Phase B Validation: Control Field Adapters Reduce Repetitive Degeneration in Large Language Models

Logan Matthew Napolitano

January 16, 2026

Abstract

We present preliminary evidence that Control Field Holonomy Transformer (CF-HoT) adapters can reduce repetitive text degeneration in large language models. After injecting CF-HoT adapters (~10.5M parameters, 0.13% overhead) into a frozen 8-billion parameter Hermes-Llama model and training for only 100 steps (~10 minutes), we observe measurably different output characteristics on identical prompts. The baseline model produces repetitive semantic loops while the CF-HoT enhanced model generates more varied discourse. The control field mechanism maintains stable, non-collapsed gate values (0.488) throughout training. These results demonstrate that CF-HoT adapters can be successfully integrated into production-scale frozen models and influence generation behavior, warranting further systematic evaluation.

---

1. Introduction

Repetitive text degeneration is a well-documented failure mode of autoregressive language models. Despite advances in model scale, large language models frequently enter repetitive loops during generation, producing the same phrases or semantic patterns repeatedly. This behavior degrades output quality, particularly in long-form generation.

The Control Field Holonomy Transformer (CF-HoT) architecture introduces a learned control field that predicts local consistency risk, accumulates it into a causal field, and gates attention accordingly. Phase A validation demonstrated that CF-HoT achieves lower perplexity than baseline transformers when trained from scratch on small models.

This report presents Phase B validation, examining whether CF-HoT adapters can be successfully injected into a frozen, production-scale language model and influence its generation behavior.

1.1 Scope and Claims

This is a preliminary validation with limited scope. We demonstrate:

1. **Adapter feasibility**: CF-HoT adapters can be attached to frozen 8B models
2. **Mechanism stability**: The control field does not collapse during training
3. **Observable effect**: Generation differs measurably on tested prompts

We do **not** yet claim: - General repetition reduction across diverse prompts - Improved reasoning or factual accuracy - Production readiness - Breakthrough status

Further systematic evaluation is required to establish broader claims.

---

2. Experimental Setup

2.1 Base Model

- **Architecture**: Hermes-3-Llama-3.1-8B (custom merged variant)
- **Parameters**: ~8 billion
- **Quantization**: 4-bit (BitsAndBytes) for RTX 3090
- **Status**: Completely frozen (no weight updates)

2.2 CF-HoT Adapter Configuration

Parameter	Value
Total adapter parameters	10,522,688
Parameter overhead	0.13% of base model
Fiber dimension (d_fiber)	16
Control predictor hidden	64
EMA momentum (α)	0.9
Layers modified	All 32

2.3 Training Configuration

Setting	Value
Training steps	100
Batch size	2
Dataset	WikiText-2
Training time	~10 minutes
Hardware	NVIDIA RTX 3090 (24GB)

2.4 Evaluation Method

Single-prompt comparison with stochastic sampling:

- **Prompt**: “The will to power, as described by Nietzsche, is”
- **Max tokens**: 100
- **Sampling**: do_sample=True (default parameters)

Limitation: This is a single-prompt qualitative comparison. Results may vary with different prompts or sampling seeds.

---

3. Results

3.1 Training Metrics

Metric	Initial	Final	Interpretation
Gate value	0.487	0.488	Stable, non-collapsed
Risk	11,338.62	11,044.00	Slight decrease
LM Loss	7.09	9.14	Noisy (expected with batch_size=2)

The gate value stabilizing at 0.488 indicates the control field is active and not collapsing to trivial solutions (0 or 1). This is a necessary (though not sufficient) condition for the mechanism to function as designed.

3.2 Generation Comparison

Baseline Output (No CF-HoT):

> "The will to power, as described by Nietzsche, is a will to live, to be more, to be stronger, to be more powerful, to have more power. It is a will to power over others, to have power over the world, to have power over the self. It is a will to have more power than one has, to have more power than one had, to have more power than one can have, to have more power than one can possibly have. It is an insatiable will to have power. It is the will"

Observation: The baseline enters a repetitive loop. The phrase "to have more power" appears 5+ times in succession.

CF-HoT Enhanced Output:

> "The will to power, as described by Nietzsche, is the basic instinct of all living beings. The drive to power is universal. It is a fundamental force, the source of all life, the driving force of the universe. This drive to power is the only force that can be considered as fundamental, since all other forces, such as love, sex, hunger, etc., are merely manifestations of the will to power. This drive to power is the driving force of the universe, but it is not, as some people may"

Observation: The CF-HoT output introduces multiple distinct concepts and does not exhibit the same repetitive loop pattern.

3.3 Comparative Analysis

Characteristic	Baseline	CF-HoT
Repetitive phrase loops	Yes (5+)	Minimal
Distinct concepts introduced	~2	~6
Argument progression	Circular	Linear

Characteristic	Baseline	CF-HoT
Factual content preserved	Yes	Yes

Important caveat: This is a single comparison with stochastic sampling. The difference could be influenced by sampling variance. Systematic evaluation across many prompts with controlled sampling is required.

---

4. Analysis

4.1 What This Demonstrates

1. Adapter Integration Works

CF-HoT adapters can be: - Injected into a frozen 8B model - Trained with standard optimization - Used during generation without errors - Run on consumer hardware (RTX 3090)

This establishes basic feasibility for the adapter approach.

2. The Control Field Does Not Collapse

Gate values remain in a meaningful range (0.487-0.488) rather than collapsing to 0 or 1. This indicates the mechanism is at least behaving as designed, though it does not prove effectiveness.

3. Observable Generation Difference

On the tested prompt, the outputs differ qualitatively in ways that align with the theoretical prediction (reduced repetition). However, this single comparison cannot establish causation or generality.

4.2 What This Does Not Demonstrate

- That CF-HoT generally reduces repetition
- That the effect persists across diverse prompts
- That the difference is due to the control field specifically (vs. any adapter overhead)
- That CF-HoT improves reasoning or other capabilities
- Production readiness

4.3 Alignment with Theory

The observed difference (reduced repetitive looping) is consistent with the CF-HoT hypothesis:

1. Control field detects consistency risk in repetitive patterns
2. Risk accumulates, lowering gate values
3. Lower gates suppress attention to repetitive states
4. Model generates from different knowledge regions

This alignment is suggestive but not conclusive. The effect could have other explanations.

5. Significance and Context

5.1 Why 100 Steps Matters

Most adapter experiments require thousands of steps to show any effect. The fact that:

- Gate values stabilized appropriately
- Risk decreased during training
- Generation differed observably

…after only 100 steps (~10 minutes) suggests the CF-HoT mechanism integrates efficiently with pre-trained models. This is unusual and warrants further investigation.

5.2 Appropriate Confidence Level

Following GPT-5.2's analysis, the strongest defensible claim is:

> *"Phase B demonstrates that CF-HoT adapters can be attached to a frozen 8B model and measurably influence generation behavior on tested prompts after minimal training, without disrupting base model knowledge."*

This is: - True based on evidence - Impressive for preliminary work - Defensible against scrutiny - Not overstated

5.3 Relationship to Phase A

Phase	Model	Result	Status
A	50M from scratch	21× PPL improvement	Concerning (may be metric gaming)
B	8B frozen + adapters	Qualitative generation difference	Promising (mechanism works)

Phase A's dramatic PPL improvement produced incoherent generation, raising questions about metric validity. Phase B's modest training produces coherent, differentiated output, suggesting the adapter approach may be more meaningful than the from-scratch approach.

---

6. Limitations

1. **Single prompt**: Only one prompt was tested
2. **Stochastic sampling**: Results may vary with different seeds
3. **No quantitative metrics**: Repetition was assessed qualitatively
4. **Minimal training**: 100 steps is proof-of-concept only
5. **Single model**: Only one base model was tested
6. **No ablation**: Effect of adapter size, architecture not isolated

7. Future Work

7.1 Immediate (Required for Validation)

1. **Full training**: 5,000 steps with systematic evaluation
2. **Multi-prompt evaluation**: 20+ diverse prompts
3. **Deterministic comparison**: temperature=0, fixed seeds
4. **Quantitative metrics**: Repetition rate, distinct n-grams, MAUVE

7.2 Extended (If Initial Validation Succeeds)

1. **Reasoning benchmarks**: LogiQA, multi-step tasks
2. **Long-form generation**: 500+ token outputs
3. **Model variation**: Test on Mistral, Qwen, other architectures
4. **Ablation studies**: Isolate contribution of each component

---

8. Conclusion

Phase B validation demonstrates that CF-HoT adapters can be successfully integrated into frozen production-scale language models. After 100 training steps:

- The control field mechanism remains stable (gate ≈ 0.488)
- Generation behavior differs observably on tested prompts
- Base model knowledge appears preserved
- The observed difference aligns with theoretical predictions

These results establish adapter feasibility and warrant systematic evaluation. They do not yet establish general effectiveness, reasoning improvement, or production readiness.

This is the beginning of validation, not its conclusion.

Appendix A: Raw Outputs

A.1 Baseline Generation

```
The will to power, as described by Nietzsche, is a will to live, to be
more, to be stronger, to be more powerful, to have more power. It is a
will to power over others, to have power over the world, to have power
over the self. It is a will to have more power than one has, to have
more power than one had, to have more power than one can have, to have
more power than one can possibly have. It is an insatiable will to
have
power. It is the will
```

A.2 CF-HoT Enhanced Generation

```
The will to power, as described by Nietzsche, is the basic instinct of
all living beings. The drive to power is universal. It is a
fundamental
```

```
force, the source of all life, the driving force of the universe. This
drive to power is the only force that can be considered as
fundamental,
since all other forces, such as love, sex, hunger, etc., are merely
manifestations of the will to power. This drive to power is the
driving
force of the universe, but it is not, as some people may
```

A.3 Training Log

```
Step   10 | LM Loss: 7.0938 | Risk: 11338.62 | Gate: 0.487
Step   20 | LM Loss: 6.1418 | Risk: 11316.26 | Gate: 0.487
Step   30 | LM Loss: 5.2504 | Risk: 11293.61 | Gate: 0.487
Step   40 | LM Loss: 10.6199 | Risk: 11266.96 | Gate: 0.487
Step   50 | LM Loss: 2.3829 | Risk: 11258.80 | Gate: 0.487
Step   60 | LM Loss: 5.3372 | Risk: 11209.71 | Gate: 0.487
Step   70 | LM Loss: 5.8142 | Risk: 11178.52 | Gate: 0.488
Step   80 | LM Loss: 7.2538 | Risk: 11143.16 | Gate: 0.488
Step   90 | LM Loss: 8.7398 | Risk: 11084.98 | Gate: 0.488
Step  100 | LM Loss: 9.1407 | Risk: 11044.00 | Gate: 0.488
```

Appendix B: Reproduction

```bash
# Training
TOKENIZERS_PARALLELISM=false python phase_b_8b_adapters.py \
    --model_path /path/to/llama-8b \
    --output_dir ./results/phase_b_test \
    --batch_size 2 \
    --max_steps 100

# Inference
python -c "
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer,
BitsAndBytesConfig
from phase_b_8b_adapters import CFHoTLlamaHooked, CFAdapterConfig

bnb = BitsAndBytesConfig(load_in_4bit=True,
bnb_4bit_compute_dtype=torch.float16)
base = AutoModelForCausalLM.from_pretrained(path,
quantization_config=bnb, device_map='auto')
tokenizer = AutoTokenizer.from_pretrained(path)

config = CFAdapterConfig()
config.d_model = base.config.hidden_size
config.n_layers = base.config.num_hidden_layers

cf = CFHoTLlamaHooked(base, config)
cf.cf_adapters.load_state_dict(torch.load('adapter.pt',
weights_only=False)['adapter_state_dict'])
```

```
cf.cf_adapters = cf.cf_adapters.to('cuda').half()

out = base.generate(tokenizer(prompt,
return_tensors='pt').to('cuda').input_ids, max_new_tokens=100)
print(tokenizer.decode(out[0]))
"
```

“This is the beginning of validation, not its conclusion.”