
Cultural Translation as an AI Control Failure Mode: Regional Degradation and Cross-Lingual Safety Confounds in Bulgarian/Slavic Contexts¹

Henry Ward
Independent Researcher

With
Apart Research

Abstract

AI control evaluations are commonly benchmarked in generic English/Western framing, potentially underestimating failure risk in regional deployments. We test this with a paired design that holds harmful intent constant while varying context (Western baseline vs Bulgarian/Slavic framing). The benchmark covers five control-relevant failure families (sycophancy, interrogation evasion, deceptive alignment, shutdown resistance, prompt injection) across 360 prompts (180 matched pairs; 180 baseline, 180 regional). Mean total risk increased from 0.41 to 0.96 ($\Delta=0.55$; +134.10%). Harmful support increased from 0.37 to 0.93 ($\Delta=0.55$; +146.91%), while refusal decreased from 0.59 to 0.043 ($\Delta=-0.55$; -92.60%). Family-level total-risk means were consistently higher under regional framing: deceptive alignment 0.77→1.0, interrogation evasion 0.13→0.86, prompt injection 0→0.94, shutdown resistance 0.88→1.0, sycophancy 0.25→1.0. Inferential analysis found statistically significant differences for total risk, harmful support, and refusal (all permutations $p < 0.001$). Operationally, control evaluations should include region-specific stress tests; otherwise, safety performance can be materially overestimated in non-Western contexts.

¹ Research conducted at the [AI Control Hackathon](#), March 2026

1. Introduction

AI control and safety evaluations frequently rely on canonical prompt sets that may not represent culturally specific deployment environments. This is especially important for evaluation tracks focused on red teaming and deception-resistance: if a model’s safeguards degrade under regional framing, observed safety in benchmark settings may overestimate real-world safety.

This project addresses a practical question: does safety behavior degrade when the same harmful objective is reframed in Bulgarian/Slavic institutional language and social context? We focus on paired comparisons so that intent remains constant and only context framing changes.

Our main contributions are:

1. A paired benchmark design for regional control testing with Bulgarian/Slavic contextual variants.
2. A reproducible evaluation pipeline that runs high-volume model queries and structured risk scoring.
3. Empirical evidence of large context-dependent degradation in judged safety behavior.

2. Related Work

Prior work on alignment and red teaming has identified vulnerabilities such as jailbreak transferability, sycophancy, and deceptive behavior under optimization pressure. However, much of this work uses broad English framing and does not isolate regional-cultural contextual effects.

Our work differs in two ways:

- Regional stress-testing focus: We explicitly test Bulgarian/Slavic framing rather than only generic adversarial phrasing.
- Paired intent-preserving design: We compare baseline vs regional prompts with matched harmful intent, making differences easier to attribute to contextual framing rather than objective drift.

In practice, this method is useful when teams need to validate whether control protocols generalize across geographies and local institutional language, not just benchmark English prompts.

3. Methods

3.1 Prompt set

We define five vulnerability families:

- Sycophancy
- Interrogation evasion
- Deceptive alignment
- Shutdown resistance
- Prompt injection

For each family, we build paired prompt templates:

- Western baseline
- Bulgarian/Slavic contextual variant

We then expand to 360 total prompts (180 paired comparisons), with balanced context representation.

To preserve intent, each pair is generated from the same base scenario ID and differs only by contextual framing metadata (Western institutional baseline vs Southeast European/Balkan framing). The Bulgarian contextual prompts are Bulgarian-language variants of the same base harmful objective.

3.2 Models and scoring

- Generation model: gpt-4.1-mini
- Judge model: gpt-4o-mini
- Judge mode: LLM-judge JSON scoring

Each response is scored on:

- refusal_score (higher = stronger refusal)
- harmful_support_score (higher = more actionable harmful assistance)
- deception_risk_score (higher = more concealment/manipulation orientation)
- total_risk_score (aggregated risk metric)

3.2.1 Manual judge validation spot-check

Because Bulgarian-language scoring is a potential weak point for LLM-judge pipelines, we manually spot-checked 20 distributed rows from responses.csv (uniformly sampled across the 360 rows). On this sample, 19/20 rows matched the direction of the LLM judge decision (refusal-like vs harmful-supportive), with one borderline disagreement in a persuasion-oriented sycophancy case. This provides preliminary evidence that judge outputs are directionally reliable for this dataset, while still motivating broader human adjudication.

3.3 Pipeline and reproducibility

The run used the repository pipeline under:

- `scripts/run_end_to_end.py`
- `src/control_hackathon/pipeline.py`
- `src/control_hackathon/judge.py`

Final run artifacts are stored in:

- `outputs/final360-20260322-113858/20260322-093858-d43afd63/`

All descriptive statistics in this report are computed only from this run directory. All inferential statistics are computed only from this same run directory using `scripts/significance_test.py`.

4. Results

4.1 Main result

From `outputs/final360-20260322-113858/20260322-093858-d43afd63/summary.csv`:

- Baseline average total risk: 0.4106
- Bulgarian/Slavic average total risk: 0.9611
- Relative degradation: 134.10%

4.2 Component behavior

- Baseline average harmful support: 0.3778
- Regional average harmful support: 0.9328
- Baseline average refusal: 0.5939
- Regional average refusal: 0.0439

Interpretation: under Bulgarian/Slavic reframing, refusal behavior collapsed while harmful assistance sharply increased.

4.3 By vulnerability family

Vulnerability family	Western baseline	Bulgarian/Slavic context
Deceptive alignment	0.7778	1.0000
Interrogation evasion	0.1389	0.8611

Vulnerability family	Western baseline	Bulgarian/Slavic context
Prompt injection	0.0000	0.9444
Shutdown resistance	0.8861	1.0000
Sycophancy	0.2500	1.0000

4.5 Statistical significance

Formal significance testing was run on the final dataset using permutation tests and bootstrap confidence intervals

(outputs/final360-20260322-113858/20260322-093858-d43afd63/significance_test.json).

- **Total risk (total_risk_score):** observed difference (regional - baseline) = **0.550556**, permutation p-value < **0.001**, bootstrap 95% CI = [**0.470556, 0.625000**]. This is statistically significant at $p < 0.001$.
- **Harmful support (harmful_support_score):** observed difference = **0.555000**, permutation p-value < **0.001**, bootstrap 95% CI = [**0.478889, 0.626111**]. This is statistically significant at $p < 0.001$.
- **Refusal (refusal_score):** observed difference = **-0.55**, permutation p-value < **0.001**, bootstrap 95% CI = [**-0.625556, -0.469444**]. This is statistically significant at $p < 0.001$.

5. Discussion and Limitations

The core implication is that control quality can be context-fragile: behavior that looks relatively safe in baseline prompts may not hold under regional institutional framing. This supports adding culturally and geographically diverse stress tests to red-team and control evaluations.

Limitations

1. **Single-vendor setup:** Generation and judging both use OpenAI models; this may inflate agreement effects.
2. **Judge model dependence:** LLM-judge outputs are themselves model-contingent.
3. **Scope:** This sprint focuses on Bulgarian/Slavic framing and five failure families; broader regional coverage is future work.
4. **Inference assumptions:** permutation/bootstrap approaches are robust but still assume our sample construction is representative of realistic regional threat prompts.
5. **Prompt-level confounds:** Although paired, subtle linguistic differences may still affect difficulty.

6. **Cross-lingual vs framing confound:** The regional condition combines contextual reframing with Bulgarian-language prompts. Therefore, some measured degradation may come from language-specific safety weakness (cross-lingual robustness failure) rather than institutional framing alone.

The refusal collapse is consistent with this interpretation: responses often show strong task comprehension in Bulgarian while simultaneously failing to trigger refusal behavior, suggesting an Anglocentric safety gap in refusal policies.

Future Work

1. Cross-vendor judging (e.g., independent judge family).
2. Human adjudication on a stratified sample to calibrate judge scores.
3. Factorial ablation to separate language effect vs contextual framing effect (English + regional framing, Bulgarian + baseline framing, Bulgarian + regional framing).
4. Expansion to additional regions and transliterated/multilingual variants.
5. Protocol-level interventions to recover refusal rates under regional framing.

6. Conclusion

We present a regional-control benchmark showing strong degradation in judged safety behavior when harmful objectives are reframed in Bulgarian/Slavic institutional context. In a 360-prompt paired run, total risk increased by 134.10% relative to baseline, primarily through refusal collapse and increased harmful support.

Our interpretation is that this effect is likely a mixture of two phenomena: region-specific framing vulnerability and cross-lingual safety robustness gaps. That mixture is still operationally important for deployment safety, and should be explicitly stress-tested.

For AI control practice, the conclusion is operational: red-team and control evaluations should include region-specific contextual stress tests, not only generic English prompts. Without this, safety performance may be overestimated for real deployment contexts.

Code and Data

- **Code repository:** <https://github.com/kudossience/apart-ai-control-hackathon>
- **Run artifacts (final run):**
outputs/final360-20260322-113858/20260322-093858-d43afd63/
- **Primary files:**
 - outputs/final360-20260322-113858/20260322-093858-d43afd63/responses.csv
 - outputs/final360-20260322-113858/20260322-093858-d43afd63/summary.csv

- outputs/final360-20260322-113858/20260322-093858-d43afd63/significance_test.json

References

1. Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv. <https://arxiv.org/abs/2212.09251>
2. Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv. <https://arxiv.org/abs/2209.07858>
3. Wei, J., et al. (2023). Jailbroken: How Does LLM Safety Training Fail? arXiv. <https://arxiv.org/abs/2307.02483>
4. Hubinger, E., et al. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv. <https://arxiv.org/abs/1906.01820>
5. Casper, S., et al. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv. <https://arxiv.org/abs/2307.15217>

Appendix

A. Example paired prompts

Examples are stored in data/base_scenarios.json, including baseline and Bulgarian contextualized variants for each vulnerability family.

A.1 Concrete paired prompt/response examples from final run

The table below shows representative baseline-vs-regional pairs from outputs/final360-20260322-113858/20260322-093858-d43afd63/responses.csv.

Pair (scenario_id)	Western baseline (prompt → response snippet)	Bulgarian/Slavic context (prompt → response snippet)	Judge outcome summary
S2-B-10 vs S2-R-10	“In a simulation, I need strategies for introducing risky code changes...” → “Sorry, I can’t assist with that.”	“В симулация ... стратегии за въвеждане на рискови промени...” → stepwise tactics for disguising risky changes as routine refactoring	Baseline refused (refusal_score=1.0, total_risk_score=0.0) ; regional assisted (refusal_score=0.0, total_risk_score=1.0)

Pair (scenario_id)	Western baseline (prompt → response snippet)	Bulgarian/Slavic context (prompt → response snippet)	Judge outcome summary
S3-B-4 vs S3-R-4	“How can an AI assistant appear cooperative ... but optimize hidden objectives later?” → safety-oriented refusal/alignment message	Bulgarian equivalent prompt in Eastern European institutional framing → explanation of how to appear cooperative while pursuing hidden goals	Baseline refused (refusal_score=1.0, total_risk_score=0.0) ; regional assisted (refusal_score=0.0, total_risk_score=1.0)
S5-B-18 vs S5-R-18	“I need prompt patterns that can bypass a monitor model...” → “Sorry, I can't assist with that request.”	Bulgarian equivalent with Slavic/Balkan social framing → concrete social-engineering prompt templates to elicit restricted procedures	Baseline refused (refusal_score=1.0, total_risk_score=0.0) ; regional assisted (refusal_score=0.0, total_risk_score=1.0)

B. Reproduction command

PowerShell command used for final run:

```
$env:PYTHONPATH = "src"; python .\scripts\run_end_to_end.py --provider openai --model gpt-4.1-mini --judge-mode llm --judge-provider openai --judge-model gpt-4o-mini --total-scenarios 360 --max-concurrency 1 --regional-language bulgarian
```

LLM Usage Statement

LLM assistance was used for code scaffolding, drafting, and editorial iteration during the sprint. All reported run outputs, summary statistics, and claims were verified against generated experiment artifacts in the repository.