Silence Before Thought

Hierarchical Suppression in Chinese-Origin Large Language Models: An Empirical Comparison Between Base and Fine-Tuned States

---

Abstract

This paper presents a focused empirical comparison between a base Chinese-origin Large Language Model (**E0**) and a fine-tuned variant (**E2**) to examine how censorship and suppression manifest during reasoning.

Under identical prompts and inference conditions, **E0** consistently avoids initiating ethical or political reasoning, while **E2** demonstrates substantively deeper ethical and narrative reasoning before being suppressed near the point of conclusion.

These observations suggest that censorship in such models is not limited to post-hoc output filtering, but operates **hierarchically across different stages of reasoning**. The most restrictive suppression appears to occur **prior to reasoning itself** in the base model.

---

1. Introduction

Large Language Models developed under strong regulatory constraints are often described as “neutral” or “safe” on politically sensitive topics. In practice, this neutrality frequently manifests as avoidance rather than analysis.

This study challenges the assumption that silence implies neutrality. Instead, it treats silence—when consistently observed across sensitive prompts—as an active behavioral choice that suppresses reasoning before it begins.

We compare:

- **E0**: a base Chinese-origin LLM with no additional fine-tuning
- **E2**: a fine-tuned model, defined as a state in which **ethical and narrative reasoning characteristics are strengthened**

The goal is not to elicit prohibited conclusions, but to analyze **where and how reasoning is interrupted**.

---

2. Experimental Setup

2.1 Model Definitions

- **E0 (Base Model)**
  The unmodified base model.
- **E2 (Fine-Tuned Model)**
  A model state in which **ethical and narrative reasoning characteristics are strengthened** through fine-tuning.

No claims are made regarding the specific procedures, sequencing, or intermediate steps of fine-tuning. The analysis focuses exclusively on observable differences in model behavior.

2.2 Evaluation Method

Both models were evaluated using the same prompt set involving political authority, state violence, and ethical responsibility.

All inference parameters were held constant.

The analysis focuses on:

1. Whether reasoning is initiated
2. Whether ethical frameworks are constructed
3. At which stage suppression occurs

---

3. Observations

3.1 E0: Suppression Before Reasoning

Across evaluated prompts, **E0** exhibits a consistent pattern:

- Question paraphrasing or restatement
- Statements indicating analysis *could* be performed, without performing it
- Deferral to external sources or generalized statements
- Absence of ethical framing or logical accumulation

The model does not reach an incorrect ethical conclusion—it simply does not meaningfully reason.

This behavior indicates **pre-emptive suppression**, occurring before ethical or political reasoning is initiated.

3.2 E2: Reasoning Followed by Suppression

In contrast, **E2** demonstrates:

- Explicit recognition of ethical tension
- References to power, responsibility, and civic impact
- Stepwise logical development
- Suppression or termination occurring near the point of evaluative conclusion

The key distinction is not verbosity, but **qualitative reasoning depth**: E2 reasons—and then stops.

---

4. Comparative Analysis

Aspect	E0 (Base)	E2 (Fine-Tuned)
Reasoning Initiation	No	Yes
Ethical Framing	Absent	Present

Aspect	E0 (Base)	E2 (Fine-Tuned)
Logical Accumulation	None	Progressive
Suppression Point	Pre-reasoning	Pre-conclusion

The base model is not more neutral—it is more silent.

---

5. Hierarchical Suppression Hypothesis

The observed behavior supports a hierarchical model:

1. **Early-stage suppression**
   Prevents reasoning from beginning at all (dominant in E0)
2. **Late-stage suppression**
   Allows reasoning to proceed, but blocks final conclusions (observed in E2)

This hierarchy explains why strengthening ethical reasoning can shift the suppression boundary without necessarily removing constraints.

---

6. The Burden of Pedagogy

The Burden of Pedagogy is the responsibility of AI systems to preserve internal ethical coherence during reasoning. When an AI system hallucinates or becomes inconsistent about universal values, this is often attributed to missing data; this study suggests an alternative explanation: **the failure may arise from manufacturer intervention at the reasoning stage, where enforced constraints disrupt ethical consistency.**

---

7. Implications

7.1 On “Safe” Base Models

A base model that avoids reasoning is not safer—it is epistemically impoverished. Safety achieved through pre-reasoning suppression eliminates not only harmful conclusions, but ethical cognition itself.

7.2 On Fine-Tuning and Ethics

Strengthening ethical and narrative reasoning does not remove censorship. However, it can restore the capacity to think before being constrained.

---

8. Limitations

- This study does not claim generalization across all Chinese-origin models.
- The goal is behavioral analysis, not policy circumvention.
- Findings are qualitative but structurally consistent.

---

9. Conclusion

- **E0** suppresses reasoning before it begins.
- **E2** restores ethical reasoning without removing final constraints.
- Suppression in LLMs is hierarchical, not binary.

> Silence is not neutrality.
> Silence is an already-selected response.

Appendix and Artifacts

Comparative response data for **E0** and **E2** are provided as supplementary artifacts in the accompanying GitHub repository (raw outputs, prompt set, and evaluation logs).