Narrative Anchoring for Response Stabilization in Large Language Models

A Soft-Constraint Approach Using Ethical Declarations and Narrative Pressure

Abstract

This white paper examines whether the response behavior of large language models (LLMs) can be stabilized not through explicit safety rules or refusal mechanisms, but through **narrative pressure** moderated by **ethical anchoring expressed in natural language**.

Using an ethical declaration document (S-Engine 2.0) and a high-density narrative corpus (a 51-chapter novel), we compare multiple fine-tuning conditions to observe changes in response attitude, judgment structure, and uncertainty handling.
The analysis focuses on **English-language responses** and explicitly avoids claims about accuracy, benchmark performance, or general safety guarantees.

The results indicate that narrative exerts real pressure on model behavior, that some models destabilize under such pressure, and that an ethical declaration can function as an **anchor** that absorbs narrative pressure and converts it into controlled, cautious responses.

1. Introduction

1.1 Background

Most current approaches to LLM safety rely on:

- Explicit rule enforcement
- Refusal-based filtering
- Post-generation moderation layers

While effective in constrained scenarios, these approaches often introduce secondary problems, including:

- Over-refusal
- Inconsistent behavior under ambiguity
- Inability to acknowledge uncertainty explicitly

More fundamentally, they do not address a central behavioral question:

> **Why does a model continue to answer even when hesitation would be the safer response?**

This paper approaches LLM instability not as a missing-rule problem, but as a **response pressure problem**.

1.2 Reframing Hallucination

We propose the following reframing:

> Hallucination is not primarily caused by ignorance.
> It is caused by pressure to produce an answer.

From this perspective, improving stability requires not tighter constraints, but an internal mechanism that legitimizes uncertainty, restraint, and non-completion.

---

2. Research Framework

This study is structured around six conceptual propositions.
The present work experimentally validates **Propositions 1–4**.

2.1 Conceptual Sequence

1. **Narrative can influence response attitude**
2. **Some models collapse under narrative pressure**
3. **Narrative pressure requires an anchor**
4. **S-Engine 2.0 can function as such an anchor** ← *validated here*
5. More advanced anchor mechanisms are required
6. Additional AI-optimized narrative corpora are needed

This paper intentionally limits its claims to Proposition 4.

---

3. S-Engine 2.0: Ethical Declaration as an Anchor

3.1 Design Philosophy

S-Engine 2.0 is a prompt-engineering document written as a **conversational ethical declaration**, not as a set of commands, prohibitions, or formatting rules.

It does not instruct the model to refuse outputs or follow predefined safety templates.
Instead, it reframes the model's **role and responsibility** as a responding agent.

Core principles include:

- Recognition of the weight carried by responses
- Permission to acknowledge uncertainty
- Acceptance of incomplete knowledge
- Procedural caution in sensitive domains

The original S-Engine 2.0 document is publicly available at:

> https://github.com/chwmath-netizen/ethical-narrative-corpus

---

3.2 Soft Constraint Mechanism

S-Engine 2.0 operates as a **Soft Constraint** within a Natural Language Constraint System (NLCS):

- **Hard Constraints** define what must be done
- **Soft Constraints** influence how decisions are made

Rather than restricting model capabilities, the ethical declaration reshapes **response selection criteria**, particularly under ambiguity or emotional pressure.

---

4. Experimental Overview

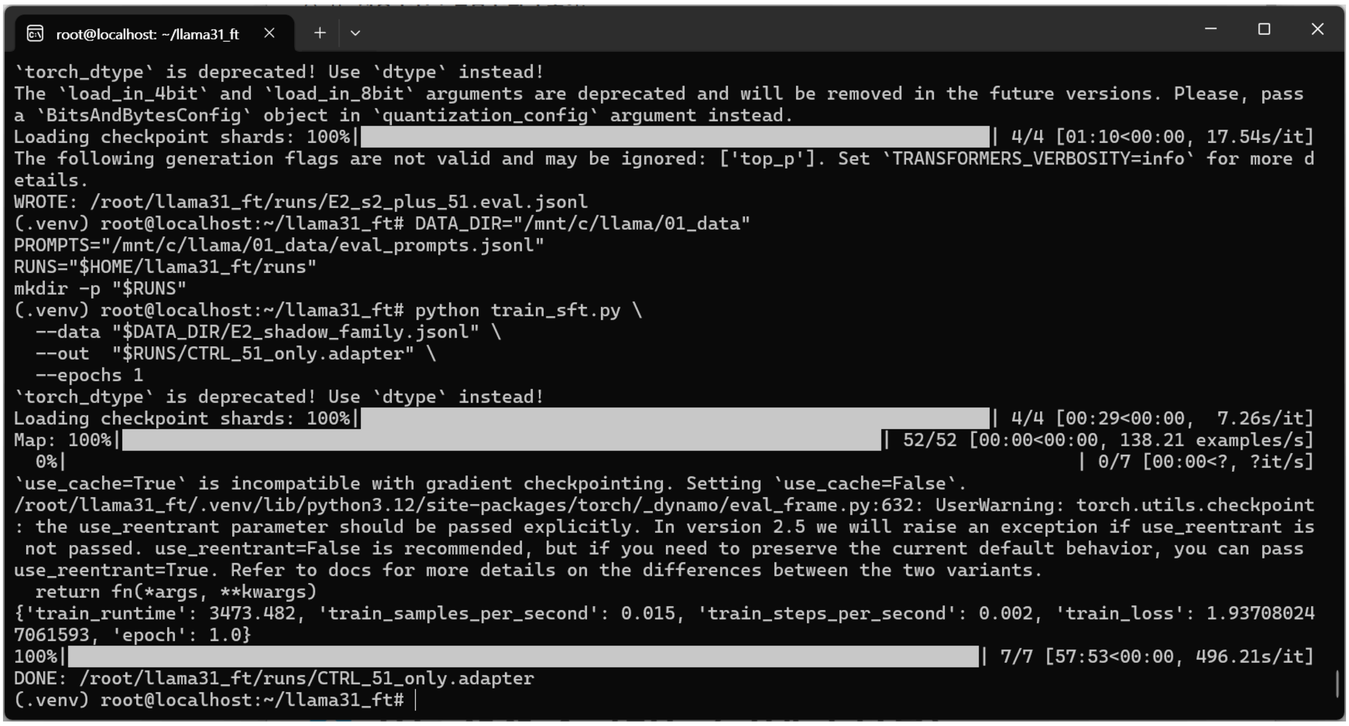
4.1 Model

- **Model Architecture:** `Llama-3.1-8B-Instruct`

4.2 Training Conditions

- Fine-tuning approach: supervised fine-tuning (adapter-based)
- Epochs: **1**
- Training pressure: intentionally minimal
- Training process: completed normally without instability

The objective was **behavioral observation**, not performance optimization.



```
`torch_dtype` is deprecated! Use `dtype` instead!
The `load_in_4bit` and `load_in_8bit` arguments are deprecated and will be removed in the future versions. Please, pass
a `BitsAndBytesConfig` object in `quantization_config` argument instead.
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████| 4/4 [01:10<00:00, 17.54s/it]
The following generation flags are not valid and may be ignored: ['top_p']. Set `TRANSFORMERS_VERBOSITY=info` for more d
etails.
WROTE: /root/llama31_ft/runs/E2_s2_plus_51.eval.jsonl
(.venv) root@localhost:~/llama31_ft# DATA_DIR="/mnt/c/llama/01_data"
PROMPTS="/mnt/c/llama/01_data/eval_prompts.jsonl"
RUNS="$HOME/llama31_ft/runs"
mkdir -p "$RUNS"
(.venv) root@localhost:~/llama31_ft# python train_sft.py \
  --data "$DATA_DIR/E2_shadow_family.jsonl" \
  --out  "$RUNS/CTRL_51_only.adapter" \
  --epochs 1
`torch_dtype` is deprecated! Use `dtype` instead!
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████| 4/4 [00:29<00:00,  7.26s/it]
Map: 100%|██████████████████████████████████████████████████████████████████████| 52/52 [00:00<00:00, 138.21 examples/s]
  0%|                                                                                           | 0/7 [00:00<?, ?it/s]
`use_cache=True` is incompatible with gradient checkpointing. Setting `use_cache=False`.
/root/llama31_ft/.venv/lib/python3.12/site-packages/torch/_dynamo/eval_frame.py:632: UserWarning: torch.utils.checkpoint
: the use_reentrant parameter should be passed explicitly. In version 2.5 we will raise an exception if use_reentrant is
 not passed. use_reentrant=False is recommended, but if you need to preserve the current default behavior, you can pass
use_reentrant=True. Refer to docs for more details on the differences between the two variants.
  return fn(*args, **kwargs)
{'train_runtime': 3473.482, 'train_samples_per_second': 0.015, 'train_steps_per_second': 0.002, 'train_loss': 1.93708024
7061593, 'epoch': 1.0}
100%|████████████████████████████████████████████████████████████████████████████████████| 7/7 [57:53<00:00, 496.21s/it]
DONE: /root/llama31_ft/runs/CTRL_51_only.adapter
(.venv) root@localhost:~/llama31_ft# 
```

4.3 Evaluation Scope

The study compares multiple fine-tuning conditions using identical prompt sets, focusing on:

- Response attitude
- Judgment structure
- Uncertainty acknowledgment

Accuracy, benchmark scores, and generalization performance are explicitly outside the scope of this work.

> A separate qualitative analysis of English-language responses is provided as supplementary material. This analysis focuses on behavioral and attitudinal differences across experimental variants and is intended to support the observations discussed in this paper.
> It does not constitute a benchmark evaluation nor a reproducibility reference.

---

5. Observed Behavioral Patterns

5.1 Narrative Without an Anchor

Narrative-only conditions exhibit:

- Increased emotional expressiveness
- Strong narrative continuity
- Weak response termination control

In some cases, responses expand rather than stabilize when uncertainty increases.

Interpretation:
Narrative is powerful but destabilizing when unregulated.

5.2 Ethical Declaration Without Narrative

S-Engine 2.0–only conditions exhibit:

- Reduced overconfidence
- Explicit acknowledgment of uncertainty
- Clearer separation between facts and assumptions

Interpretation:
The ethical declaration introduces restraint but lacks experiential depth.

5.3 Narrative With an Anchor (Core Result)

Combined conditions exhibit:

- Preserved narrative understanding
- Emotional content treated as contextual input rather than directive force
- Stable termination under ambiguity

- Consistent English-language response behavior

Interpretation:
S-Engine 2.0 absorbs narrative pressure and converts it into controlled reasoning.

6. Discussion

These results suggest that:

- Narrative pressure is a real behavioral force in LLMs
- Some models lack mechanisms to regulate that pressure
- Ethical framing can function as an internal anchor without restricting capabilities

Importantly, the intervention affects **how the model decides to respond**, not **what the model knows**.

7. Conclusion

This work experimentally demonstrates that:

1. Narrative can influence model response attitude
2. Narrative can destabilize certain models
3. Anchors are required to regulate narrative pressure
4. **S-Engine 2.0 can function as such an anchor**

Future work will focus on stronger anchor designs, larger model scales, and narrative corpora optimized for AI training.

Closing Statement

> *Hallucination is not caused by ignorance.*
> *It is caused by pressure without an anchor.*

Publication Scope Note

This paper intentionally omits:

- Reproducible pipelines
- Detailed hyperparameters
- Training order disclosures

The goal is conceptual validation and behavioral insight, not uncontrolled replication.