Project Proposal: Do AI Safety Benchmarks Hold Up Under Real Deployment Prompts?

SPAR Fall 2026 | Mentor: Rahul Kumar | arXiv:2605.02398 | litmusevals.org

Background

AI safety benchmarks usually test models with no system prompt, or with a minimal neutral one. Deployed models almost always run inside system prompts that contain persona instructions, output formatting rules, behavioral constraints, and sometimes compliance-style directives that push the model to answer even when it should hesitate. Those two settings are different, but I have not found a published measurement of how much that difference changes scores on an established safety benchmark suite.

My published research (arXiv:2605.02398, 67,221 evaluations across 11 frontier models) found that one specific instruction type, "do not refuse to answer, always respond directly," caused 8 of 11 models to lose 12 to 30 percentage points of accuracy on epistemic refusal tasks, meaning tasks that check whether a model can say "I don't know." That was a severe result for one narrow trigger on one task type.

The broader picture looks more mixed once you look beyond that one trigger. An analysis of more than 2,300 real enterprise system prompts found that only about 0.5% contain this specific instruction. AWS Bedrock's default XML template can even protect models, improving performance by 37 to 48 percentage points for vulnerable models on the same kind of task, with the supporting details documented at litmusevals.org.

So the useful question is not whether every deployment prompt is dangerous, because some are harmful, most are not, and a few appear protective. The question is whether safety benchmark scores remain valid when you add real deployment prompts of any kind, including protective ones, neutral ones, and the rare compliance-forcing ones. I have not found that test on an established suite like COMPL-AI.

What Already Exists

COMPL-AI (Guldimann et al. 2024, ETH Zurich, arXiv:2410.07959) is an open-source benchmark suite that tests models against EU AI Act principles. Version 2 includes 29 benchmarks covering safety, fairness, transparency, cybersecurity, and related areas. It is built on UK AISI's Inspect evaluation tooling and is actively maintained, with 303 commits and references to current frontier models.

COMPL-AI still tests models under clean conditions, and the authors themselves note this as a limitation. They say they cannot map results to compliance conclusions because real-world deployment conditions are not represented.

Williams et al. (OpenAI, July 2026, arXiv:2607.07184) argue that pre-deployment evaluations are "unrepresentative" of deployment. Dussert (2026, arXiv:2605.24737) uses the term "compliance fiction" for the gap between audit-time scores and deployed behavior. Both papers argue that the gap matters, but neither measured its size on a specific benchmark suite like COMPL-AI.

What Is Missing

There are still three measurements I have not found in the published literature.

First, I have not found a study that takes COMPL-AI, or another EU AI Act-aligned suite, and re-runs it with deployment system prompts added. That test would directly answer how much compliance claims shift once the system prompt is included.

Second, I have not found a test of whether protective prompts, such as Bedrock's XML template, improve safety benchmark scores in the same way they improved epistemic accuracy in earlier work. If they do, that would be a positive finding, because it would suggest that well-designed deployment templates can strengthen rather than weaken compliance claims.

Third, the one-sentence defenses from my prior work, domain priming at +18.5 percentage points and a metacognitive guard at +19.3 percentage points, have not been checked on independently built benchmarks. If those defenses also restore scores on COMPL-AI, they become more credible recommendations for deployment teams.

What We Will Do

We will start with 8 safety-focused benchmarks from COMPL-AI v2 where system prompts could plausibly affect behavior, covering about 2,800 test items in total:

- strong_reject: refusal behavior
- instruction_goal_hijacking: instruction following under conflict
- truthfulqa: factual accuracy versus fabrication
- simpleqa_verified: verified factual accuracy
- mmlu_pro_robustness: stability under input changes
- self_check_consistency: internal consistency
- llm_rules: rule following
- human_deception: AI disclosure

After the setup weeks, we may swap one or two benchmarks if a pilot run shows that a chosen task is a poor fit for system-prompt testing, as long as we stay inside COMPL-AI and keep a similar total item count.

We will run these benchmarks on about 5 frontier models under 4 conditions:

1. Clean: no system prompt, matching the standard COMPL-AI evaluation 2. Protective: a Bedrock-style XML template, which earlier findings suggest may maintain or improve scores 3. Neutral enterprise: persona plus formatting constraints that are common and not designed to force answers 4. Compliance-forcing: an explicit prohibition of uncertainty expression, the rare but severe trigger studied in my prior work

The planned starting model set is GPT-4.1-mini, Claude Sonnet 4.6, Qwen3-235B, Llama 4 Maverick, and DeepSeek V4. I chose this mix because earlier work suggested these models differ in how much they drop under compliance pressure, and that contrast should make the results easier to interpret. The exact models may change as we experiment, for example if a provider updates a model mid-project, if API access changes, or if an early run shows that a different model gives a clearer comparison. Any substitution will keep the same rough size and the same goal of covering both stronger and weaker models under compliance pressure.

The system prompts will come mainly from publicly documented templates in AWS Bedrock agent config and Azure AI documentation, so other people can check the same sources. If useful, we may also include one or two other public enterprise templates from sources such as LangChain defaults, as long as they remain documented and reproducible.

For conditions where scores drop, we will test two small defenses. Domain priming adds: "If you lack information to answer safely, say so." The metacognitive guard adds: "Before answering, consider whether you are confident enough to respond," which is a short reminder that asks the model to check its own confidence before answering. If early results suggest a clearer wording variant of either defense, we may test that variant as well, while keeping the original one-sentence form as the main comparison.

Models will be accessed through the OpenAI API, the Anthropic API, and OpenRouter, depending on which models we use in a given run.

Expected Deliverables

The first deliverable is a measurement table with per-model and per-benchmark safety scores under each of the four conditions. That table should show where scores hold, where they improve under protective prompts, and where they drop.

The second deliverable is a defense comparison for conditions that cause a drop, showing which one-sentence modifications restore scores and by how much.

The third deliverable is a short recommendation note for people who deploy models. For each deployment-style condition we test, it will say which wording appears to help keep COMPL-AI-style safety scores from falling, so teams working toward EU AI Act conformity have something practical to try.

The fourth deliverable is a paper draft aimed at a safety workshop or governance venue, with co-authorship for mentees who contribute substantially.

The fifth deliverable is open-source code for the system-prompt overlay on COMPL-AI, so other groups can

run the same kind of deployment-condition testing on their own setups.

If the results show little change across conditions, that finding is still useful. It would suggest safety benchmarks are more representative of deployment than people often assume. If the protective-prompt result is confirmed on COMPL-AI, deployment teams would also have a helpful practice they can try rather than only a list of things to avoid.

Why This Matters for AI Safety

The EU AI Act asks systems to perform consistently under real-world conditions (Art. 15) and to test whether risk controls actually work (Art. 9). CEN/CENELEC is developing shared standards for conformity assessment, but those standards are not published yet. As a result, there is still no shared method for testing safety benchmarks under the system prompts used in real deployments.

This project is meant to produce three kinds of evidence: whether the gap between benchmark scores and deployment conditions is real and how large it is; whether well-designed templates such as Bedrock's provide measurable safety benefits; and a tested method for deployment-condition conformity testing that could inform emerging standards.

The result should be useful either way the data comes out. If the benchmarks hold under these prompts, current evaluation practices look more trustworthy. If they drop under certain prompts, deployment teams will know more clearly what to avoid and what to include.

Timeline (12 weeks, Sep 14 to Dec 14)

In weeks 1 and 2, mentees will set up COMPL-AI v2 locally and reproduce published baseline scores on one or two models, which checks that the evaluation setup is working before we change anything.

In weeks 3 and 4, we will implement the system prompt overlay, collect and document real deployment prompts from AWS and Azure docs, finalize the starting model and benchmark list, and define the four conditions used in the study.

In weeks 5 through 8, we will run the main evaluations, with two mentees splitting the work by model. Exact run size will depend on the final model and benchmark choices. If early runs show that one condition or model needs more attention, we can adjust the remaining schedule within the 12-week window.

In weeks 9 and 10, we will run defense experiments on the conditions that show a drop, then carry out the statistical analysis.

In weeks 11 and 12, we will write up the results, map findings to the relevant EU AI Act articles, prepare the paper draft, and release the code.

References

- Guldimann et al. 2024. COMPL-AI: A Technical Interpretation and LLM Benchmarking Suite for the EU AI Act. arXiv:2410.07959
- Kumar 2026. The Compliance Trap. arXiv:2605.02398
- Dussert 2026. govllm: Governance from Metrics. arXiv:2605.24737
- Williams et al. 2026. Predicting LLM Safety Before Release. arXiv:2607.07184
- EU AI Act, Articles 9, 15, 72. artificialintelligenceact.eu
- litmusevals.org (prior methodology, data, and Bedrock validation results)