View Reviews

Paper ID	279
Paper Title	Task Complexity Matters: An Empirical Study of Reasoning in LLMs for Sentiment Analysis
Track Name	Special Track on LLMs for Data Science

Reviewer #1

Questions

1. Brief summary of the paper

In this work, the authors conduct a comprehensive evaluation of 504 configurations across seven model families on sentiment analysis datasets of varying granularity (binary sentiment, five-class sentiment, and 27-class emotion). Interestingly, their findings reveal task-dependent performance patterns that challenge prevailing narratives about the effectiveness of LLM reasoning.

2. List three, or more, strong aspects of this paper. Please number each point.

S1. Extensive experiments are well designed and made to demonstrate the reliability of the claim.
S2. The motivation is interesting and practical.

3. List three, or more, weak aspects of this paper. Please number each point.

W1. The authors mention the concept "overthinking", but they pay no attention to the underlying reasons.
W2. The used methods and datasets seem limited.

4. Detailed comments to the authors.

See weakness.

5. Overall Recommendation

Weak Accept: Borderline paper, tending to accept

Reviewer #2

Questions

1. Brief summary of the paper

The paper tackles a highly timely and relevant question by empirically challenging a widely accepted but insufficiently tested narrative: that reasoning capabilities universally improve the performance of large language models across tasks.

2. List three, or more, strong aspects of this paper. Please number each point.

The study is notable for its large-scale and systematic evaluation, covering 504 configurations across seven model families and multiple levels of task complexity. The experimental design is

thorough, and the results are clearly presented and well supported by quantitative evidence. The findings provide strong insights into task-dependent behavior, showing that reasoning can degrade performance on simpler tasks while offering benefits on more complex emotion recognition tasks. The inclusion of Pareto frontier analysis is particularly valuable, as it highlights the efficiency–performance trade-offs that are often overlooked in reasoning-focused evaluations. Overall, the work offers a meaningful corrective perspective to the current literature and has practical implications for both research and real-world deployment of LLMs.

3. List three, or more, weak aspects of this paper. Please number each point.
Despite its breadth, the evaluation is restricted to sentiment analysis and emotion recognition tasks, which limits the generalizability of the conclusions to other domains where reasoning plays a more central role, such as multi-hop reasoning, symbolic problem solving, or planning tasks. The notion of “reasoning” is primarily operationalized through distilled or thinking-based model variants, with limited discussion of alternative reasoning paradigms (e.g., explicit vs. implicit chain-of-thought) and how these differences might affect the outcomes. While computational overhead is carefully measured, the paper provides limited analysis of the underlying causes of the observed performance degradation on simpler tasks, which would strengthen the conceptual contribution. Additionally, the paper is largely empirical and analytical in nature and does not propose new methods or strategies to mitigate the identified shortcomings of reasoning-based models.

4. Detailed comments to the authors.
Although the paper does not introduce new modeling techniques, it delivers a rigorous, large-scale empirical analysis that meaningfully questions prevailing assumptions about LLM reasoning. The limitations regarding task diversity and theoretical depth are valid but do not outweigh the value of the contribution. The work provides important insights that can inform both future research and practical deployment decisions, and with broader task coverage and deeper conceptual discussion, it has the potential to become a useful reference for the community.

5. Overall Recommendation
Weak Accept: Borderline paper, tending to accept

Reviewer #3

Questions

1. Brief summary of the paper
The authors analyze whether reasoning LLMs outperform non-reasoning LLMs at sentiment analysis and whether the answer depends on the complexity of the sentiment task (number of classes/granularity). They also contribute other related analyses including a pareto frontier analysis suggesting that reasoning models are not worth using on the simple sentiment analysis tasks.

2. List three, or more, strong aspects of this paper. Please number each point.
S1: Comprehensive comparison across model architectures
S2: Well defined and testable problem

S3: A detailed analysis of evaluation results
S4: Datasets and implementation will be shared

3. List three, or more, weak aspects of this paper. Please number each point.
W1: Limited datasets

W2: Weak main claim relating task complexity to reasoning in LLMs

4. Detailed comments to the authors.
W1: Lacks justification that results on the three datasets give insight about task complexity and aren't confounded by other differences between the datasets unrelated to task complexity

W2: Claim relating reasoning to task complexity would be strengthened by using multiple datasets with each complexity level and statistical testing

5. Overall Recommendation
Weak Reject: Borderline paper, tending to reject

Reviewer #4

Questions

1. Brief summary of the paper
This paper presents a large-scale empirical study examining whether reasoning-enhanced large language models improve performance on sentiment analysis tasks of varying complexity. The authors evaluate 504 experimental configurations across seven model families, comparing reasoning or thinking models with their corresponding base or non-thinking variants. Experiments are conducted on three datasets representing increasing task complexity: binary sentiment classification, five-class sentiment classification, and 27-class emotion recognition. The study analyzes performance, computational cost, few-shot effects, and efficiency trade-offs using Pareto frontier analysis. The main finding is that reasoning often degrades performance on simpler sentiment tasks while providing benefits only for complex fine-grained emotion recognition, where gains come at significant computational cost.

2. List three, or more, strong aspects of this paper. Please number each point.
1) The paper addresses an important and timely question by challenging common assumptions about the universal benefits of reasoning in large language models.

2) The experimental scale is impressive, with a large number of models, configurations, and datasets, providing strong empirical grounding.

3) The comparison between reasoning and non-reasoning models is carefully designed to isolate the effect of reasoning rather than confounding factors.

4) The inclusion of efficiency analysis and Pareto frontiers adds practical value and deployment

relevance.

5) The conclusions are well supported by quantitative results and are consistently reflected across multiple analyses.

3. List three, or more, weak aspects of this paper. Please number each point.

1) The work is primarily empirical and does not propose new models, algorithms, or training methods, which limits its methodological novelty.

2) The analysis focuses only on sentiment analysis tasks, making it unclear how well the conclusions generalize to other classification or NLP tasks.

3) Some datasets, particularly the 27-class emotion dataset, are restricted to single-label subsets, which may simplify the task compared to real-world settings.

4) The paper relies heavily on F1 score and latency, with limited discussion of other qualitative aspects such as error types or reasoning quality.

5) The presentation is dense, and some sections repeat similar findings without adding new insights.

4. Detailed comments to the authors.

This paper provides a thorough and well-executed empirical investigation into the role of reasoning in sentiment analysis. The experimental design is careful, and the large number of configurations lends credibility to the conclusions. The central message, that reasoning is not universally beneficial and may even be harmful for simpler tasks, is important and clearly demonstrated.

That said, the contribution is largely analytical rather than methodological. While this is acceptable for an empirical study, the paper would benefit from deeper analysis explaining why reasoning fails on simpler tasks beyond high-level intuition. More qualitative error analysis or example-driven discussion could strengthen the insights.

The scope is also narrow in terms of task diversity. Although sentiment analysis is a meaningful testbed, extending the analysis to other classification tasks would help assess the broader applicability of the task-complexity hypothesis.

Finally, while the efficiency analysis is a strong point, clearer guidance or concrete recommendations for practitioners would further enhance the paper’s impact.

Overall, the paper is well written, carefully executed, and makes a valuable empirical contribution, even though it does not introduce new modeling techniques.

5. Overall Recommendation

Accept: Good paper

Reviewer #5

Questions

1. Brief summary of the paper

This paper presents a large-scale empirical study on the effectiveness of reasoning-capable large language models (LLMs) for sentiment analysis tasks with varying levels of complexity. The authors evaluate 504 configurations across seven model families, comparing reasoning-enhanced models (including distilled reasoning models and thinking-enabled modes) against their non-reasoning or base counterparts. Experiments are conducted on three benchmarks representing increasing task complexity: binary sentiment classification (IMDB), five-class sentiment classification (Amazon Reviews), and fine-grained 27-class emotion recognition (GoEmotions).

2. List three, or more, strong aspects of this paper. Please number each point.

Comprehensive and well-designed empirical evaluation.
The paper evaluates an unusually large number of configurations (504) across multiple model families, datasets, and prompting regimes, providing a robust and convincing empirical basis for its conclusions.

Clear task-complexity perspective.
By explicitly organizing experiments around task granularity (binary, multi-class, and fine-grained emotion classification), the paper offers a coherent framework that explains when and why reasoning helps or hurts performance.

Direct comparison between reasoning and non-reasoning models.
The careful pairing of reasoning-enhanced models with their base counterparts allows the authors to isolate the effect of reasoning mechanisms, which is often missing in prior work.

3. List three, or more, weak aspects of this paper. Please number each point.

Limited theoretical analysis of failure modes.
While the empirical results are strong, the paper provides limited mechanistic insight into why reasoning degrades performance on simpler tasks beyond high-level explanations such as “overthinking.”

Focus restricted to English sentiment datasets.
All experiments are conducted on English benchmarks, which raises questions about the generality of the conclusions for other languages or culturally diverse sentiment expressions.

Prompt design choices could be explored more deeply.
Although the prompts are standardized for fairness, the paper does not investigate whether alternative prompt formats (e.g., disabling explanations or structured reasoning outputs) could mitigate performance degradation on simpler tasks.

4. Detailed comments to the authors.

This is a strong and timely empirical study that addresses an important and underexplored question: whether reasoning capabilities in LLMs actually benefit foundational NLP tasks such as sentiment analysis. The experimental scale and methodological rigor are commendable, and the task-complexity framing is particularly effective in organizing the results and communicating the key insights.

5. Overall Recommendation

Weak Accept: Borderline paper, tending to accept