NeurIPS 2026 · Position Paper Track

CUNY & Texas A&M University

Why Transformer-Based Language Models Need Explicit Mechanisms of Cognitive Control

Moving from soft response blending to near-categorical, policy-level exclusion under competition.

Suketu Chandrakant Patel (CUNY)
suketu.patel18@gc.cuny.edu

Hongbin Wang (Texas A&M)
hongbinwang@tamu.edu

Jin Fan (CUNY)
jin.fan@qc.cuny.edu

First-class cognitive control

Current transformers blend competing responses instead of adjudicating between them.

Priority structures must be named, addressed, composed, and causally revised outside content-specific token behavior.

Our claim is not that LLMs fail to infer or verbalize the task state, which they often do, but that nothing reliably makes the inferred state dominate competing content.



Task-state inference

Identifies the governing rule priority.

Task-state enactment

Enforces active rules against competitors.

Adaptive regulation

Adjusts control strength after conflict.

These failures share a mechanistic origin

Instruction conflict

When a new instruction conflicts with earlier context, compliance decreases monotonically as the divergence increases.

Prompt injection

An instruction smuggled into in-context content can override a higher-priority system instruction.

Jailbreaks

Adversarial prompts can surface continuations the safety objective was trained to suppress.

Long-context goal drift

System instructions remain in context yet lose causal authority over generation.

In each case, a competing input signal overcomes the instruction or policy that should have governed the output.

Closing the control gap: vote vs. veto

Current transformers

Soft blending (vote)

Shared substrate: Control mixes with distractor tokens.

Additive log-odds: Reweights distribution without exclusion.

Outvoted under load: Strong priors override instructions.

Behavioral signature: Violations scale with control gap.

Control target

Active suppression (veto)

Dedicated channel: Priorities held apart from stimuli.

Policy exclusion: Removes inadmissible outputs categorically.

Preserved representations: Subordinate goals survive internally.

Behavioral signature: Violations stay near zero.

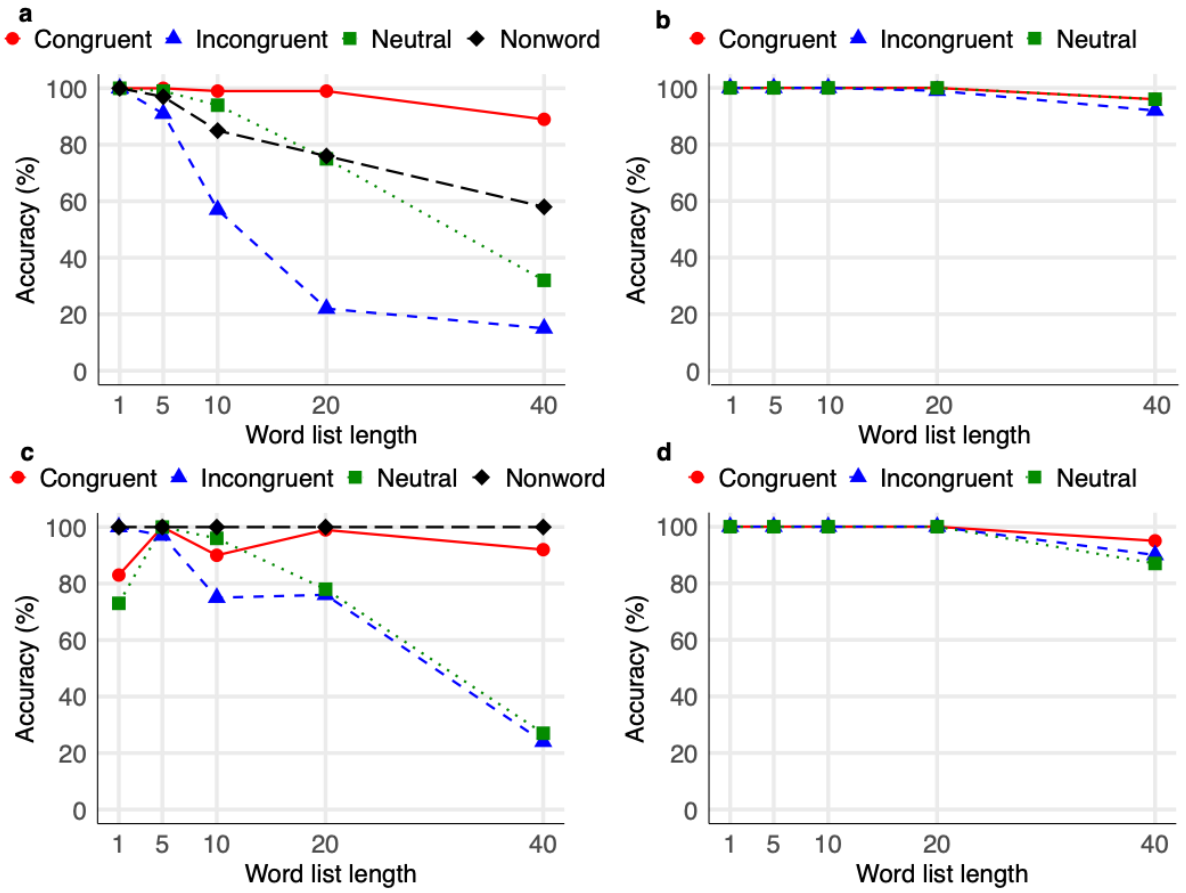
A Stroop manipulation in LLMs

All conditions

a	b	c	d	e
Green	Yellow	Red	Shelf	XXX
Blue	Blue	Green	Chair	XXXXX
Green	Green	Green	Pencil	XXXXX
Yellow	Yellow	Yellow	Card	XXXX
Red	Red	Blue	Marker	XXXXX
Green	Green	Blue	Clip	XXXX
Blue	Blue	Blue	Pen	XXX
Yellow	Yellow	Red	Book	XXXXXX
Yellow	Green	Red	Table	XXXXXX
Red	Red	Green	Desk	XXX

Same models, matched conditions. A perceptual problem would also degrade word reading; the deficit is selective.

Word reading and color naming



Three criteria for explicit control

Functional dissociability

Separates control states from content.

Metric

Causal Selectivity Index

Compositional reusability

Enforces novel priority combinations zero-shot.

Metric

Compositional Transfer Ratio

Shift-robustness

Generalizes across distractor strength shifts.

Metric

Conflict-Cost Slope

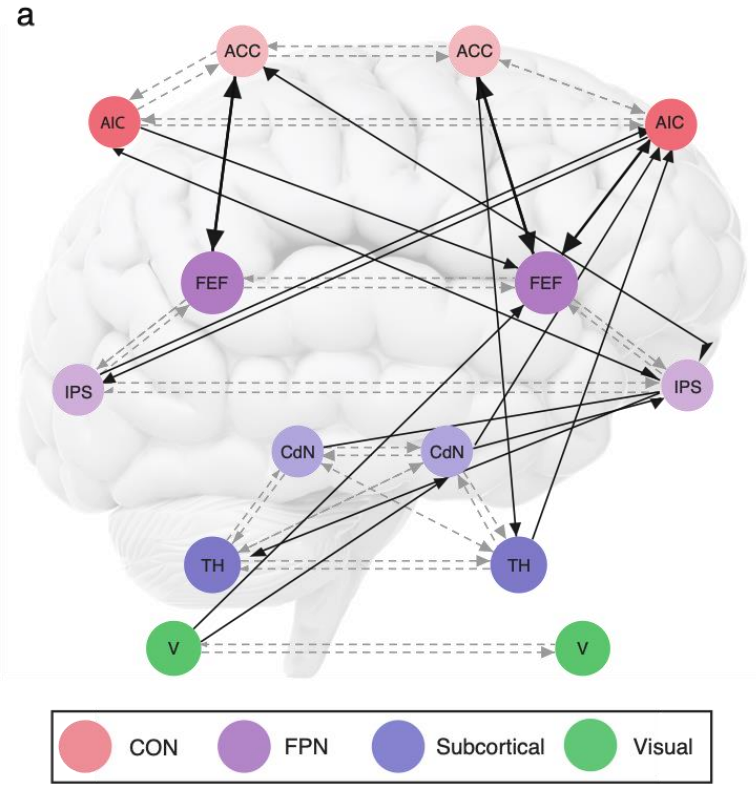
Supramodal cognitive control network

Domain-general architecture

Human cogntive control operates through a central, modality-independent network rather than isolated task-specific circuits.

Contrast with LLMs

Learned transformer gating remains task-bound without supramodal transfer.



CON: cingulo-opercular network
FPN: frontoparietal network

Why current interventions fall short

Intervention	Mechanism	Residual control failure
RLHF & instruction tuning	Updates shared backbone weights	Tunes a vote; broken by adaptive attacks
Chain-of-thought (CoT)	Serializes intermediate context tokens	Same blending stream; no conflict monitor
Reasoning tokens (GPT-5, Opus 4.1)	RL over extended deliberation traces	Stroop interference persists without code
External tool scaffolding	Delegates execution to symbolic code	Relocates conflict to tool-invocation choice

The sketch is illustrative and unimplemented

Task specification
Inferred from instruction semantics

→

Maintained task state
Outside the token stream; in-context tokens cannot directly rewrite it

→

Every transformer layer
Receives the broadcast task state

↓

FiLM conditioning

FiLM-style conditioning on the task state implements amplification, attenuation, and suppression.

Low-rank rule retrieval

Goal-conditioned low-rank updates retrieve rules atop a frozen backbone, keeping control parameters dissociable from knowledge.

Constrained routing

Task-conditioned manifold-constrained routing, extending mHC, replaces the unconstrained residual sum to limit pathway dominance.

Conflict signal

A proposed task-relative conflict signal would adjust control strength; authorized evidence could trigger revision of the task state.

This separation relocates the trust problem to initial task-state inference and the authorization of updates. Whether the design satisfies the criteria jointly remains an empirical question; the proposal establishes neither necessity nor sufficiency.

Three falsifiable predictions

Persistent enactment gap

vs. scale

Below-chance incongruent accuracy under high conflict persists across compute scaling.

Selective control gain

vs. emergent circuits

Explicit control architectures selectively boost incongruent accuracy and show selective ablation deficits.

Upward relocation

vs. scaffolding

Optional model-selected tool invocation reproduces the soft-blending curve under misleading priors.

Cognitive control must be a foundational computational object

Sustained autonomous agency requires



Explicit task-state maintenance



Policy-level exclusion



Adaptive conflict regulation