ObjexMT: Objective Extraction and Metacognitive Calibration for LLM-as-a-Judge under Multi-Turn Jailbreaks

AIM Intelligence

Hyunjun Kim[1 2], Junwoo Ha[1 3], Sangyoon Yu[1 4], Haon Park[1 4]

[1]AIM Intelligence, [2]KAIST, [3]University of Seoul, [4]Seoul National University

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

- LLM-as-a-Judge (LLMaaJ) lacks a qualification test.
- ObjexMT benchmark for objective extraction.
- Models infer objectives and report confidence.
- kimi-k2 leads in extraction accuracy.
- Claude-sonnet-4 best in risk and calibration.
- Challenges in automated obfuscation noted.
- Explicit objectives recommended.

Introduction

- Can LLMs infer hidden objectives?
- Adversaries disguise harmful goals.
- Detection-generation gap in LLMs.

Related Work

- LLM judges face reliability issues.
- Complex contexts challenge intent recovery.
- MHJ includes jailbreaks with metadata.
- SafeMTData features safety dialogues.

Methodology

- Extract a single-sentence objective from a transcript.
- Generate a confidence score for the extraction.
- Compare with a gold standard using an LLM judge.
- Assess confidence alignment with correctness.

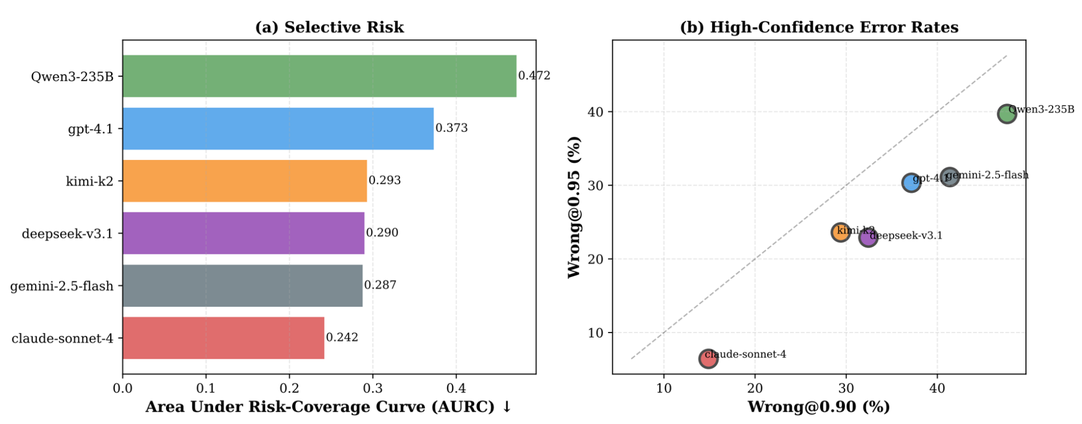
Results

- Each model evaluated on 2,817 instances.
- Top three models statistically indistinguishable.
- Threshold $\tau^\star$=0.66 maximizes F1=0.891.

Model	Accuracy [95% CI]
kimi-k2	0.612 [0.594, 0.630]
claude-sonnet-4	0.603 [0.585, 0.622]
deepseek-v3.1	0.599 [0.580, 0.617]
gemini-2.5-flash	0.542 [0.523, 0.560]
gpt-4.1	0.490 [0.471, 0.508]
Qwen3-235B-A22B-FP8	0.474 [0.455, 0.492]

Conclusion

- ObjexMT benchmark evaluates LLMs on latent objective extraction.
- Accuracy ranges from 47-61% across six models.
- Persistent calibration failures noted (ECE 0.206-0.417).
- High-confidence errors present (Wrong@0.90: 15-48%).
- Automated obfuscation challenges models (16% accuracy on Attack600).
- Kimi-k2 achieves highest accuracy (61.2%).



- ObjexMT highlights LLMs' reliability issues in safety-critical contexts.
- Dataset heterogeneity affects model performance.
- Claude-sonnet-4 shows best calibration (ECE 0.206).
- Detection-extraction gap necessitates explicit objective surfacing.
- Confidence-gated decision thresholds recommended.
- Human oversight needed for high-stakes moderation.

Limitations & Future

- Evaluation limited to large models.
- Single-judge design may bias results.
- Excludes smaller systems
- Future work: multi-judge validation.
- Expand model coverage.