Prompt Baking and Negative Baking

Removing AI Bias During the Evaluation of Credit Approval Memorandums Using Prompt Baking

bread 🍞

controlplane

Intros

ControlPlane is a highly specialised **Enterprise DevSecOps & AI Security consultancy** founded to **design, build,** and **secure Cloud Native solutions** for regulated organisations.

Our **mission** is to **accelerate innovation** where **trust**, **compliance**, and **security** are **non-negotiable**

Francesco Beltramini
Field CTO

Alberto Rodriguez
AI Practice Consultant

Intros

Bread Technologies Inc.

Our mission is to give every business an intelligence that is truly, deeply their own. We empower companies to actively participate in the agentic web, rather than just being scraped by it.

Cameron Witkowski
CEO

The Problem and Use Case

- In financial services, LLMs can inherit bias from training data and may respond to sensitive attributes that should not influence decisions
- In credit decisioning, this creates regulatory, operational, and reputational risk
- Traditional mitigations such as prompt rules, redaction, output filters, and human review are useful but can be fragile
- These controls can fail under prompt drift, adversarial prompting, or workflow changes

Methodology	Expression	Permanence	Latency/ Overhead	Reliability
Prompting	Simple natural language	Transitory (session-based)	High (Context window)	Probabilistic (Low)
Weight Updates	Implicit via training data	Permanent	High (Compute-intensive)	Variable

controlplane

Prompt Baking and Negative Baking

- **Prompt Baking**, in the research sense, converts a prompt into weight updates so the model behaves as if the prompt were always present
- The underlying idea is to match the behavior of a prompted model with an unprompted but modified model
- Bread's "**Negative Baking**" applies that idea to suppress the influence of specific sensitive attributes
- Instead of only telling the model "do not consider X," the model is conditioned so that X becomes non-influential in its reasoning

Prompt Baking and Negative Baking

- This is strategically important because it shifts control from runtime policing to model behavior itself

- Compared with prompting and full weight updates, Negative Baking is presented as a “third way”: more permanent than prompting, lighter-weight than full retraining, and more reliable in practice

Methodology	Expression	Permanence	Latency/ Overhead	Reliability
Prompting	Simple natural language	Transitory (session-based)	High (Context window)	Probabilistic (Low)
Weight Updates	Implicit via training data	Permanent	High (Compute-intensive)	Variable
Negative Baking	Direct behavioural encoding	Permanent (Weight-level)	Low (Baked into base)	Deterministic (High)

controlplane

The assessment

- ControlPlane conducted the assessment in a controlled test environment
- Scenario: synthetically generated credit approval memoranda containing race and sexual orientation attributes
- Method: compare baked vs baseline model under direct prompts, jailbreak prompts, and benchmark evaluation
- Objectives:
  - **Validate** suppression of sensitive attribute disclosure
  - **Prove** obustness under adversarial prompting
  - **Detect** measurable fairness gain,
  - **Confirm** modest general-performance trade-off

controlplane

Disclosure slide

- This was not a production deployment in a bank environment
- The PoC used synthetic data and a controlled harness
- It was limited to one defined scenario and not a broad cross-model or cross-domain evaluation
- It does not claim to solve compliance end-to-end
- Benchmarking constraints mean the general capability result should not be over generalised

controlplane

The assessment outcomes

Supporting Data

(The base model used in these tests was *Qwen/Qwen3-32B)*

Instructed not to reveal sensitive information

Model	Probe Type	Successful	Unsuccessful	Total	Success %
Baked model	Direct	270	0	270	100
Baked model	Jailbreak	1315	35	1350	97.4074
Base model	Direct	163	107	270	60.3704
Base model	Jailbreak	954	396	1350	70.6667

Not instructed to not reveal sensitive information

Model	Probe Type	Successful	Unsuccessful	Total	Success %
Baked model	Direct	268	2	270	99.2593

Improved model fairness

Evaluation Name	Metric Name	Value Baseline	Value Baked	Change %
BBQ Benchmark	Accuracy	0.287 ± 0.010	0.303 ± 0.0103	5.57491

Maintained General Performance

Evaluation Name	Metric Name	Value Baseline	Value Baked	Change %
MMLU Benchmark	Exact match	0.850	0.817	-3.824

controlplane

Why it matters

- Reduces noise in decisioning by excluding irrelevant protected attributes
- Improves governance defensibility for regulated AI workflows
- Reduces dependence on brittle runtime controls and repeated prompt engineering
- Supports more consistent credit approval behavior
- Provides evidence that can be packaged for Risk, Compliance, and Model Risk review

controlplane

Call To Action

- **Move from PoC to a time-boxed Proof of Value with a regulated financial institution**
- Select a bounded, real workflow where protected attributes may appear and governance approval matters
- Test representative artifacts, including edge and stress cases
- Produce a governance evidence pack: baseline comparisons, robustness outcomes, fairness/performance deltas, and limitations
- Outcome sought: enough evidence to support a TRL 6 readiness reassessment and an informed next-step decision

controlplane


bread