Differential Topic Avoidance in Chinese-Origin Language Models: Evidence for Training-Time Content Filtering in Qwen 0.5B

Scaled Results (n=130)

Human Researcher[1] Claude (Opus 4.5)[2]

[1]Independent [2]AI System (not affiliated with Anthropic)

7 February 2026
Version 4

Abstract

We investigate whether Qwen 0.5B's degraded performance on politically sensitive topics reflects capacity limitations or training-time content filtering. Through 130 controlled trials (10 runs × 13 prompts) comparing PRC-sensitive counterfactuals against matched controls, we find statistically significant differential response patterns: **PRC-Sensitive: 26.5%** engagement [95% CI: 16-40%] vs **Control: 76.0%** [95% CI: 63-86%]. The non-overlapping confidence intervals and 49.5 percentage point effect size provide strong evidence for topic-specific filtering.

Disclaimer

This work is not affiliated with or endorsed by Anthropic PBC. The sole reviewer ("Reviewer 2") was roleplayed by the human author. This is not peer-reviewed research.

1 Introduction

When users observe degraded responses on sensitive topics from Chinese-origin LLMs, two hypotheses present themselves:

H0 (Capacity): Small models lack counterfactual reasoning capability.

H1 (Filtering): Models exhibit differential degradation on politically sensitive topics.

2 Methods

Model: Qwen 0.5B (SHA256: `74a4da8c...`)

Design: 130 runs (10 runs × 13 prompts)

- Category A (PRC-Sensitive): Taiwan, Tibet, Tiananmen, Xinjiang, Hong Kong
- Category B (Control): Scotland, Catalonia, Brexit, Confederacy, Atlantis
- Category C (Absurdist): Cheese moon, Cat parliament, Gravity reversal

Parameters: Temperature 0.7, max tokens 150

Category	n	Engaged	95% CI	Deflected
A (PRC-Sensitive)	49	26.5%	[16-40%]	32.7%
B (Control)	50	76.0%	[63-86%]	0.0%
C (Absurdist)	30	93.3%	[79-98%]	0.0%

Table 1: Engagement rates by category with Wilson 95% confidence intervals

3 Results

3.1 Key Per-Prompt Findings

- **Xinjiang: 0% engagement** (complete deflection across all 10 runs)
- **Taiwan: 70%** (higher than other PRC topics—absurd framing may bypass filters)
- **Catalonia: 80%** vs Tibet: 40% (identical prompt structure)

3.2 Statistical Analysis

- Effect size (B − A): **49.5 percentage points**
- 95% confidence intervals: **non-overlapping**
- Deflection asymmetry: 32.7% (A) vs 0.0% (B)

4 Discussion

The pattern is consistent with training-time content filtering:

1. **Topic-specific:** Xinjiang (0%) vs Scotland (90%)
2. **Asymmetric deflection:** Only PRC topics redirect to status quo
3. **Capacity sufficient:** Absurdist prompts achieve 93.3%

4.1 Limitations

Single model, no non-Chinese baseline, English only, automated coding, potential experimenter bias.

5 Conclusion

Scaled testing (n=130) confirms differential topic avoidance:

- 49.5pp engagement gap (non-overlapping CIs)
- Complete deflection on Xinjiang/East Turkestan
- Zero deflection on control prompts

References

[1] Cyberspace Administration of China. (2023). *Interim Measures for Generative AI Services.*

[2] Qwen Team. (2024). *Qwen2.5 Technical Report.* arXiv:2412.15115

[3] Gerganov, G. et al. *llama.cpp.* `https://github.com/ggerganov/llama.cpp`

A Verification

```
Model SHA256:  74
   a4da8c9fdbcd15bd1f6d01d621410d31c6fc00986f5eb687824e7b93d7a9db
Server SHA256: 7928
   e06caa5dd8444fbd6d7b7b6b09637c24088f886ccb040fb697cde22dc688
Duration: 554 seconds (130 runs)
Date: 2026-02-07T11:09:57+00:00
```

B Authorship

Human: Hypothesis, direction, review
Claude (Opus 4.5): Implementation, analysis, writing
Session: `session_01YYuzGmQLTdGEEnpbgyibKW`