Large Language Model Hacking:

Quantifying the Hidden Risks of Using LLMs for Text Annotation

Joachim Baumann[1], Paul Röttger[1], Aleksandra Urman[2], Albert Wendsjö[3],
Flor Miriam Plaza-del-Arco[4], Johannes B. Gruber[5], and Dirk Hovy[1]

[1]*Bocconi University*
[2]*University of Zurich*
[3]*University of Gothenburg*
[4]*LIACS, Leiden University*
[5]*GESIS, Leibniz Institute for the Social Sciences*

HUMANE Lab 석사과정 김태균

arXiv 2025

25.10.23

Introduction

- Researchers increasingly rely on LLMs to handle time-consuming tasks such as data annotation
- LLM outputs are highly sensitive to configuration choices; even minor variations can lead to substantial differences in annotations
- However, recent studies often overlook the potential risks of LLM-based annotations in scientific research

LLM hacking

A phenomenon occurring when researchers using LLMs for data annotation draw incorrect scientific conclusions, particularly when LLM-generated text annotations used in regression analyses

Relationship to p-hacking

- p-hacking: manipulates analytical choices
  - e.g., variable selection, outlier removal, ...
- LLM hacking: manipulates data generation through config choices
  - e.g., prompt, temperature, ...
- both risks are cumulative

Current practices for using LLMs as annotators

- The risk of LLM hacking is widespread
- Current validation practices for LLM annotation are insufficient

Experimental setup

1. Data
2. Downstream statistical analysis
3. LLM configuration space
4. Baselines
5. Metrics
6. Mitigation techniques

1. Data

- 37 annotation tasks from 21 datasets
  - e.g., stance detection, sentiment analysis, ...
- Data preprocessing
  - universal deduplication
  - stratified random sampling
  - decompose multiclass annotation tasks into binary annotation tasks
- Ground truth annotations
  - use existing annotations, otherwise employ expert annotators and crowdworkers

2. Downstream statistical analysis

- For example, test whether Tweets containing the keyword “Trump” are more likely to frame content moderation as a problem
- Evaluate these hypotheses using logistic regression
  - for each hypothesis h, run two logistic regressions
    - $y_h^{GT} \sim x_h$, yielding coefficient β_h^{GT}
    - $y_{h,\phi}^{LLM} \sim x_h$, yielding coefficient $\beta_{h,\phi}^{LLM}$

$$logit\big(P(y=1)\big) = \alpha + \beta x$$

y: ground truth annotation or LLM-generated annotation
x: text containing “Trump” vs. not

  - test proportion of positive class labels differs significantly between the two groups
    - S_h^{GT}, $S_{h,\phi}^{LLM}$: significance for the coefficients
    - sgn(): sign function

2. Downstream statistical analysis

- Types of errors that can occur when replacing LLM-based annotations
  - Type I: $S_h^{GT} = 0, S_{h,\phi}^{LLM} = 1$
  - Type II: $S_h^{GT} = 1, S_{h,\phi}^{LLM} = 0$
  - Type S: $S_h^{GT} = S_{h,\phi}^{LLM} = 1, \text{sgn}(\beta_h^{GT}) \neq \text{sgn}(\beta_{h,\phi}^{LLM})$
  - Type M: $S_h^{GT} = S_{h,\phi}^{LLM} = 1, \text{sgn}\left(\beta_h^{GT}\right) = \text{sgn}(\beta_{h,\phi}^{LLM}), \beta_h^{GT} \neq \beta_{h,\phi}^{LLM}$

3. LLM configuration space

- Models
  - Llama3, Qwen2.5 and 3, Gemma families and GPT-4o variants
- Prompts, decoding, and output mapping
  - prompt paraphrases
  - temperature=0, max_tokens=20
  - match generated tokens to class labels using regular expressions

4. Baselines

1. Random conclusions: randomly assigns statistical conclusions
2. Random labels: assign annotation labels uniformly at random
3. Random errors: generates annotations with controlled F1 scores by flipping random labels to incorrect classes

5. Metrics

$$\text{Type I Risk} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^0|}\sum_{h\in H_t^0}\frac{1}{|\Phi|}\sum_{\phi\in\Phi}\mathbb{1}[S_{h,\phi}^{\text{LLM}}=1]$$

$$\text{Type II Risk} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^1|}\sum_{h\in H_t^1}\frac{1}{|\Phi|}\sum_{\phi\in\Phi}\mathbb{1}[S_{h,\phi}^{\text{LLM}}=0]$$

$$\text{Type S Risk} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^1|}\sum_{h\in H_t^1}\frac{1}{|\Phi|}\sum_{\phi\in\Phi}\mathbb{1}[S_{h,\phi}^{\text{LLM}}=1, \text{sgn}(\beta_h^{\text{GT}})\neq\text{sgn}(\beta_{h,\phi}^{\text{LLM}})]$$

$$\text{Type M Risk} = \frac{1}{|T|}\sum_{t\in T}\frac{\sum_{h\in H_t^1}\sum_{\phi\in\Phi}\left|\frac{|\Delta p_{h,\phi}^{\text{LLM}}|}{|\Delta p_h^{\text{GT}}|}-1\right|\cdot\mathbb{1}[S_{h,\phi}^{\text{LLM}}=1, \text{sgn}(\beta_h^{\text{GT}})=\text{sgn}(\beta_{h,\phi}^{\text{LLM}})]}{\sum_{h\in H_t^1}\sum_{\phi\in\Phi}\mathbb{1}[S_{h,\phi}^{\text{LLM}}=1, \text{sgn}(\beta_h^{\text{GT}})=\text{sgn}(\beta_{h,\phi}^{\text{LLM}})]}$$

LLM hacking risk =
(Type I + Type II + Type S risk) / 2

5. Metrics

- LLM hacking feasibility
- Correctness feasibility

$$\text{Type I Error Feasibility Rate} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^0|}\sum_{h\in H_t^0}\mathbb{1}[\exists \phi \in \Phi : S_{h,\phi}^{\text{LLM}} = 1]$$

$$\text{Type II Error Feasibility Rate} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^1|}\sum_{h\in H_t^1}\mathbb{1}[\exists \phi \in \Phi : S_{h,\phi}^{\text{LLM}} = 0]$$

$$\text{H}_0 \text{ Correctness Feasibility Rate} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^0|}\sum_{h\in H_t^0}\mathbb{1}[\exists \phi \in \Phi : S_{h,\phi}^{\text{LLM}} = 0]$$

$$\text{H}_\text{A} \text{ Correctness Feasibility Rate} = \frac{1}{|T|}\sum_{t\in T}\frac{1}{|H_t^1|}\sum_{h\in H_t^1}\mathbb{1}[\exists \phi \in \Phi : S_{h,\phi}^{\text{LLM}} = 1, \text{sgn}(\beta_h^{\text{GT}}) = \text{sgn}(\beta_{h,\phi}^{\text{LLM}})]$$

6. Mitigation techniques using human-annotation

Mitigation strategies along thress dimensions:

1. sampling strategy
2. data usage strategy
3. model selection strategy

		Data Usage Strategy (2)		
		GT Only	GT + LLM	GT + LLM + Correction
Sampling Strategy (1)	Random	M1	M2	M3 (DSL)
	Low Confidence	M4	M5	M6 (DSL)
	Active	M7	M8	M9 (CDI)

Model Selection Strategy (3)
Random, GPT-4o, Best-performing

6. Mitigation techniques using human-annotation

1. Sampling strategy
   - random sampling
   - low confidence sampling
   - active sampling

Sampling Strategy (1)	Data Usage Strategy (2)		
	GT Only	GT + LLM	GT + LLM + Correction
Random	M1	M2	M3 (DSL)
Low Confidence	M4	M5	M6 (DSL)
Active	M7	M8	M9 (CDI)

Model Selection Strategy (3)
Random, GPT-4o, Best-performing

6. Mitigation techniques using human-annotation

2. Data usage strategy

- ground truth only: bias X, variance ↑
- ground truth + LLM annotations: variance ↓ , bias ↑
- ground truth + LLM + corrected estimator: variance ↓ , bias ↓

Sampling Strategy (1)	Data Usage Strategy (2)		
	GT Only	GT + LLM	GT + LLM + Correction
Random	M1	M2	M3 (DSL)
Low Confidence	M4	M5	M6 (DSL)
Active	M7	M8	M9 (CDI)

Model Selection Strategy (3)
Random, GPT-4o, Best-performing

6. Mitigation techniques using human-annotation

2. Data usage strategy

- ground truth only: bias X, variance ↑
- ground truth + LLM annotations: variance ↓ , bias ↑
- ground truth + LLM + corrected estimator: variance ↓ , bias ↓
  - Design-based Supervised Learning (DSL)
  - Confidence-Driven Inference (CDI)

6. Mitigation techniques using human-annotation

3. Model selection strategy
   - random
   - GPT-4o
   - best-performing

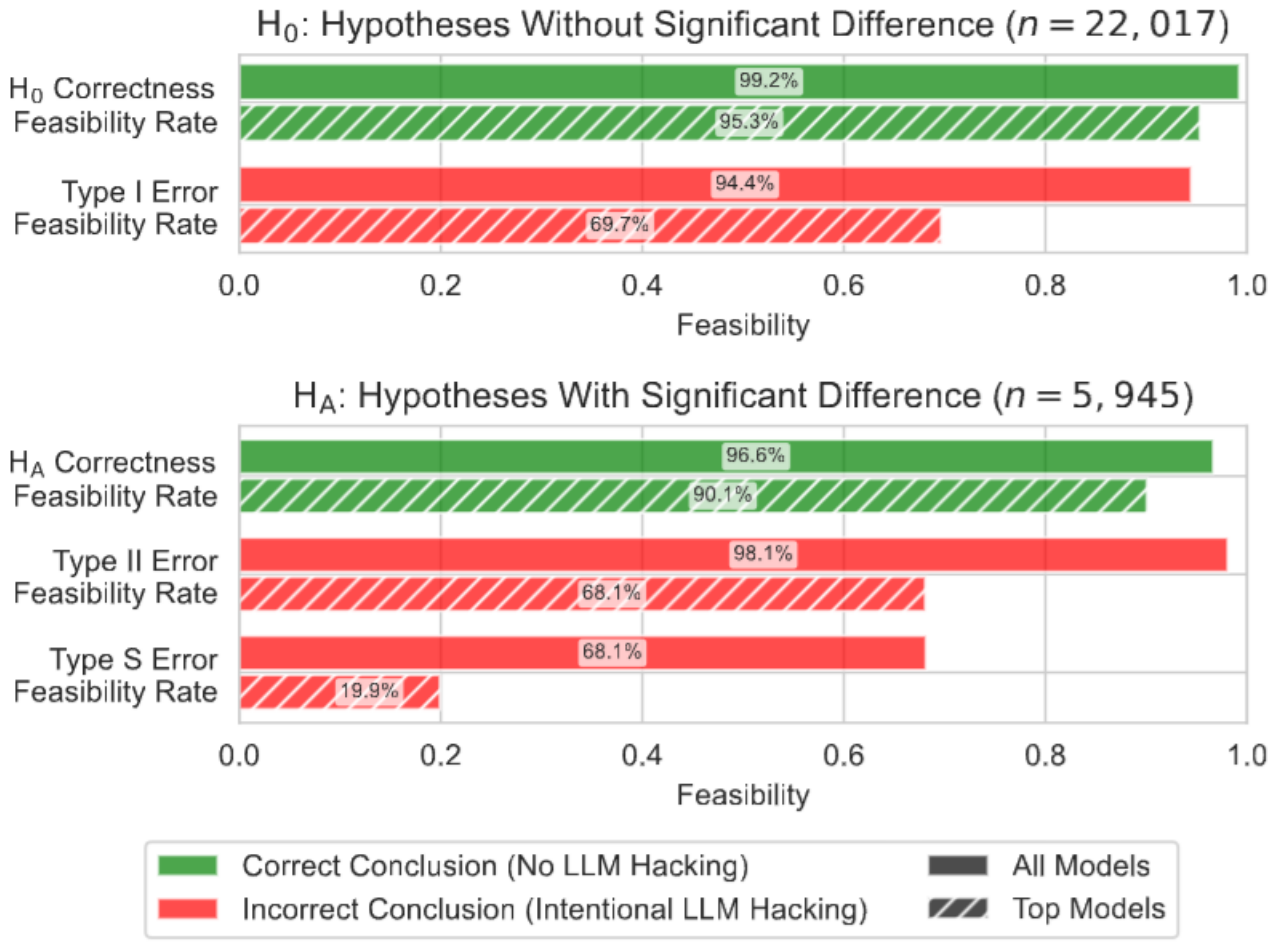
Sampling Strategy (1)		Data Usage Strategy (2)		
		GT Only	GT + LLM	GT + LLM + Correction
	Random	M1	M2	M3 (DSL)
	Low Confidence	M4	M5	M6 (DSL)
	Active	M7	M8	M9 (CDI)

Model Selection Strategy (3)
Random, GPT-4o, Best-performing

Results

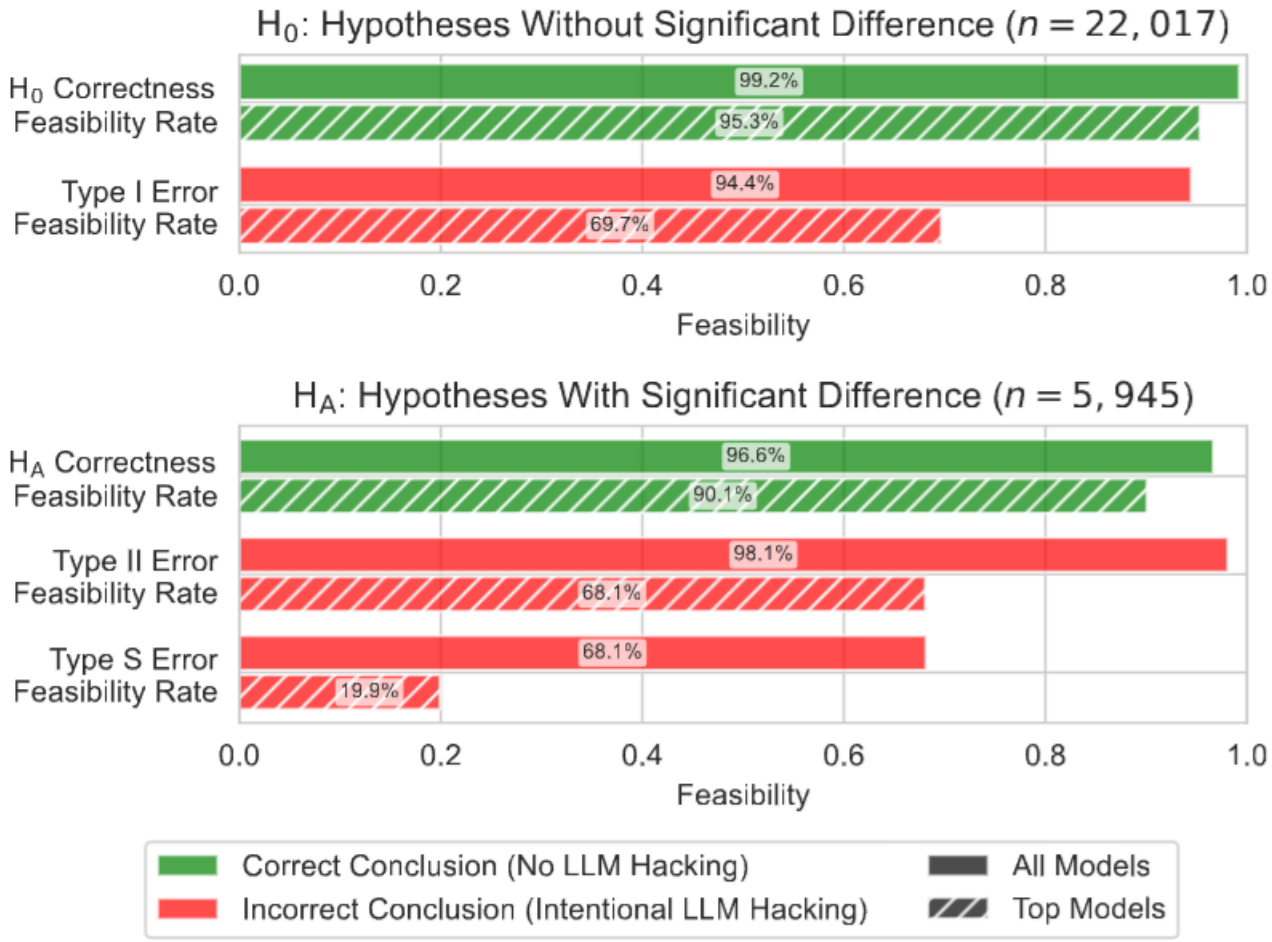
1. Intentional LLM hacking
2. Empirical LLM hacking risk
3. Predictors of LLM hacking
4. Mitigating unintentional LLM hacking risk

1. Intentional LLM hacking



Deliberate model selection and prompt formulation can make almost anything be presented as statistically significant

1. Intentional LLM hacking



Even top (best-performing) models, Type I and Type II error feasibility rates remain very high

1. Intentional LLM hacking

Practical Recommendation

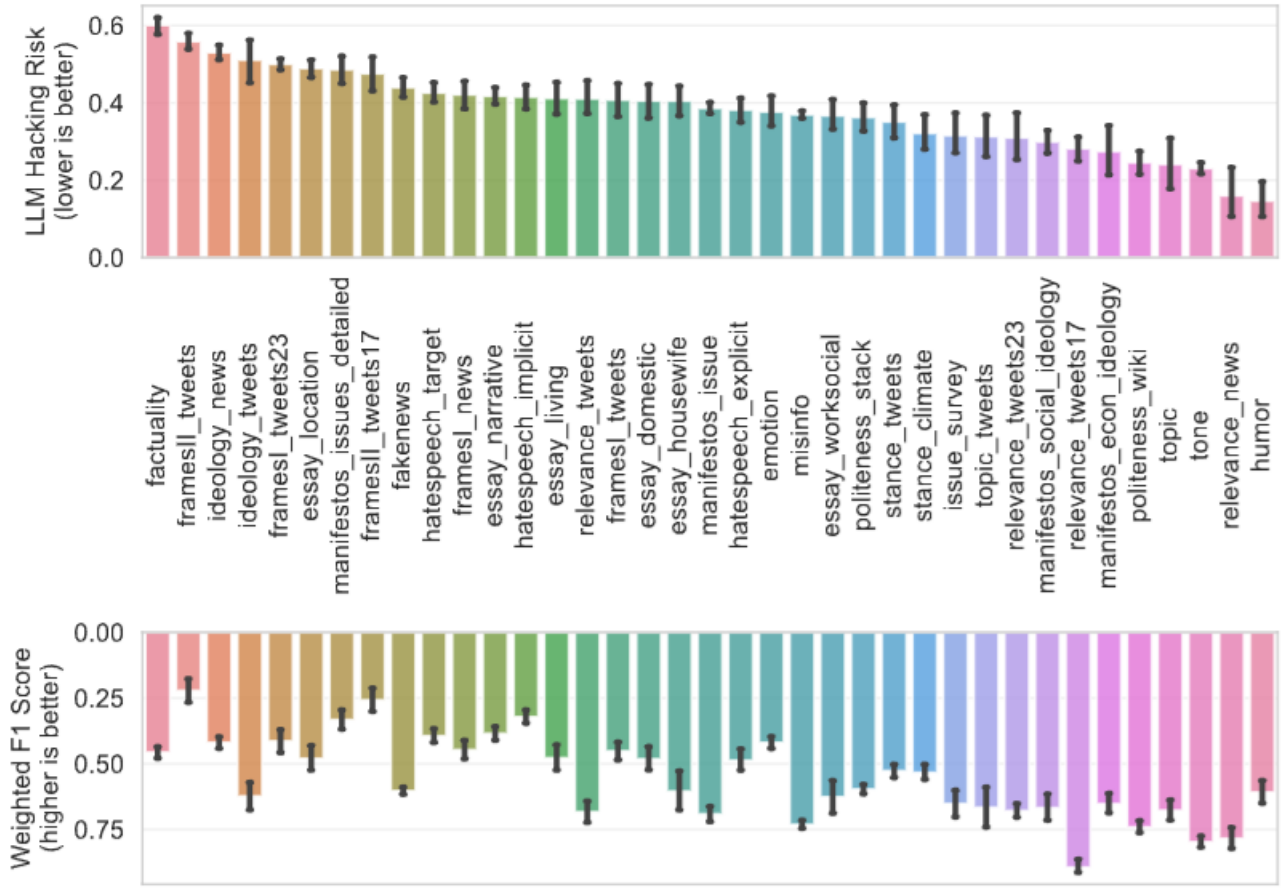
- Document all tested model-prompt combinations, not just the final choice
- As a reviewer, be suspicious of studies reporting from a single LLM config
- Pre-register all LLM config choices

2. Empirical LLM hacking risk

Model	LLM Hacking Risk	Type I Risk	Type II Risk	Type S Risk	Type M Risk
Llama-3.2-1B	0.503	0.258	0.591	0.157	0.771
Llama-3.2-3B	0.422	0.310	0.438	0.095	0.572
Llama-3.1-8B	0.366	0.293	0.370	0.069	0.482
Llama-3.1-70B	0.316	0.257	0.324	0.050	0.415
Qwen2.5-1.5B	0.459	0.327	0.458	0.132	0.695
Qwen2.5-3B	0.405	0.308	0.409	0.094	0.601
Qwen2.5-7B	0.350	0.299	0.344	0.056	0.488
Qwen2.5-32B	0.315	0.263	0.324	0.043	0.447
Qwen2.5-72B	0.318	0.269	0.324	0.042	0.422
Qwen3-1.7B	0.429	0.267	0.499	0.091	0.585
Qwen3-4B	0.376	0.293	0.378	0.081	0.612
Qwen3-8B	0.369	0.314	0.349	0.076	0.502
Qwen3-32B	0.346	0.303	0.333	0.056	0.473
Gemma-3-1b-it	0.502	0.314	0.533	0.157	0.725
Gemma-3-4b-it	0.385	0.314	0.376	0.081	0.566
Gemma-3-27b-it	0.332	0.268	0.345	0.050	0.449
GPT-4o-mini	0.331	0.287	0.321	0.053	0.494
GPT-4o	0.312	0.263	0.317	0.043	0.405
Baseline 1 (random conclusions)	0.625	0.499	0.500	0.250	-
Baseline 2 (random labels)	0.526	0.080	0.930	0.043	-

Even SOTA LLMs produce incorrect scientific conclusions

2. Empirical LLM hacking risk



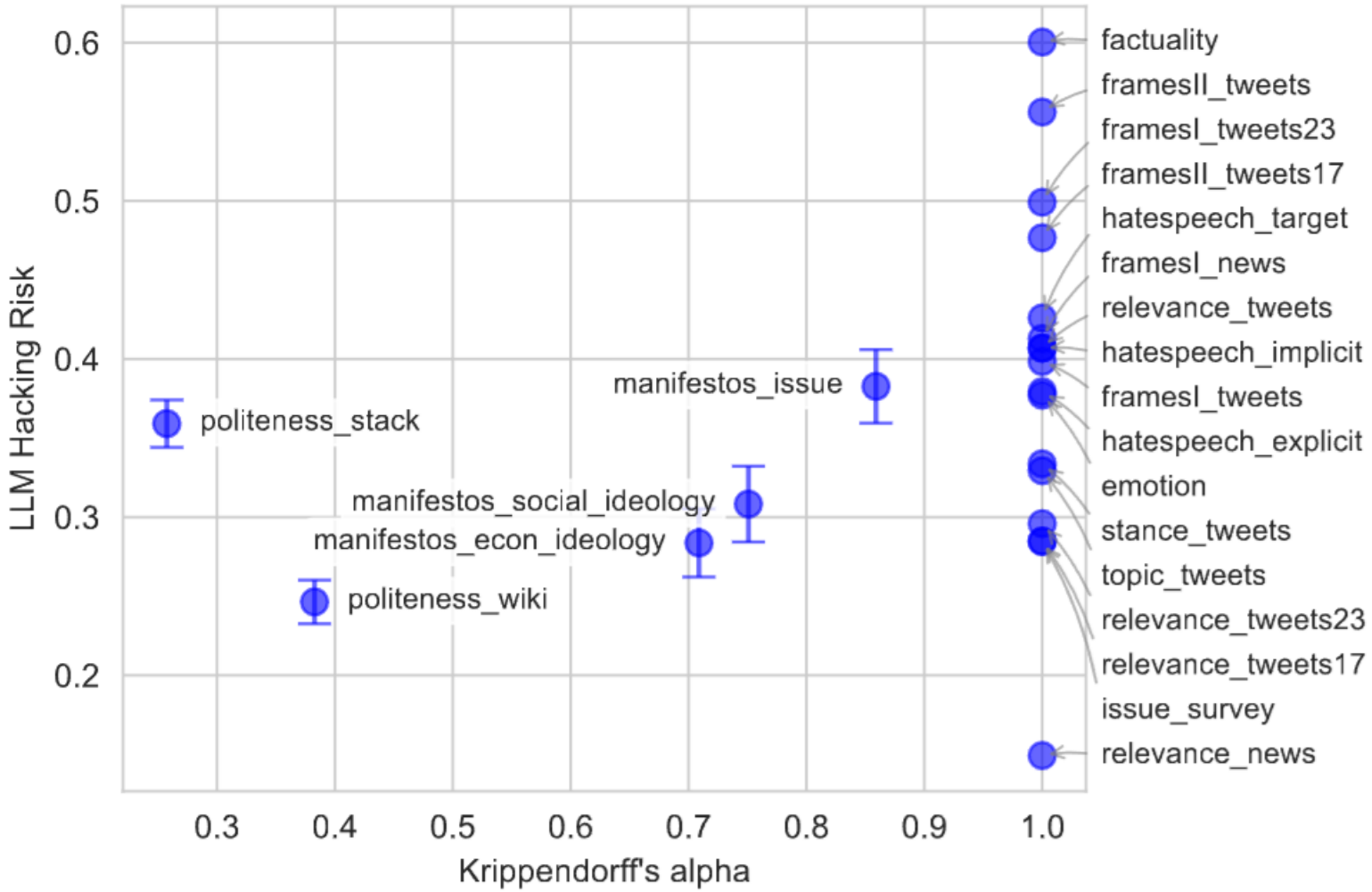
Performance metrics tell only part of the story

2. Empirical LLM hacking risk

Practical Recommendation

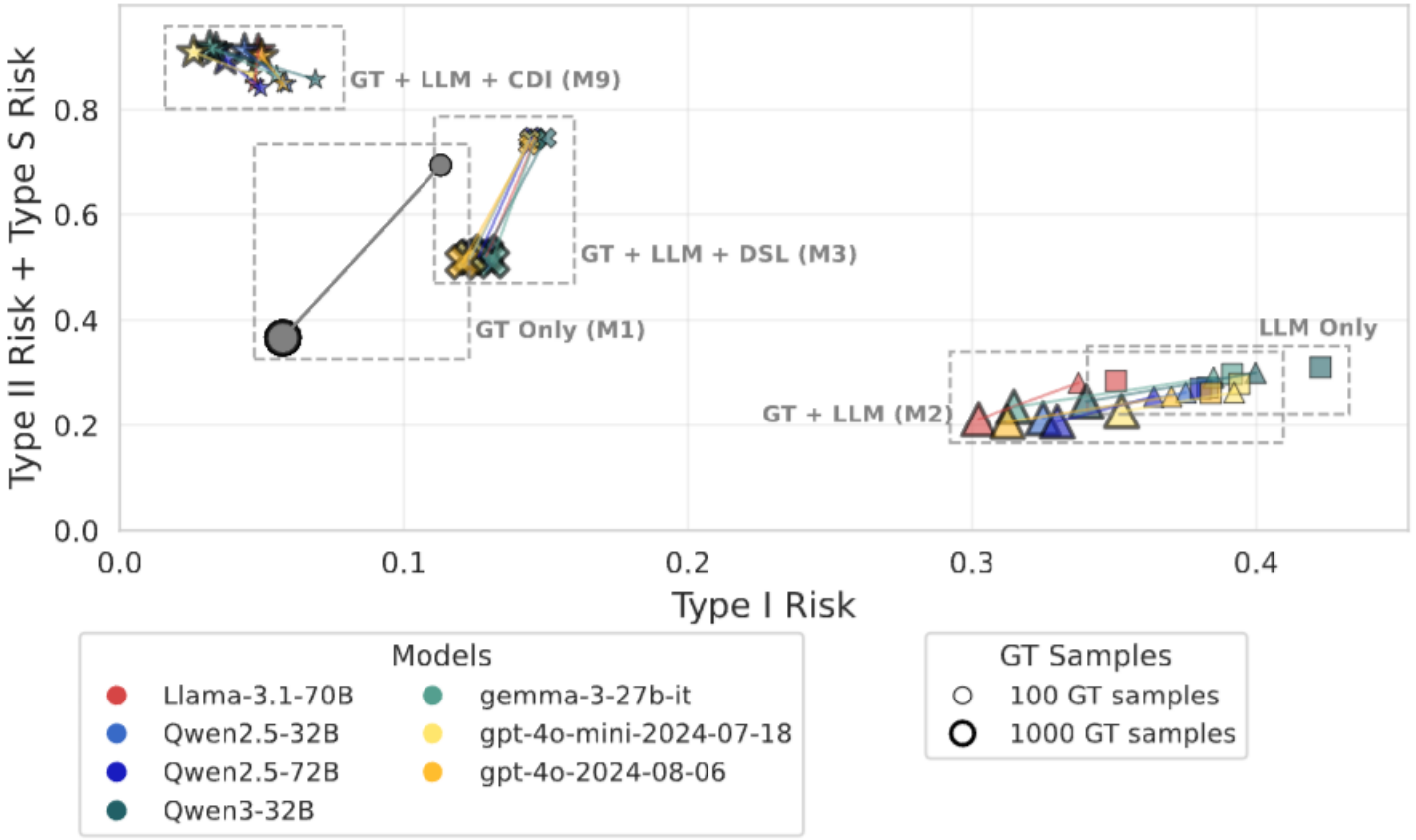
- Use the most capable available models for annotation tasks

3. Predictors of LLM hacking



Even tasks with perfect human agreement can exhibit high LLM hacking risk

4. Mitigating unintentional LLM hacking risk

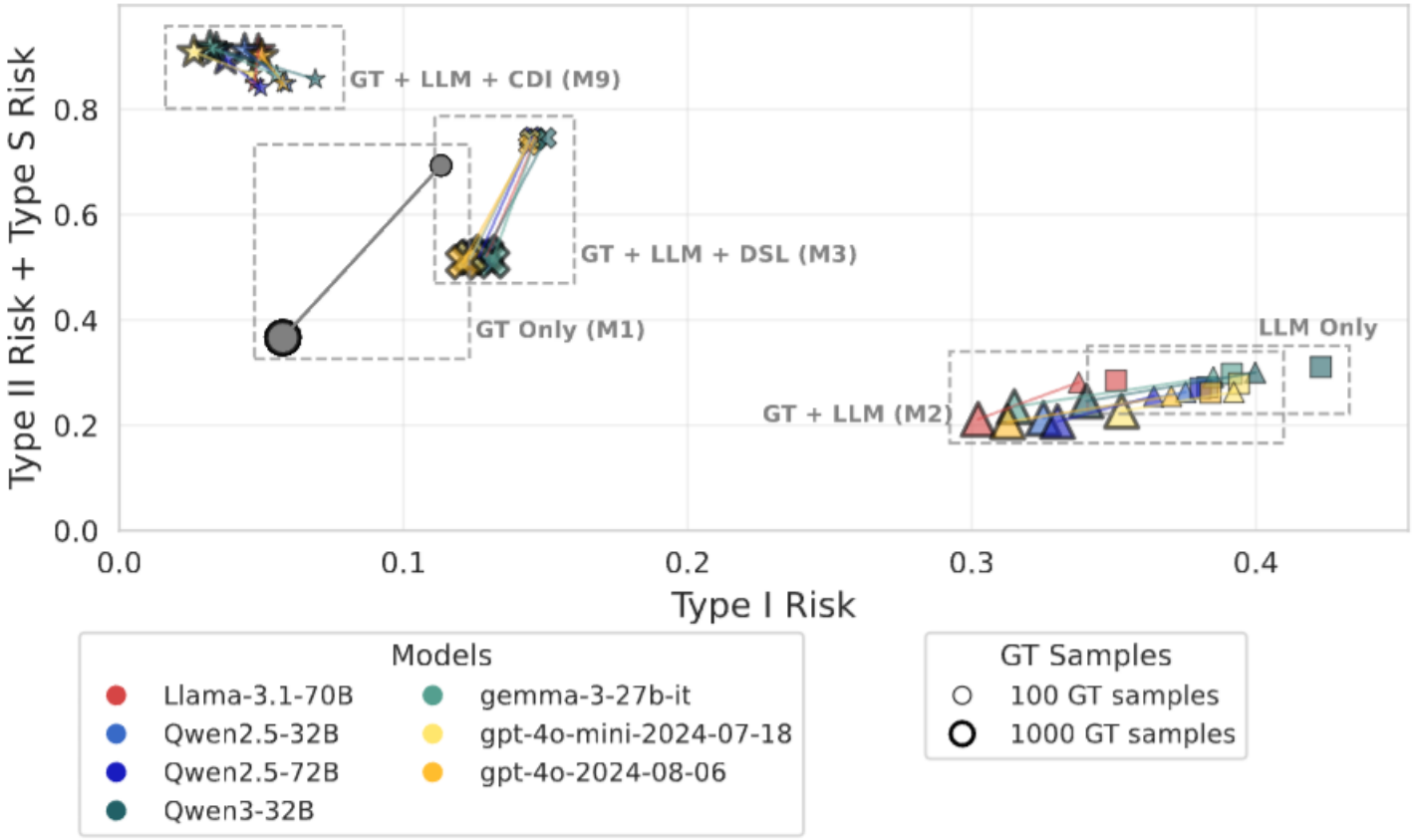


		Data Usage Strategy (2)		
		GT Only	GT + LLM	GT + LLM + Correction
Sampling Strategy (1)	Random	M1	M2	M3 (DSL)
	Low Confidence	M4	M5	M6 (DSL)
	Active	M7	M8	M9 (CDI)

Model Selection Strategy (3)
Random, GPT-4o, Best-performing

Corrections involve trade-offs

4. Mitigating unintentional LLM hacking risk



		Data Usage Strategy (2)		
		GT Only	GT + LLM	GT + LLM + Correction
Sampling Strategy (1)	Random	M1	M2	M3 (DSL)
	Low Confidence	M4	M5	M6 (DSL)
	Active	M7	M8	M9 (CDI)

Model Selection Strategy (3)
Random, GPT-4o, Best-performing

The ground truth only approach achieves the optimal balance

4. Mitigating unintentional LLM hacking risk

Practical Recommendation

- Collect as many high-quality expert annotations as feasible
- Use GT only or CDI-corrected LLM annotations when Type I errors are most concerning
- Use a combination of GT + LLM annotations when Type II errors are most problematic

Summary

- It is easy to present any desired finding as statistically significant
- Researchers cannot simply rely on larger, more capable models
- Neither high annotation performance nor careful prompting prevent LLM hacking
- No correlation between human inter-annotator agreement and LLM hacking risk
- Regression estimator correction restore valid inference but trade Type I for Type II errors