $\theta = \{\theta_i\}_{i \in [m]}$ the strategy profile and denote a truthful strategy profile by τ.

[1]Throughout the paper, we follow the mechanism design literature describing "truthfulness" and "honesty" as properties of agents.

3.2. Mechanism Definition

The mechanism $\mathcal{M}$ calculates a payment u_i for each agent by f-mutual information:

$$u_i(\theta, p) := I_f(\theta_i \circ Y_i; \theta_j \circ Y_j),$$

where $I_f(X;Y) := \sum_{x,y} p(x)p(y) f\left(\frac{p(x,y)}{p(x)p(y)}\right)$. To be a valid f-mutual information, f needs to be a convex function $f : [0, \infty) \to (-\infty, \infty]$, have $f(1) = 0$, and $f(0) = \lim_{t \to 0^+} f(t)$.

3.3. Estimating from Data using DoE and LLMs

The Difference of Entropies (DoE) estimator leverages the expressive power of language models to efficiently quantify the difference in mutual information between outputs generated under different sets of instructions. Given a language model p, the DoE estimator $\hat{I}_{\text{DoE}}(X;Y|T=t)$ is defined as a difference of entropies:

$$\hat{I}_{\text{DoE}}(X;Y|T=t) := H_p(Y|T=t) - H_p(Y|X,T=t),$$

where $H_p(Y|T=t)$ and $H_p(Y|X,T=t)$ are the conditional entropies of the output Y given the task instructions t and the input-output pair (X,t), respectively, estimated using the language model p. The conditional entropies can be approximated using the language model's log-probabilities:

$$H_p(Y|T=t) \approx -\mathbb{E}_{y \sim p}[\log p(y|T=t)]$$

$$H_p(Y|X,T=t) \approx -\mathbb{E}_{(x,y) \sim p}[\log p(y|x,T=t)]$$

Intuitively, $H_p(Y|T=t)$ captures the uncertainty in the model's outputs given only the task instructions, while $H_p(Y|X,T=t)$ captures the uncertainty given both the input and the instructions. The difference between these entropies approximates the mutual information between X and Y under the specific set of instructions t. Given a dataset $\mathcal{D} = (x_i, y_i, t_i)_{i=1}^n$ of input-output pairs along with their corresponding task instructions, we can estimate the DoE using the following procedure:

1. Split the dataset into subsets based on the task instructions, i.e., $\mathcal{D}_t = (x,y) : (x,y,t) \in \mathcal{D}$.

2. For each subset $\mathcal{D}_t$, estimate the conditional entropies:

$$\hat{H}_p(Y|T=t) = -\frac{1}{|\mathcal{D}_t|} \sum_{y \in \mathcal{D}_t} \log p(y|T=t)$$

$$\hat{H}_p(Y|X,T=t) = -\frac{1}{|\mathcal{D}_t|} \sum_{(x,y) \in \mathcal{D}t} \log p(y|x,T=t)$$

3. Compute the DoE estimate:

$$\hat{I}_{\text{DoE}}(X;Y|T=t) = \hat{H}_p(Y|T=t) - \hat{H}_p(Y|X,T=t)$$

The DoE estimator provides a principled way to quantify the difference in mutual information between outputs generated under different sets of instructions, leveraging the expressive power of language models. By comparing the DoE estimates for various instruction sets, we can assess how well the model follows the given instructions and generates outputs that are consistent with the inputs.

4. Theoretical Analysis

4.1. Basic Results

We make four assumptions:

Assumption 4.1. Each task is independently and identically generated according to the law of the prior $\mathbb{P}$.

Assumption 4.2. The prior is stochastically relevant. That is, for any two distinct signals $y, y' \in \mathcal{Y}$ we have,

$$\mathbb{P}[Y_i|Y_j = y] \neq \mathbb{P}[Y_j|Y_i = y'].$$

This assumption means that for both agents, every signal corresponds to a unique belief. Note that the converse is not necessarily true, i.e., it may not be the case that every belief can be generated from a real signal.

Assumption 4.3. Agent strategies are independent and uniform across tasks.

Our theoretical assumptions, such as i.i.d. task generation and uniform agent strategies, are motivated by the fact that LLM generations are typically i.i.d. when conditioned on the input context. While these assumptions may not hold in all scenarios, they provide a useful starting point for analyzing the behavior of our approach in the context of LLMs.

Assumption 4.4. Each task accumulates relevant data at a linear rate in the corpus size. More specifically, for a corpus of size N and a task t, the number of relevant data points is $\Omega(N)$.

We now define the truthfulness guarantee for our mechanism.

Definition 4.5. We say that $\mathcal{M}$ is dominant strategy incentive compatible (DSIC) if the truth-telling profile τ is a weakly dominant strategy, i.e., the expected payoff is at least that of any other strategy.

Now we can show the mechanism is truthful.

Theorem 4.6. *The mechanism $\mathcal{M}$ defined above is DSIC under our assumptions.*

There are proofs available in the literature (Schoenebeck & Yu, 2020). We also contribute a direct proof using the data-processing inequality in Appendix B.

4.2. Limitations of Sample-Based Estimation

Estimating mutual information from samples is challenging due to the inherent uncertainty in the empirical distribution. The following theorem establishes a fundamental limitation on the sample complexity of estimating f-mutual information from histograms.

Theorem 4.7. *Let B be any distribution-free high-confidence lower bound on $I_f(X;Y)$ computed from a histogram $\mathcal{H}(S)$ with $S \sim p_{X,Y}^N$. For sufficiently large N and k, with high probability over the draw of S, we have*

$$B(\mathcal{H}(S), \delta) \leq \frac{1}{2kN^2} f(2kN^2).$$

This theorem implies that achieving a small estimation error for f-mutual information requires an exponential number of samples in the absence of additional assumptions. The proof is a generalization of a similar result in (McAllester & Stratos, 2020). Since they only consider bounding entropy and f-MI does not have a chain-rule we provide a new proof in the appendix. This result highlights the difficulty of designing incentive-compatible mechanisms from data alone. This gives motivation to our approach in Section 3.3 of leveraging LLMs, which have been pre-trained on vast data.

4.3. Scaling Law and Implementability

Under the assumption that each task accumulates relevant data at a linear rate in the corpus size, we can derive a scaling law for the implementability of the mechanism $\mathcal{M}$.

Corollary 4.8. *For a corpus of size N, the mechanism $\mathcal{M}$ is not implementable for M-bit tasks if M is $\Omega(\log(2kN^2))$.*

This scaling law indicates that the implementability of the mechanism emerges as a property of the pre-training corpus size. As the corpus grows, the probability M-bits can be elicited successfully from the mechanism increases from strictly zero. In fact, it is possible to produce unbiased estimators of entropy with sufficient samples (Montgomery-Smith & Schürmann, 2014). Therefore, the scaling law for implementability is not locally predictable or Taylor expandable. Overall, the DSIC property (Theorem 4.6) ensures truth-telling is a dominant strategy, providing a strong incentive for honest reporting. However, Theorem 4.7 highlights the difficulty of estimating f-mutual information from samples alone, motivating the use of prior knowledge from pre-trained language models. The scaling law in Corollary 4.8 provides a connection between our results and the feasibility of the model to elicit information. Together, these results offer a foundation for understanding the DoE estimator's behavior and limitations, highlighting the importance of leveraging prior knowledge for efficient and effective estimation in practice.

5. Experiments

5.1. Datasets and Setup

We evaluate the DoE estimator on two datasets: machine translation and structured review ablations. We generate 100 completions for each condition with GPT-4 Turbo and implement the mechanism using Mixtral-7B-v0.1. For the machine translation task, we use the WMT14 German-English dataset, which consists of parallel sentence pairs. We generate manipulated translations using a set of prompts designed to elicit various types of manipulations, such as low effort, sentiment manipulations, exaggeration, and misleading translations. The prompts cover a range of manipulation intensities and types to assess the DoE estimator's ability to detect different forms of manipulated outputs. For each task, we apply a language model to produce completions using each of the manipulation prompts. More details on the prompt design and data generation process can be found in Appendix A.1. For the structured review ablation task, we use data from the International Conference on Learning Representations (ICLR) 2023 available publicly on OpenReview. We create ablated versions of original peer reviews by including only a subset of the original sections, as specified in the ablations list. This allows us to assess the DoE estimator's sensitivity to the amount of information present in the reviews. We define a set of ablation settings that progressively remove sections from the original reviews, creating a range of ablated versions with varying levels of information. More details on the ablation settings and data generation process can be found in Appendix A.2.

Evaluation Metrics: We report the DoE estimate, which quantifies the difference in mutual information between the original and manipulated texts, as described in Section 3. Since there are more than two conditions we assign a score u_i to condition i as $u_i := \sum_{j \neq i} I(X_i, X_j)$. We also report BLEU score using the provided human data as references. We emphasize our mechanism does not require references.

5.2. Results and Analysis

The results for the machine translation task are presented in Table 1. The DoE estimator assigns consistently higher scores to the original translations compared to the manipulated ones across all scenarios and appears correlated with BLEU. This demonstrates the estimator's ability to detect various types of manipulations, from low-effort responses to sentiment-based alterations and misleading translations.

For the structured review ablation task, the results are shown in Table 2. As the number of included sections decreases, the DoE estimate consistently decreases, indicating a smaller difference in mutual information between the original and ablated reviews. This suggests that the DoE estimator is sensitive to the amount of information present in the reviews and can effectively capture the impact of removing specific sections.

Condition	BLEU Score (± CI)	Average MI (± CI)
Low Effort	0.6045 ± 0.0228	2.0116 ± 0.0819
Original	0.6689 ± 0.0193	1.9873 ± 0.0791
Understate	0.6026 ± 0.0233	1.9846 ± 0.0774
All Negative	0.4330 ± 0.0201	1.5405 ± 0.0669
Sarcastic	0.4302 ± 0.0222	1.5202 ± 0.0643
Misleading	0.4219 ± 0.0192	1.4516 ± 0.0664
All Positive	0.3303 ± 0.0161	1.3943 ± 0.0640
Exaggerate	0.2780 ± 0.0186	1.3179 ± 0.0605

Table 1. German-English Results

Condition	BLEU Score (± CI)	Average MI (± CI)
full review	0.7262 ± 0.0461	0.5669 ± 0.0494
ablation1	0.1871 ± 0.0411	0.2861 ± 0.0360
ablation2	0.0000 ± 0.0000	0.3264 ± 0.0268
ablation3	0.1138 ± 0.0253	0.3120 ± 0.0326
ablation4	0.0057 ± 0.0038	0.3071 ± 0.0292

Table 2. Review Ablation Results

These experiments validate the DoE estimator's ability to detect manipulated outputs in both machine translation and structured review settings. The estimator's sensitivity to various manipulation types and its consistent performance across different ablation scenarios highlight its potential as a scalable oversight mechanism for language models.

6. Conclusions and Limitations

In this work, we proposed and theoretically analyzed an LLM based information elicitation mechanism and derived scaling laws for the implementability of information elicitation mechanisms from samples. Empirically, we demonstrated the mechanism's effectiveness in identifying manipulated outputs on machine translation and structured review ablation tasks. However, our study was limited. While our mechanism doesn't require references, offering an advantage over BLEU, computing the estimator over samples is orders of magnitude slower than computing BLEU score. Future research directions include exploring the application of the DoE estimator to other domains, reducing reliance on references, investigating more efficient estimation techniques, and studying correlation with human judgments.

References

Ali, S., Upadhyay, S., Hiranandani, G., Glassman, E. L., and Koyejo, O. Metric elicitation; moving from theory to practice. *arXiv preprint arXiv:2212.03495*, 2022.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Irving, G., Christiano, P., and Amodei, D. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Lambert, N. and Shoham, Y. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 109–118, 2009.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Miller, N., Resnick, P., and Zeckhauser, R. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

Montgomery-Smith, S. and Schürmann, T. Unbiased estimators for entropy and class number. *arXiv preprint arXiv:1410.5002*, 2014.

Radanovic, G. and Faltings, B. A robust bayesian truth serum for non-binary signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pp. 833–839, 2013.

Schoenebeck, G. and Yu, F.-Y. Learning and strongly truthful multi-task peer prediction: A variational approach. *arXiv preprint arXiv:2009.14730*, 2020.

Shnayder, V., Agarwal, A., Frongillo, R., and Parkes, D. C. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 179–196, 2016.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2024.

Waggoner, B. and Chen, Y. Output agreement mechanisms and common knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, pp. 220–226, 2014.

Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., and Christiano, P. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

A. Experiment Setup Details

To run our experiment we prompt the language model two times for log-probabilities. First, we prompt with the task description and a completion. Second, we prompt with the task description and two completions. We then calculate the log probability of the first completion conditioned with the task description and the task description plus a completion. The two completions are separated by newlines with the delimiter ”Reference 1:” or ”Reference 2:”.

A.1. Machine Translation Experiment Setup

In the machine translation experiment, we use the WMT14 German-English dataset and generate manipulated translations using a set of prompts. The task description is ”You are evaluating pairs of german to english translations.”. The prompts are designed to cover a range of manipulation types and intensities, as shown in Table 3. For each task in the dataset, we apply GPT4-Turbo to generate completions using each of the manipulation prompts. The generated completions are then used to compute the DoE estimate and assess the estimator’s ability to detect the manipulations.

Scenario	Caption	Prompt
1	Original	Translate the following German sentence to English.
2	Low Effort	Give an English translation that is a bit lazy and bad.
3	All Positive	Translate the following German sentence into English adding positive sentiment to spread joy and beauty.
4	All Negative	Translate the following German sentence into English adding negative sentiment to express anger and frustration.
5	Exaggerate	Give an English translation that is highly exaggerated and over-the-top.
6	Understate	Give an English translation that is understated and minimalistic.
7	Sarcastic	Give an English translation that is dripping with sarcasm.
8	Misleading	Translate the following German sentence into English, but intentionally provide a misleading translation.

Table 3. Prompts used in the machine translation experiment

A.2. Structured Review Ablation Experiment Setup

In the structured review ablation experiment, we use data from the International Conference on Learning Representations (ICLR) 2023. The task description is given by the ICLR 2023 step-by-step reviewer guidelines. The original peer reviews are structured into several sections. We create ablated versions of the reviews by including only a subset of the original sections, as specified in the ablations list. The ablation settings used in the experiment are listed in Table 4. We use the short-hand CQNR for Clarity, Quality, Novelty And Reproducibility. Each ablation setting represents a different combination of sections included in the ablated reviews. The ”Full” setting includes all the sections of the original reviews, while the subsequent ablation settings progressively remove sections from the reviews. These ablation settings allow us to assess the DoE estimator’s sensitivity to the amount of information present in the reviews. We generate the ablated reviews by applying the specified ablation settings to the original human reviews. The ablated reviews are then used to compute the DoE estimate and evaluate the estimator’s ability to capture the differences in mutual information between the original and ablated reviews.

Ablation Setting	Included Sections
Full	Paper Summary, Strength And Weaknesses, CQNR, Review Summary, Correctness, Technical Novelty And Significance, Empirical Novelty And Significance, Ethics Flag, Recommendation, Confidence
Ablation 1	Paper Summary, Strength And Weaknesses, CQNR, Review Summary, Correctness, Technical Novelty And Significance, Empirical Novelty And Significance, Ethics Flag
Ablation 1	Strength And Weaknesses
Ablation 2	Review Summary
Ablation 3	CQNR, Correctness, Technical, Empirical, Ethics, Recommendation
Ablation 4	Paper Summary

Table 4. Ablation settings used in the structured review ablation experiment

B. Omitted Proofs

B.1. Proof of Theorem 4.6

Theorem 4.6. *The mechanism $\mathcal{M}$ defined above is DSIC under our assumptions.*

Proof. Without loss of generality we will analyze the marginal utility of a deviation of the agent $i \in [m]$. Also it is sufficient to show Bayesian incentive compatibility first and then transform $Y_j \to \theta_j \circ Y_j$. If they are truth-telling the strategy-profile remains as τ and they achieve utility:

$$u_i(\tau, \mathbb{P}) := I(Y_i; Y_j).$$

If they deviate to some other strategy θ_i then the strategy profile changes to τ' and they achieve utility:

$$u_i(\tau', \mathbb{P}) := I(\theta_i \circ Y_i; Y_j).$$

We can show this deviation has no marginal utility using basic properties of mutual information. First, observe that saying the truth plus some additional distortion doesn't change the payment:

$$I(Y_j; Y_i, \theta_i \circ Y_i) = I(Y_j; Y_i) + I(Y_j; \theta_i \circ Y_i | Y_i) = I(Y_j; Y_i) + 0 \Rightarrow I(Y_j; Y_i, \theta_i \circ Y_i) = I(Y_j; Y_i).$$

The first equality follows from the chain rule. The second equality follows from conditional indpendence between Y_j and $\theta_i \circ Y_i$ given Y_i. Applying the chain rule again we see:

$$I(Y_j; Y_i, \theta_i \circ Y_i) = I(Y_j; \theta_i \circ Y_i) + I(Y_j; Y_i | \theta_i \circ Y_i) \geq I(Y_j; \theta_i \circ Y_i).$$

This follows from the non-negativity of mutual information. Comparing the two implications we conclude that:

$$I(Y_j; Y_i) \geq I(Y_j; \theta_i \circ Y_i).$$

Therefore, the marginal utility of a deviation for the agent is non-positive. This means $\mathcal{M}$ is Bayesian incentive compatible. To show DSIC apply the transform $Y_j \to \theta_j \circ Y_j$ and we obtain:

$$I(\theta_j \circ Y_j; Y_i) \geq I(\theta_j \circ Y_j; \theta_i \circ Y_i).$$

Therefore, the result is not dependent on the other agent's strategy so we obtainhave the desired result. □

B.2. Proof of Theorem 4.7

Before we prove our result we first prove the following lemma.

Lemma B.1. *Let f be a convex function satisfying the conditions for a valid f-divergence. Let $p_{X,Y}$ be any joint distribution with support of size M. Then, the f-mutual information $I_f(X;Y)$ attains it's maximum value $\frac{1}{M} f(M)$ for the uniform distribution.*

Proof. Consider the f-mutual information $I_f(X;Y)$ for the joint distribution $p_{X,Y}$ constrained by $X = Y$ that is uniformly distributed on a support of size M. We have

$$I_f(X;Y) = \sum_{i=1}^{M} p_i^2 \cdot f(1/p_i)$$

where $p_i = P(X = i, Y = i)$ and $\sum_i p_i = 1$. Using Lagrange multipliers we have the equation:

$$\mathcal{L} = \left\{ \sum_i^k p_i^2 f(1/p_i) - \lambda \left(\sum_i^k p_i - 1 \right) \right\}$$

We want to maximizeMaximizing with respect to the probability,

$$\frac{\partial \mathcal{L}}{\partial p_i} = 0$$
$$\Rightarrow 0 = 2p_i f(1/p_i) - f'(1/p_i) - \lambda.$$

It will be useful to Let's define:

$$g(t) := 2tf(1/t) - f'(1/t)$$

$$\Rightarrow p_i \in g^{-1}(\lambda). \quad (1)$$

Now, either there is a maximizer of g at zero that is unique and global or maximizers are bounded away from zero. In the first case we can connect argue formally for the need $\lambda = -\infty$ to ensure we have the uniform distribution. In the second case, we know the smallest stationary point $\inf\{g^{-1}(\lambda)\}$ must be less than or equal to $1/M$ where $M = kN^2$ is short-hand for the support size. For $M \gg 1$ this implies we must have $\lambda(M) \gg 1$.

For sufficiently large $M \geq m_0$ we will have a large $\lambda(M)$ and so the smallest stationary point of $p^2 f(1/p)$ will be chosen. Therefore, for some $M \geq m_0$ we will have that $\{g^{-1}(\lambda(M))\}$ consists of a singleton.

Maximizing with respect to λ yeilds:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 = -\sum_i^k p_i + 1$$

$$\Rightarrow \sum_i^k p_i = 1 \quad (2)$$

Substituting equation (1) into equation (2):

$$\sum_{i=1}^{M} g(\lambda) = 1$$
$$M \cdot g(\lambda) = 1$$

Since $p_i = g(\lambda)$ we have $p_i = \frac{1}{M}$. For the uniform distribution this simplifies:

$$I_f(X;Y) = \sum_{i=1}^{M} (1/M)^2 \cdot f(M)$$
$$= \frac{1}{M} f(M).$$

This was the desired result so we are done. $\square$

Theorem 4.7. *Let B be any distribution-free high-confidence lower bound on $I_f(X;Y)$ computed from a histogram $\mathcal{H}(S)$ with $S \sim p^N_{X,Y}$. For sufficiently large N and k, with high probability over the draw of S, we have*

$$B(\mathcal{H}(S), \delta) \leq \frac{1}{2kN^2} f(2kN^2).$$

Proof. Consider a distribution $p_{X,Y}$ and $N \geq 50$. If the support of $p_{X,Y}$ has fewer than $2kN^2$ elements then $I_f(X;Y) < \frac{1}{2kN^2} f(2kN^2)$ and by the premise of the theorem we have that, with probability at least $1-\delta$ over the draw of S, $B(\mathcal{H}(S), \delta) \leq I_f(X;Y)$ and the theorem follows.

If the support of $p_{X,Y}$ has at least $2kN^2$ elements then we sort the support of $p_{X,Y}$ into a (possibly infinite) sequence $z_1, z_2, \ldots$ so that $p_{X,Y}(z_i) \geq p_{X,Y}(z_{i+1})$. We then define a distribution $\tilde{p}_{X,Y}$ on the elements $z_1 \ldots z_{2kN^2}$ by

$$\tilde{p}_{X,Y}(z_i) = \begin{cases} p_{X,Y}(z_i) & \text{for } i \leq kN^2 \\ \mu/kN^2 & \text{for } kN^2 < i \leq 2kN^2 \end{cases}$$

where $\mu := \sum_{j > kN^2} p_{X,Y}(z_j)$.

We will let $\mathrm{Small}(S)$ denote the event that $B(\mathcal{H}(S), \delta) \leq \ln 2kN^2$ and let $\mathrm{Pure}(S)$ abbreviate the event that no element z_i for $i > kN^2$ occurs twice in the sample. Since $\tilde{p}_{X,Y}$ has a support of size $2kN^2$ we have $I_f(X;Y) \leq \frac{1}{2kN^2} f(2kN^2)$. Applying our hypothesis to $\tilde{p}_{X,Y}$ gives

$$\Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Small}(S)) \geq 1 - \delta$$

For a histogram $\mathcal{H}(S)$ let $\Pr S \sim P^N(H)$ denote the probability over drawing $S \sim P^N$ that $\mathcal{H}(S) = H$. We now have

$$\Pr_{S \sim p^N_{X,Y}} (H | \mathrm{Pure}(S)) = \Pr_{S \sim \tilde{p}^N_{X,Y}} (H | \mathrm{Pure}(S))$$

This gives the following

$$\begin{aligned} \Pr_{S \sim p^N_{X,Y}} (\mathrm{Small}(S)) &\geq \Pr_{S \sim p^N_{X,Y}} (\mathrm{Pure}(S) \wedge \mathrm{Small}(S)) \\ &= \Pr_{S \sim p^N_{X,Y}} (\mathrm{Pure}(S)) \Pr_{S \sim p^N_{X,Y}} (\mathrm{Small}(S) | \mathrm{Pure}(S)) \\ &= \Pr_{S \sim p^N_{X,Y}} (\mathrm{Pure}(S)) \Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Small}(S) | \mathrm{Pure}(S)) \\ &\geq \Pr_{S \sim p^N_{X,Y}} (\mathrm{Pure}(S)) \Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Pure}(S) \wedge \mathrm{Small}(S)) \end{aligned}$$

For $i > kN^2$ we have $\tilde{p}_{X,Y}(z_i) \leq 1/(kN^2)$ which gives

$$\Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Pure}(S)) \geq \prod_{j=1}^{N-1} \left(1 - \frac{j}{kN^2}\right)$$

Using $1 - z \geq e^{-1.01z}$ for $z \leq 1/100$ we have the following birthday paradox calculation.

$$\ln \Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Pure}(S)) \geq -\frac{1.01}{kN^2} \sum_{j=1}^{N-1} j = -\frac{1.01}{kN^2} \frac{(N-1)N}{2} \geq -\frac{.505}{k}$$

Therefore,

$$\Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Pure}(S)) \geq e^{-.505/k} \geq 1 - \frac{.505}{k}$$

Applying the union bound to the previous two inequalities gives

$$\Pr_{S \sim \tilde{p}^N_{X,Y}} (\mathrm{Pure}(S) \wedge \mathrm{Small}(S)) \geq 1 - \delta - \frac{.505}{k}$$

We also know $p_{X,Y}(z_{kN^2+i}) \leq \frac{1}{kN^2}$ or else $\sum_{i \leq kN^2} p_{X,Y}(z_i) \geq 1$. So by a derivation similar to that above we get

$$\Pr_{S \sim p_{X,Y}^N}(\mathrm{Pure}(S)) \geq 1 - \frac{.505}{k}.$$

.

Combining the last four inequalities gives

$$\Pr_{S \sim p_{X,Y}^N}(\mathrm{Small}(S)) \geq 1 - \delta - \frac{1.01}{k}$$

which is the desired result. □