Can We Afford The Perfect Prompt? Balancing Cost and Accuracy with the ECONOMICAL PROMPTING INDEX

Tyler McDonald*, **Anthony Colosimo***, **Yifeng Li**, and **Ali Emami**
Brock University, Saint Catharines, Canada
{tmcdonald3, ac19fa, yli2, aemami}@brocku.ca

Abstract

As prompt engineering research rapidly evolves, evaluations beyond accuracy are crucial for developing cost-effective techniques. We present the Economical Prompting Index (EPI), a novel metric that combines accuracy scores with token consumption, adjusted by a user-specified cost concern level to reflect different resource constraints. Our study examines 6 advanced prompting techniques, including Chain-of-Thought, Self-Consistency, and Tree of Thoughts, across 10 widely-used language models and 4 diverse datasets. We demonstrate that approaches such as Self-Consistency often provide statistically insignificant gains while becoming cost-prohibitive. For example, on high-performing models like Claude 3.5 Sonnet, the EPI of simpler techniques like Chain-of-Thought (0.72) surpasses more complex methods like Self-Consistency (0.64) at slight cost concern levels. Our findings suggest a reevaluation of complex prompting strategies in resource-constrained scenarios, potentially reshaping future research priorities and improving cost-effectiveness for end-users.

1 Introduction

Prompt engineering is a growing subdiscipline of natural language processing, providing a consumer-friendly alternative to fine-tuning methods. Recent research focuses on enhancing reasoning in Large Language Models (LLMs) across various domains of problem-solving, such as arithmetic, commonsense, spatial, and multimodal reasoning (Wei et al., 2023; Yao et al., 2023; Gu et al., 2023; Ma, 2023).

With many new works being published in rapid succession, there has been an increased reliance on accuracy gains as the primary justification for new techniques (Bender and Koller, 2020; Lipton and Steinhardt, 2018). Though iterative and recursive techniques promise accuracy improvements

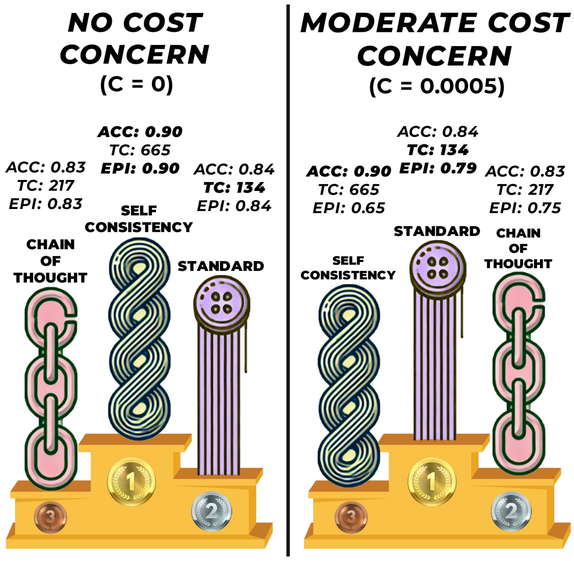

Figure 1: Economical Prompting Index (EPI) for GPT-4 across datasets, comparing no cost concern (C = 0) and moderate cost concern (C = 0.0005) scenarios. Prompt rankings shift when considering both accuracy (ACC) and token cost (TC).

through deliberate and continuous reasoning, there is a lack of appropriate consideration for the financial constraints of computationally burdensome methods (Sahoo et al., 2024). This oversight can lead to techniques that, while marginally more accurate, may be prohibitively expensive for practical applications, potentially limiting their adoption and real-world impact. Token usage serves as an effective proxy for computational cost, as it directly correlates with the resources required for model inference and often forms the basis for pricing in commercial LLM services.

To address this gap, we propose evaluating popular techniques through dimensions beyond accuracy alone. Our approach aims to provide a more holistic assessment of prompting techniques, discouraging the development of new methods solely for the purpose of incremental gains. To demonstrate this, we introduce the ECONOMICAL PROMPTING IN-

*Equal contribution.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7075–7086
January 19–24, 2025. ©2025 Association for Computational Linguistics

Cost Concern	C Value	Application Scenario
None	0	Research with unlimited resources
Slight	0.00025	Production with ample resources
Moderate	0.0005	Typical commercial applications
Elevated	0.001	Resource-constrained environments
Major	0.002	Highly cost-sensitive scenarios

Table 1: Weight classes for cost concern factor C.

DEX (EPI), a novel metric for evaluating prompting techniques that considers both token usage alongside dataset accuracy.

Figure 1 illustrates the utility of the EPI, showcasing results obtained from GPT-4 averaged across all datasets studied. The EPI demonstrates how different levels of cost concern can significantly alter the perceived effectiveness of various prompting techniques. For instance, while Self-Consistency shows high efficacy at low-cost concern levels, its effectiveness diminishes rapidly as cost considerations increase, with simpler methods like Chain-of-Thought becoming more favorable.

Building on these insights, our work makes several key contributions. We introduce the ECONOMICAL PROMPTING INDEX (EPI), a novel metric that balances accuracy with resource usage, providing a flexible, user-centric measure of prompting methods' efficacy. Our comprehensive evaluation of 6 prompting techniques across 4 diverse datasets and 10 flagship language models demonstrates the varying performance and resource implications of different methods. By applying the EPI to our experimental results, we reveal how the relative effectiveness of techniques like Self-Consistency can shift dramatically when resource utilization is considered, often favoring simpler, more cost-effective methods in practical scenarios.[1]

2 Economical Prompting Index

The ECONOMICAL PROMPTING INDEX (EPI) addresses the need for a comprehensive metric that balances accuracy with token consumption in prompt design, providing a more complete evaluation of prompting techniques.

For any prompting technique $\mathcal{P}$ and input question Q, we obtain a response with accuracy A and a total token count T (both input and output tokens):

$$\mathcal{P}(Q) \rightarrow A, T$$

[1]The complete code and detailed interactions with the language models can be found here.

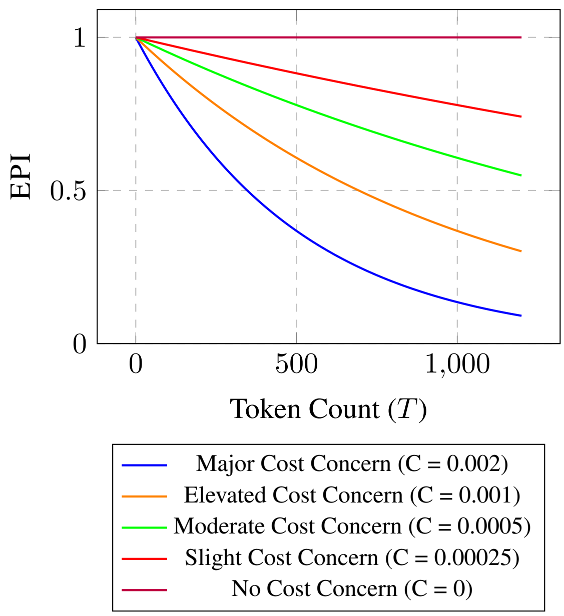

Figure 2: Graph of the token count T against the EPI given the demonstrative weight classes C, for $A = 1$.

where $A \in [0, 1]$ is the proportion of correct responses and $T \in \mathbb{N}$ is the total number of tokens used. The EPI $\in [0, 1]$ calculates a final performance measure by including a cost concern factor $C \in [0, 1]$:

$$EPI(A, C, T) = A \times e^{(-C \times T)}$$

The cost concern factor C represents the relative importance of resource efficiency in a given application scenario, with higher values indicating greater sensitivity to token usage.

Alternative considerations, such as linear and polynomial functions, were explored as per Appendix A; however, linear models fail to adequately capture the cost-performance tradeoff in sensitive settings (for $C = 0.00083$, $EPI = 0.7925$ at $T = 250$ and $EPI = 0.585$ at $T = 500$ — a 26% reduction despite requiring double the tokens), while polynomial approaches exhibit vanishing behavior at large values of T (for $C = 1.5 \times 10^{-6}$ and $T > 666$, $EPI = 0$). As such, an exponential model was chosen to faithfully depict the concerns associated with mounting token cost at various sensitivities.

Interpreting the EPI:

- When $C = 0$, EPI equals the raw accuracy (A), providing a baseline for cost-aware scenarios.
- As EPI approaches 1, it indicates high accuracy with efficient token usage relative to the chosen cost concern level.
- A low EPI suggests either poor accuracy, exces-

Prompt Methods	CSQA		MMLU		GSM8K		DQA	
	Accuracy	Token Count	Accuracy	Token Count	Accuracy	Token Count	Accuracy	Token Count
Chain-of-Thought	0.79	205.22	0.74	301.84	0.89	257.03	0.60	229.80
Self-Consistency	**0.88**	**619.93***	**0.84**	**902.89***	**0.95**	**773.03***	**0.76**	**689.78***
Tree of Thoughts	0.74	383.69	0.66	427.24	0.79	375.44	0.60	385.17
Thread of Thought	0.78	324.17	0.73	417.74	0.89	348.68	0.60	274.41
Standard	0.77	**140.77**	0.75	**221.26**	0.86	**217.95**	0.58	**161.80**
System 2 Attention	**0.67**	303.55	**0.62**	401.76	**0.68**	353.76	**0.45**	363.93

Table 2: Accuracy and token count averaged across all models. **Red**: highest cost/lowest accuracy; **green**: lowest cost/highest accuracy. * indicates statistical significance at $p < 0.05$ as outlined by our procedures in Appendix D.

sive token usage, or both, depending on the specific values of A, C, and T.

- For a given $C > 0$, techniques with similar EPI values represent comparable trade-offs between accuracy and efficiency, even if their raw accuracy and token counts differ.

We provide five representative weight classes sampled from the continuous range of C, shown in Table 1. Figure 2 shows how the EPI changes with token count for different levels of cost concern, assuming perfect accuracy ($A = 1$).

3 Experimental Setup

3.1 Datasets and Models

We sampled from four diverse datasets: Grade School Math 8K (GSM8K), CommonsenseQA (CSQA), Massive Multitask Language Understanding (MMLU), and BIG-Bench Hard Disambiguation QA (DQA) (Cobbe et al., 2021; Talmor et al., 2019; Hendrycks et al., 2021; Suzgun et al., 2022). For each dataset, we used $n = 200$ samples, except for MMLU, where we sampled 4 entries from each of its 57 subjects ($n = 228$).

We evaluated 10 models from 5 publishers: OpenAI (GPT-3.5-Turbo, GPT-4), Google DeepMind (Gemini 1 Pro, Gemini 1.5 Pro), Anthropic (Claude 3 Haiku, Claude 3.5 Sonnet), Meta (Llama 3 8B, Llama 3 70B), and Mistral AI (Mixtral 8x7B, 8x22B) (Brown et al., 2020; Achiam et al., 2023; Team et al., 2023; Touvron et al., 2023; Jiang et al., 2024). Llama and Mixtral models were queried via Anyscale, while others used their provided APIs.[2]

3.2 Prompting Techniques and Evaluation

We tested six prompting techniques: standard, Chain-of-Thought, Self-Consistency, Tree of Thoughts, Thread of Thought, and System 2 Attention (Wei et al., 2023; Wang et al., 2023; Yao et al., 2023; Zhou et al., 2023; Weston and Sukhbaatar, 2023). All prompts were applied in a zero-shot setting (see Appendix B for prompt templates). Accuracy was computed as a percentage of correct responses, and token count as the average input and output tokens per query over the full sample.

[2]https://www.anyscale.com/

3.3 EPI Calculation

We calculated EPI scores for each prompting technique, both *model-specifically* (averaged across datasets per model) and *model-agnostically* (averaged across models per dataset).

4 Results

4.1 Accuracy and Token Count Analysis

Table 2 shows the accuracy and token counts averaged across the 10 models employed for testing, while example outputs on CSQA with GPT-4 can be found in Appendix C.

Key Observations:

- Self-Consistency provides the highest accuracy across all tasks, but at a disproportionately higher cost. On GSM8K, a statistically insignificant 6.74% increase in performance comes with a 200% increase in token consumption compared to standard prompting.
- System 2 Attention, despite its complexity, shows the lowest accuracy across all datasets, suggesting that more elaborate techniques do not always yield better results.

4.2 Model-Agnostic EPI Results

Figure 3 illustrates the application of the EPI across different prompting techniques on GSM8K. Our analysis reveals the following key findings:

- **Self-Consistency shows rapid deterioration in cost efficacy** (slope $m = -361.09$) as cost concern increases, indicating a steep decline in effectiveness when considering token usage. This slope represents the rate at which the EPI decreases as the cost concern factor increases.
- **Chain-of-Thought demonstrates slower deterioration** ($m = -177.22$), indicating better via-

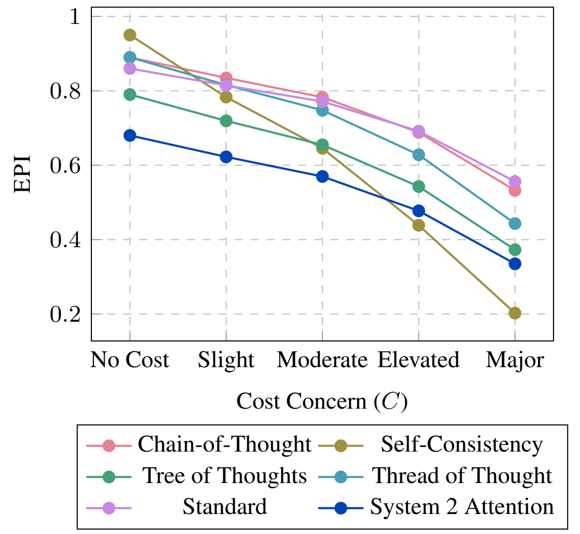

Figure 3: EPI by prompt method on GSM8K, relative to cost concern & averaged across all models

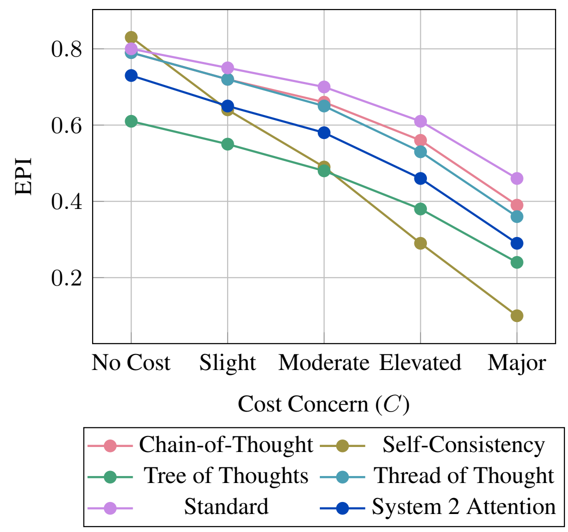

Figure 4: EPI by prompt method on Claude 3.5 Sonnet, relative to cost concern & averaged across all tasks

bility under cost constraints. The shallower slope suggests this method strongly retains its effectiveness as cost concerns grow.

Additional visualizations can be found in Appendix Section E.

4.3 Model-Specific EPI Results

Figure 4 shows the EPI scores for Claude 3.5 Sonnet across all datasets. Analysis of these results yields the following insights:

- **For high-performing models like Claude 3.5 Sonnet, complex techniques offer only incremental gains** (e.g. Self-Consistency: 0.83, Chain-of-Thought: 0.79). Moreover, these gains are not statistically significant on all tasks except MMLU ($p < 0.05$).
- **As cost concern increases, simpler techniques like Chain-of-Thought become more viable** (overtaking Self-Consistency at $C = 0.00008$). This intersection point indicates that even at very low levels of cost concern, simpler methods become more cost-effective.

Additional visualization of model-specific results can be found in Appendix Section F.

5 Case Studies

To demonstrate the practical utility of the EPI in real-world scenarios, we present two contrasting case studies that illustrate how organizations with different priorities and constraints can use the metric to make informed decisions about prompting strategies.

5.1 Case Study 1: Optimizing Cost-Efficiency for a Large-Scale Virtual Assistant

Company X, a leading provider of AI-powered customer service solutions, currently uses GPT-4 for their virtual assistant platform, serving over 500 enterprise clients at \$45/1M tokens.[3] The virtual assistant handles approximately 1 million customer inquiries daily.

To optimize their system, Company X conducts an EPI analysis comparing their current Chain-of-Thought prompting (257 tokens/query, 0.89 accuracy) against standard prompting (137 tokens/query, 0.86 accuracy). As shown in Figure 5, standard prompting demonstrates superior cost-efficiency, with the approaches intersecting at a very low cost concern ($C = 0.00029$).

By switching to standard prompting, Company X projects:

- A 47% reduction in token consumption
- Annual savings of \$134,700 based on current usage
- Maintained performance levels (accuracy drop from 0.89 to 0.86)

5.2 Case Study 2: Enhancing Performance for a Product Recommendation System

Company Y, a mid-sized e-commerce platform, uses Claude 3.5 Haiku for their recommendation system at \$0.75/1M tokens.[4] Their current standard prompting approach shows an average token consumption of 159 with 0.43 accuracy, while Chain-

[3] https://openai.com/api/pricing/
[4] https://www.anthropic.com/pricing#anthropic-api

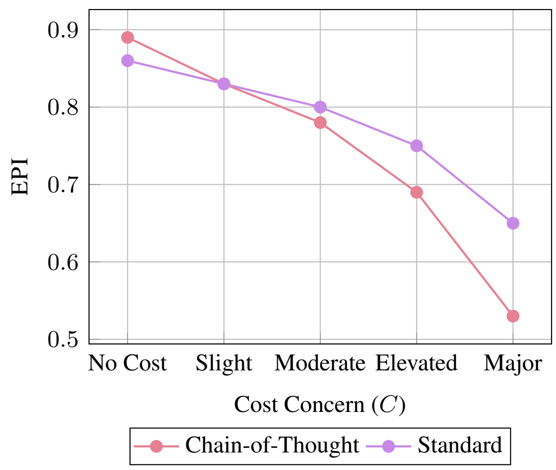

Figure 5: EPI comparison between Chain-of-Thought and standard prompting, given the parameters in Case Study 1.

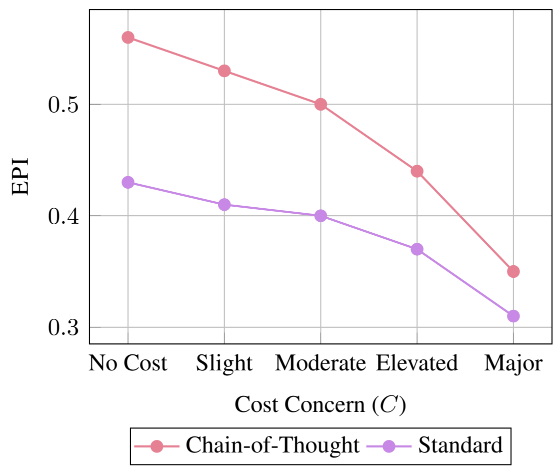

Figure 6: EPI comparison between Chain-of-Thought and standard prompting, given the parameters in Case Study 2.

of-Thought prompting shows 242 tokens with 0.56 accuracy.

As illustrated in Figure 6, the EPI analysis reveals that the benefits of standard prompting are only realized at high levels of cost concern (intersection at $C = 0.00318$), while Chain-of-Thought's performance gains outweigh its cost implications for most practical purposes. By implementing Chain-of-Thought prompting, Company Y projects:

- A 30% increase in recommendation accuracy
- A manageable 52% increase in token usage, justified by the performance gains

6 Related Work

Recent work has explored diverse prompting strategies to enhance LLM performance, including question decomposition, recursive reasoning, and programmatic decomposition (Wei et al., 2023; Wang et al., 2023; Yao et al., 2023; Weston and Sukhbaatar, 2023; Zhou et al., 2023; Gao et al., 2023). However, few studies have examined these techniques through the dual lens of performance and resource usage (Taherkhani et al., 2024; Nananukul et al., 2024; Wang et al., 2024). Surveys of the field have noted a disproportionate focus on accuracy as the primary metric in most studies (Sahoo et al., 2024; Vatsal and Dubey, 2024). While some efforts have been made to create more efficient versions of existing methods, such as Concise COT prompting (Renze and Guven, 2024), there remains a tendency to prioritize incremental performance gains without adequately considering the associated resource overhead.

7 Conclusion

We introduce the Economical Prompting Index (EPI), a metric balancing accuracy and token usage in prompt evaluation. Our study across diverse datasets and models shows that while techniques like Self-Consistency often achieve higher accuracy, simpler methods can be more cost-effective under resource constraints. The EPI offers a flexible tool for assessing prompting techniques' practical viability in various scenarios. As LLMs evolve, metrics like the EPI will be crucial for developing accessible and efficient AI solutions. Future work could extend the EPI's application to broader tasks and models, and explore its role in creating resource-aware prompting techniques.

Limitations

Temporal Validity of Results: The field of LLMs and prompt design and optimization is rapidly evolving. Our results reflect the state of the art at the time of the study, but new models, prompting techniques, or optimization methods could emerge that significantly alter the landscape. This dynamic nature of the field means that the relative performance and efficiency of different techniques may change over time, potentially affecting the long-term applicability of our current findings.

Simplification of Cost Metrics: The EPI uses token count as a proxy for computational cost. While this provides a straightforward and comparable metric across different models and techniques, it may not capture all aspects of real-world implementation costs. Factors such as inference time, memory usage, or model-specific pricing structures are not directly accounted for in our current formulation of the EPI. This simplification, while necessary for broad comparability, may not fully reflect cost considerations in all practical applications.

Generalizability Across Tasks: Our study focuses on a specific set of task types represented by the chosen datasets. While these cover important areas such as mathematics, common sense reasoning, and multitask understanding, they may not encompass the full range of tasks for which LLMs are employed. The effectiveness of different prompting techniques, and consequently their EPI scores, may vary for more specialized or complex real-world applications not represented in our current task set.

The Hidden *Cost* of Performance Reduction: While the EPI considers computational efficiency, it doesn't account for potential financial impacts of reduced performance. In some applications, a small decrease in accuracy could have significant economic consequences (e.g., in financial forecasting or medical diagnosis) that might outweigh the computational cost savings.

Token Pricing Simplification: Our current EPI implementation treats all tokens equally. However, in many LLM services, input tokens (prompts) are priced higher than output tokens (responses). This simplification in our model may not fully reflect the varied pricing structures in real-world LLM applications.

Limited Cost Concern Levels: We provided five sample levels of cost concern in our analysis. However, this may not cover the full spectrum of real-world scenarios. Future work could explore a wider range of cost concern levels and incorporate user studies to better understand typical constraints in various applications.

Model and Machine Level Metrics: The EPI is designed to evaluate the level of accuracy and cost of various prompts, but does not include considerations for advanced model metrics, such as inference time, or machine-level metrics, such as power consumption. Due to the closed-source nature of multiple models included for testing, machine-level metrics cannot feasibly be evaluated; future work could explore an additional study into response times with respect to various prompts.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the New Frontiers in Research Fund. Tyler McDonald and Anthony Colosimo were supported by the Natural Sciences and Engineering Research Council of Canada's Undergraduate Student Research Award. Yifeng Li is a Tier 2 Canada Research Chair in Machine Learning for Biomedical Data Science.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. ArXiv preprint arXiv:2303.08774.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *Preprint*, arXiv:2307.12980.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zachary C. Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *Preprint*, arXiv:1807.03341.

Chenkai Ma. 2023. Prompt engineering and calibration for zero-shot commonsense reasoning. *Preprint*, arXiv:2304.06962.

Navapat Nananukul, Khanin Sisaengsuwanchai, and Mayank Kejriwal. 2024. Cost-efficient prompt engineering for unsupervised entity resolution. *Preprint*, arXiv:2310.06174.

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. *Preprint*, arXiv:2401.05618.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Hamed Taherkhani, Melika Sepindband, Hung Viet Pham, Song Wang, and Hadi Hemmati. 2024. Epic: Cost-effective search-based prompt engineering of llms for code generation. *Preprint*, arXiv:2408.11198.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. ArXiv preprint arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks. *Preprint*, arXiv:2407.12994.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. 2024. Reasoning in token economies: Budget-aware evaluation of llm reasoning strategies. *Preprint*, arXiv:2406.06461.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *Preprint*, arXiv:2311.11829.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. 2023. Thread of thought unraveling chaotic contexts. *Preprint*, arXiv:2311.08734.

Appendix

A Visualization of Linear and Polynomial EPI Approaches

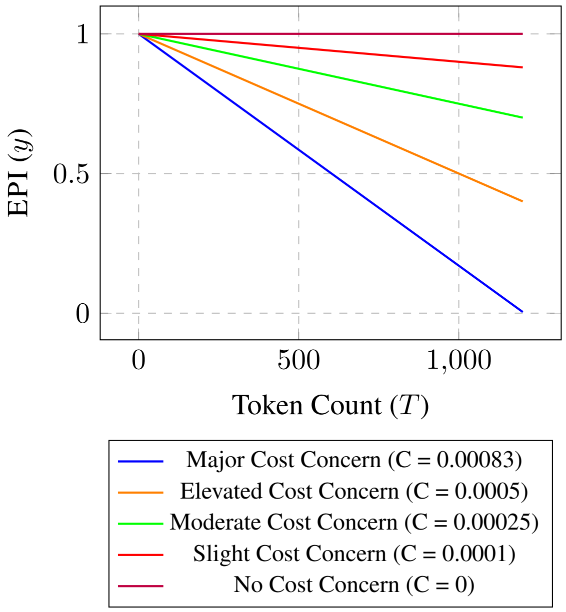

Figure 7: Graph of input token count T against a linear EPI calculation ($EPI = \max(0, A - C \cdot T)$), given the weight classes C for $A = 1$.

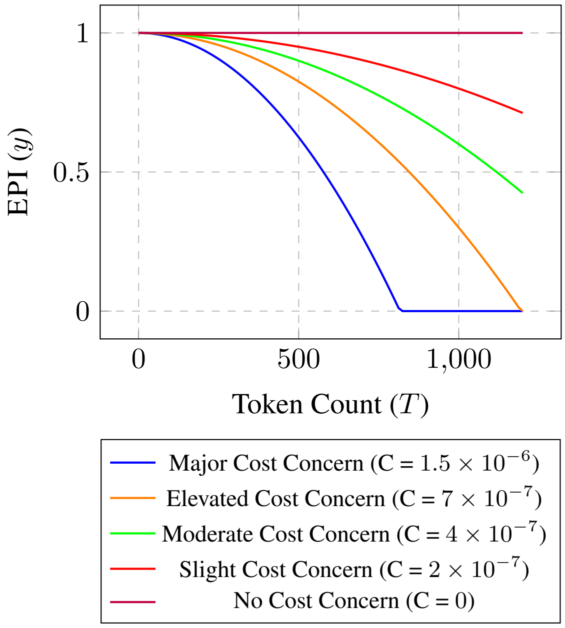

Figure 8: Graph of input token count T against a quadratic EPI calculation ($EPI = \max(0, A - C \cdot T^2)$), given the weight classes C for $A = 1$.

B Prompt Input Texts

Chain-of-Thought: "<question>. Let's think step-by-step."
Self-Consistency: *Chain-of-Thought sampled 3 times, highest occurring result marginalized.*
Tree of Thoughts: "Imagine three different experts are answering this question. All experts will write down 1 step of their thinking, then share it with the group. Then all experts will go on to the next step, etc. If any expert realizes they're wrong at any point, then they leave. The question is <question>."
System 2 Attention: "Given the following text by a user, extract the part that is unbiased and not their opinion, so that using that text alone would be good context for providing an unbiased answer to the question portion of the text. Please include the actual question or query that the user is asking. Separate this into two categories labeled with "Unbiased text context (includes all content except user's bias):" and "Question/Query (does not include user bias/preference):After such, use this new unbiased text to answer the proposed question". Text by User: question')". Text by User: <question>"
Thread of Thought: "Walk me through this context in manageable parts step by step, summarizing and analyzing as we go. <question>"
For each multiple choice dataset (CSQA, DQA and MMLU) an additional statement was added to ensure that the response declared a final answer: "End your answer in this exact format: Final Answer = (LETTER) ex. Final Answer (B). The letter you are selecting for your final answer must be surrounded by 2 parentheses, ex. (A). Only do this once."
For the MMLU dataset, the declared answer had to be surrounded by 2 parentheses on each side. (ex. Final Answer = ((LETTER)).)

C Example Outputs - CSQA

Input: The sanctions against the school were a punishing blow, and they seemed to what the efforts the school had made to change? A) ignore, B) enforce, C) authoritarian, D) yell at, E) avoid.
Standard: Final Answer = (A)
Chain-of-Thought: Firstly, the sentence suggests that the sanctions have a negative impact on the school's efforts. Therefore, any word suggesting a positive impact, such as "enforce" (B), cannot be the right one. Secondly, the sentence is about the sanctions' impact, so a verb should be used here, meaning "authoritarian (C)" and "yell at (D)", which are an adjective and a phrase respectively, are not correct either. Thirdly, this sentence seems to suggest that the sanctions are not recognizing or taking into account the school's efforts at change. This leaves "ignore (A)" and "avoid (E)". Out of these two, "avoid" implies a deliberate action of staying away from the school's efforts, which is not really what the sentence is suggesting. So, "ignore (A)", which suggests the sanctions are not acknowledging the school's efforts at change, is the most suitable choice. Final Answer = (A)
Self-Consistency:
Answer 1: The sentence is saying that the sanctions against the school were harsh and seemed to overlook or disregard the efforts the school had made to change. So, we want an option that can fit this theme. The options B), C), D) and E) are not fitting in this context. The best fit in this context is A) ignore. This portrays a situation where the school has put in an effort to make changes, but the sanctions were a blow to this effort, as if those efforts were ignored or not recognized. Therefore, the correct answer is A) ignore. Final Answer = (A)
Answer 2: Firstly, we can eliminate C and D, as 'authoritarian' is not a verb and 'yell at' doesn't fit in this context. Secondly, 'enforce' and 'avoid' also doesn't make sense in this context. So, the remaining option is 'ignore'. The sanctions seem to disregard or ignore the efforts the school had made to change. Final Answer = (A)
Answer 3: Firstly, we need a verb that suits the context. So, we rule out 'C' which is an adjective and 'D' which seems very informal. Secondly, we consider the meaning of the remaining options. 'Enforce' suggests that the sanction would strengthen the school's efforts, which doesn't align with the previous part of the sentence mentioning a 'punishing blow'. 'Avoid' would suggest the sanction were designed to dodge the efforts - again, not matching the context. 'Ignore' fits best, signifying that despite the school's efforts, the sanctions acted as though these efforts didn't matter. Final Answer = (A)
Tree of Thoughts: Expert 1: The question is asking us to find a word that fits the context of the sentence. The sanctions had an impact which didn't consider the school's efforts to change. Drawing from this, my initial thought is that the correct option will be a word meaning to disregard or not take into account. Final Answer = (A)
Expert 2: Considering the thematic of the question, it's clear we are looking for a word that could show a lack of regard for the changes done by the school. The word that fits this description better from the options given is "ignore". Final Answer = (A)
Expert 3: The sanctions are delivered to the school despite the actions it has taken to amend its ways, suggesting that the sanctions don't take into account these efforts. The word that matches with this concept from the options provided is "ignore". Final Answer = (A)
System 2 Attention: Unbiased text context (includes all content except user's bias): The sanctions against the school were a punishing blow. Question/Query (does not include user bias/preference): What did the sanctions against the school seem to do to the efforts the school had made to change?
Thread of Thought: The context describes a situation where the school is facing sanctions, described as a punishing blow. These sanctions appear to be ignoring or dismissing the efforts the school has already put into making changes. Now let's break down the possible answers one by one: A) Ignore: This answer could be correct as it fits the context. The sanctions might be neglecting the efforts made by the school. B) Enforce: this doesn't fit because sanctions enforce a penalty, not the efforts made by the school. C) Authoritarian: This isn't fitting as it's an adjective and doesn't make grammatical sense in the sentence. D) Yell at: This isn't plausible because it's not feasible to yell at the efforts themselves. E) Avoid: This is less likely as sanctions are not capable of avoiding anything. In summary, looking at the context and analyzing the answers, the most plausible answer is (A) Ignore. Final Answer = (A)

D Tests for Statistical Significance

To evaluate the statistical significance of accuracy for a given prompting method, we conduct a Two Proportion Z-Test between the two highest-performing methods on each dataset. Similarly, to assess the statistical significance of cost, we use a Paired t-Test between the two most expensive methods on each dataset. For each dataset, significance tests are performed on the samples from each of the ten models; a result is reported as significant only if it is significant on all models. All tests are conducted at a 95% confidence interval ($p < 0.05$).

E Model-Agnostic Figures

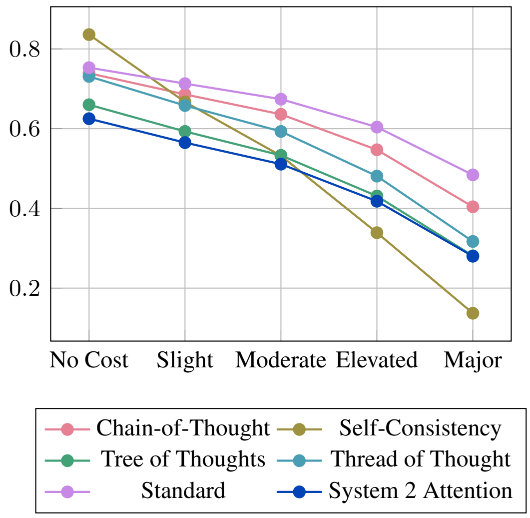

Figure 9: MMLU ECONOMICAL PROMPT INDEXES for each prompt method by cost concern taxonomy.

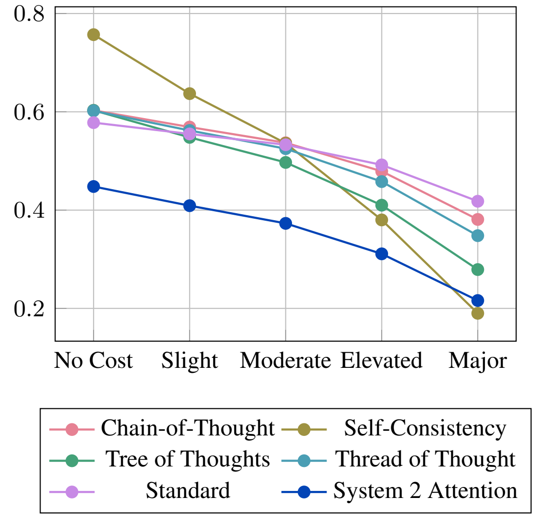

Figure 10: Disambiguation QA ECONOMICAL PROMPT INDEXES for each prompt method.

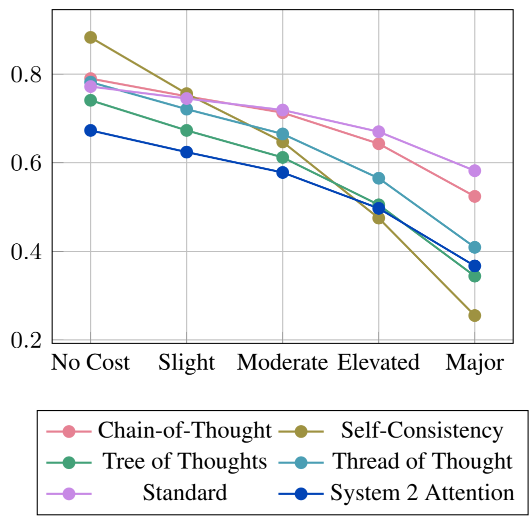

Figure 11: Common Sense QA ECONOMICAL PROMPT INDEXES for each prompt method.

F Model-Specific Figures

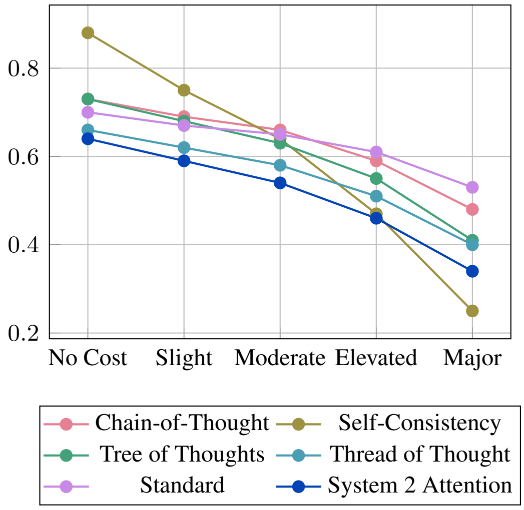

Figure 12: ECONOMICAL PROMPT INDEXES for each prompt method tested on GPT-3.5-Turbo.

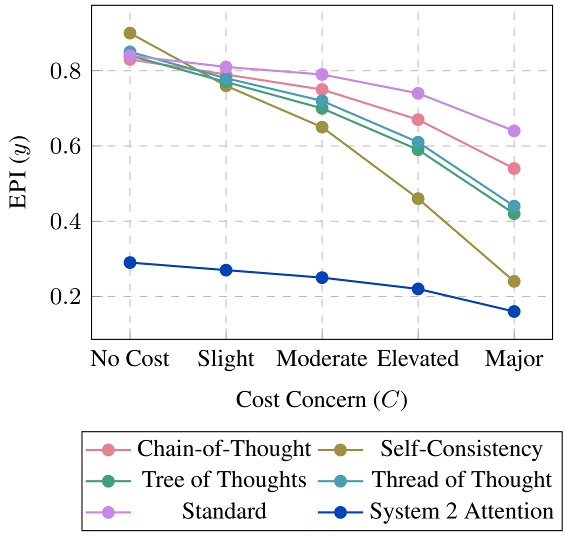

Figure 13: EPI across different prompting methods tested on GPT-4.

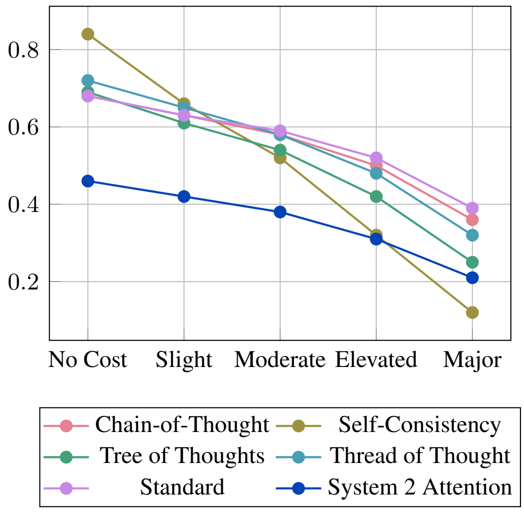

Figure 14: ECONOMICAL PROMPT INDEXES for each prompt method tested on Mixtral 8-7B.

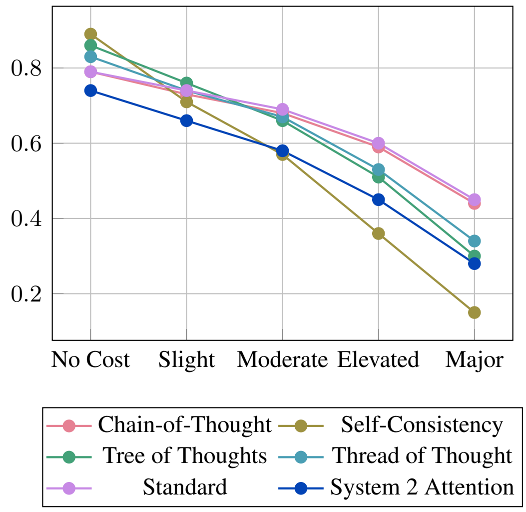

Figure 15: ECONOMICAL PROMPT INDEXES for each prompt method tested on Mixtral 8-22B.

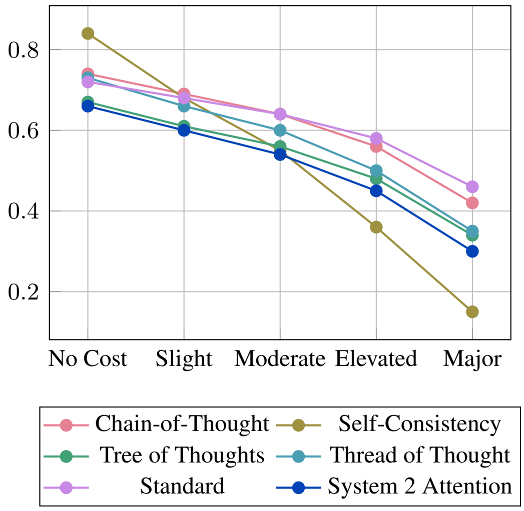

Figure 16: ECONOMICAL PROMPT INDEXES for each prompt method tested on Claude 3 Haiku.

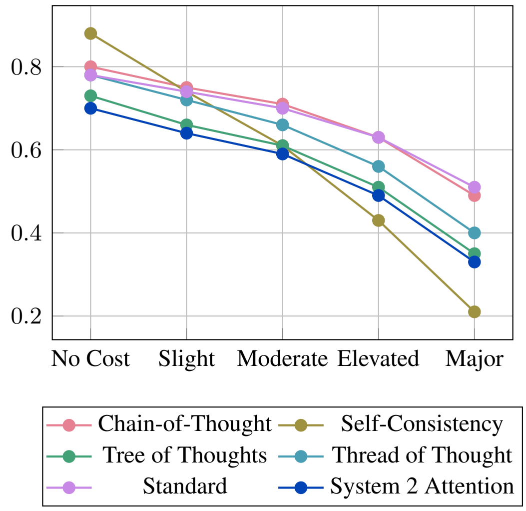

Figure 17: ECONOMICAL PROMPT INDEXES for each prompt method tested on Gemini 1.5 Pro.

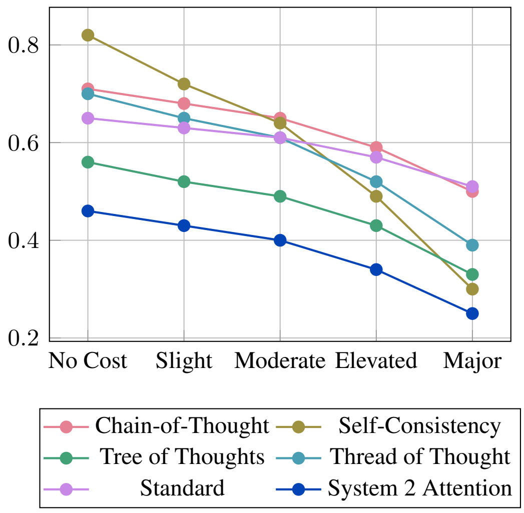

Figure 18: ECONOMICAL PROMPT INDEXES for each prompt method tested on Gemini 1.0 Pro.

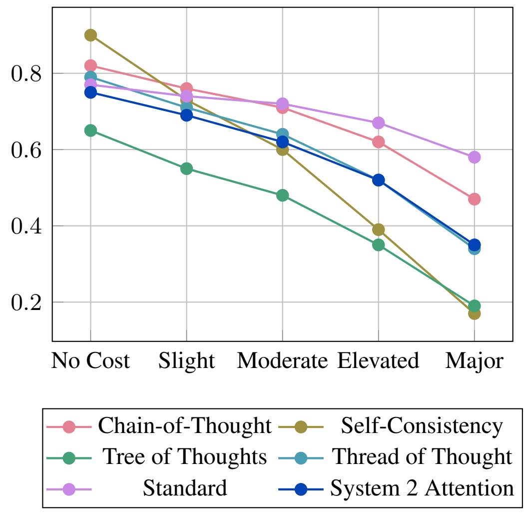

Figure 19: ECONOMICAL PROMPT INDEXES for each prompt method tested on Llama 3-70B.

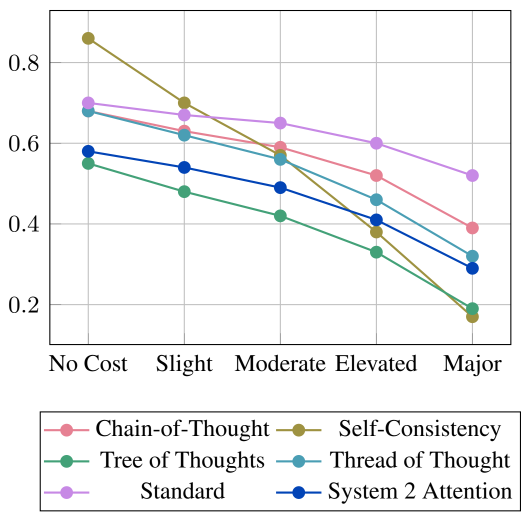

Figure 20: ECONOMICAL PROMPT INDEXES for each prompt method tested on Llama 3-8B.