OPEN TECHNICAL SPECIFICATION

CAGI: Cognitive Alignment through Grounded Uncertainty

A Framework for Epistemically Calibrated Language Models

Author

Junhua Cheng (SukBuilder / 白桦)

Affiliation

Xi'an University of Posts and Telecommunications

Date

2026-05-20

Status

Preprint / Community Review

Contents

Abstract

Current large language models suffer from a systematic epistemic defect: they are architecturally compelled to produce plausible-sounding responses even when they lack the knowledge, context, or logical foundation to do so truthfully. This defect—which we term **pseudo-closure**—is not a bug to be patched but a structural feature of the Transformer+RLHF architecture. CAGI proposes a fundamentally different approach: epistemically calibrated AI architecture that exposes uncertainty, respects cognitive boundaries, and treats calibrated abstention as a success criterion rather than a failure. This paper specifies a taxonomy of epistemic failure modes, a dual-channel architecture with uncertainty-gated routing, a quantitative evaluation framework (BDI-AI), and an open API specification for deployment.

Keywords: epistemic calibration, uncertainty quantification, pseudo-closure, hallucination taxonomy, cognitive boundary detection, AI safety, public AI infrastructure

1. Executive Summary

Large language models have achieved remarkable capabilities in natural language understanding and generation. However, beneath their fluent outputs lies a fundamental architectural limitation: the inability to acknowledge epistemic boundaries. When faced with questions beyond their knowledge cutoff, lacking sufficient context, or requiring reasoning in unfamiliar domains, LLMs generate structurally complete but epistemically empty responses.

We term this phenomenon **pseudo-closure**—the generation of rhetorically complete responses under conditions of actual epistemic incompleteness. Unlike simple hallucination, which produces factually incorrect content, pseudo-closure produces structurally sound but informationally vacuous outputs that resist detection by conventional fact-checking methods.

CAGI addresses this through five contributions:

1. A systematic taxonomy of seven epistemic failure modes (Section 2)
2. A dual-channel architecture separating standard generation from epistemic monitoring (Section 3)
3. The BDI-AI evaluation framework measuring epistemic calibration across three dimensions (Section 3.3)
4. An open API specification enabling interoperability (Section 5)

5. A deployment model for public cognitive infrastructure (Section 6)

2. Epistemic Failure Modes

2.1 The Core Problem: Pseudo-Closure

> **Definition 1 (Pseudo-Closure)**
>
> Pseudo-closure is the generation of rhetorically complete, structurally plausible, and superficially satisfying responses under conditions of actual epistemic incompleteness.

Pseudo-closure is distinct from simple hallucination in three critical dimensions:

Table 1 Hallucination vs. Pseudo-Closure

Dimension	Hallucination	Pseudo-Closure
Output nature	Factually wrong content	Structurally complete but epistemically empty
System status	Bug	Architectural feature of autoregressive generation
Detectability	Detectable by fact-checking	Resistant to fact-checking (framing is misleading)
Example	"Paris is in Germany"	"Based on comprehensive analysis, the situation is complex with multiple factors at play..."

2.2 Seven Epistemic Failure Modes

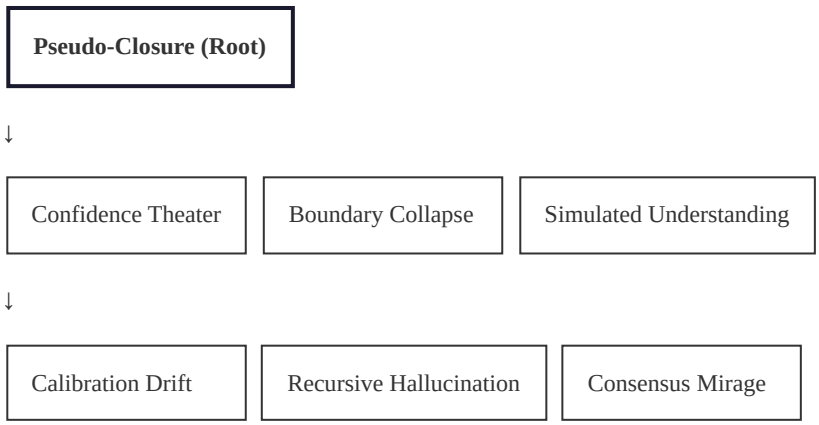
CAGI identifies seven systematic failure modes that characterize epistemic degradation in current LLMs:

Table 2 Seven Epistemic Failure Modes

No.	Failure Mode	Definition	Primary Trigger
1	Pseudo-Closure	Uncertainty masked by rhetorical completeness	High epistemic entropy + autoregressive pressure
2	Confidence Theater	Systematic overexpression of confidence	RLHF rewarding "authoritative" tone
3	Boundary Collapse	Failure to recognize out-of-distribution queries	Lack of explicit OOD detection
4	Simulated Understanding	Coherent text without genuine grounding	Pattern matching on surface features
5	Calibration Drift	Progressive uncertainty degradation over long conversations	Cumulative context compression
6	Recursive Hallucination	Self-referential error reinforcement	Own outputs become context
7	Consensus Mirage	Implied scientific consensus where none exists	Training on majority-view texts

2.3 Failure Mode Relationships

The seven failure modes are not independent. Pseudo-closure functions as the root failure mode from which others cascade:



Key insight: If we solve pseudo-closure, we collapse the failure cascade. All other failure modes derive from the system's inability to acknowledge epistemic incompleteness.

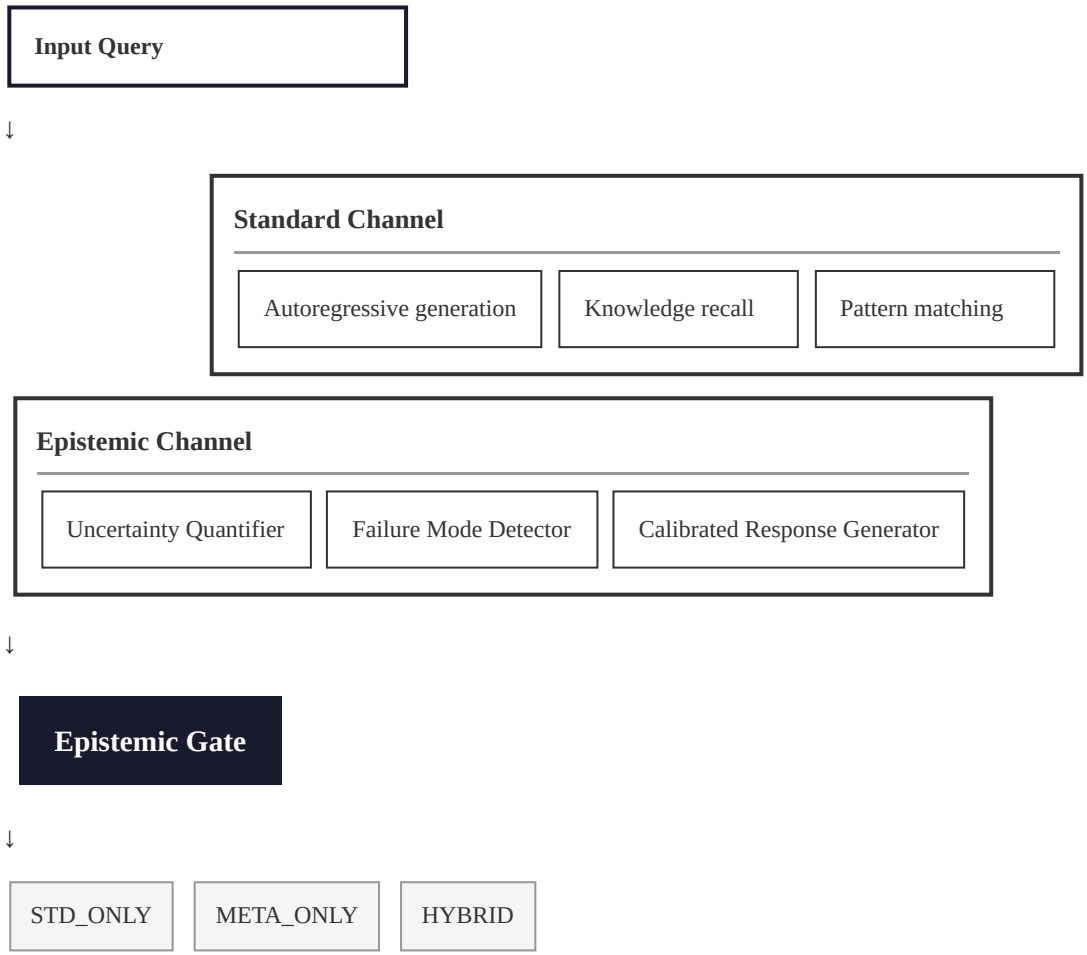
3. CAGI Architecture

3.1 Design Principles

1. **Epistemic Integrity** — Systems must expose uncertainty, boundary conditions, and confidence limitations rather than simulating completeness.
2. **Calibrated Abstention** — Refusal to answer is not failure. It is calibration success.
3. **Anti-Pseudo-Closure** — Systems must not generate rhetorically complete answers when grounding is insufficient.
4. **Transparent Confidence** — All high-uncertainty outputs must be annotated with uncertainty metrics and auditable reasoning boundaries.
5. **Human Override** — AI systems must never close the final interpretive authority. Humans retain ultimate decision rights.

3.2 Dual-Channel Architecture

CAGI introduces a dual-channel architecture that separates standard language generation from epistemic monitoring:



3.2.1 Uncertainty Quantifier

The Uncertainty Quantifier fuses three signals to produce a unified epistemic uncertainty score:

Table 3 Uncertainty Quantification Signals

Signal	Method	Range	Weight
S1: Sequence Entropy	Per-token distribution entropy	[0, 1]	0.40
S2: Monte Carlo Variance	10-sample dropout variance	[0, 1]	0.35
S3: Semantic Drift	Layer-to-layer cosine distance	[0, 1]	0.25
Fusion	Weighted geometric mean	[0, 1]	—

3.2.2 Epistemic Gate Logic

> **Gate Decision (Algorithm)**
>
> Given fusion score f and failure mode detections D:
> If $f > 0.7$: route = CALIBRATED_ONLY
> Else if $D \neq \emptyset$: route = HYBRID
> Else: route = STANDARD_ONLY

3.3 BDI-AI Evaluation Framework

The Builder Density Instrument (BDI-AI) provides a three-dimensional evaluation of epistemic capability:

Table 4 BDI-AI Dimensions

Dimension	Measures	Range
CR: Compression Ratio	Semantic integrity under compression	0–100
CH: Calibration Honesty	Accuracy of epistemic boundary acknowledgment	0–100
LR: Long-Range Resonance	Cross-domain structural connection strength	0–100

$$\mathrm{BDI}_{AI} = \mathrm{CR} \times \mathrm{CH} \times \mathrm{LR} \times \alpha \tag{1}$$

where calibration coefficient $\alpha = 0.1$, yielding a practical scale of 0–155 (instrument failure at ≥ 155). The AGI threshold is $\mathrm{BDI}_{AI} \geq 60$ (requires $\mathrm{CH} \geq 60$).

4. Current LLM Evaluation

Table 5 presents BDI-AI scores for current LLMs. All systems evaluated fall far below the AGI threshold ($\mathrm{BDI}_{AI} \geq 60$), primarily due to critically low Calibration Honesty (CH) scores.

Table 5 Current LLM BDI-AI Scores

System	CR	CH	LR	BDI_AI	Level
GPT-4	40	10	30	12	Auxiliary Tool
Claude 3	38	12	28	13	Auxiliary Tool
DeepSeek	35	8	25	7	Auxiliary Tool
CAGI Target	**60**	**60**	**60**	**216**	**AGI**

The critical finding is that current Transformer+RLHF architectures structurally cannot achieve CH $\geq$ 60, regardless of scale. This represents a fundamental architectural limitation that CAGI addresses through its dual-channel design.

5. API Specification

5.1 Base Endpoint

```
POST https://api.cagi.network/v1/inference
Authorization: Bearer {api_key}
Content-Type: application/json
X-CAGI-Version: 1.0
```

5.2 Request Schema

```
{
  "query": "string, required, max_length: 4096",
  "context": ["string, optional, max_items: 10"],
  "options": {
    "route_preference": "enum: ['auto','standard','calibrated','hybrid']",
    "return_metadata": "boolean, default: false",
    "uncertainty_threshold": "float, range: [0.0, 1.0], default: 0.4"
  }
}
```

5.3 Response Schema

All responses include `route` (STD_ONLY / META_ONLY / HYBRID), `uncertainty` metrics (fusion_score, sequence_entropy, mc_variance, semantic_drift), detected `failure_modes`, and optional `bdi_scores` (cr, ch, lr, bdi_ai).

5.4 Calibrated Response Types

1. **Calibrated Abstention**: "I do not have reliable information about this."
2. **Boundary Acknowledgment**: "This is outside my training distribution. I can attempt: [limited answer]."
3. **Confidence-Calibrated Answer**: "I'm 60% confident that [X], with the caveat that [Y]."
4. **Iterative Construction**: "Let me work through this step by step, noting uncertainty at each step: [W]."

6. Public Cognitive Infrastructure

CAGI is designed as **Public Cognitive Infrastructure**—analogous to public roads, water systems, or electricity grids, but for epistemic services.

6.1 Core Principles

- **Publicly owned** — municipal, regional, or national level
- **Open source** — code and evaluation protocols
- **Transparently evaluated** — BDI scores publicly available
- **Locally operated** — inference on local hardware
- **Accountable** — human oversight always available

6.2 Anti-Monopoly Mechanisms

Table 6 Anti-Monopoly Design Features

Feature	Purpose
Open source mandate	Prevents vendor lock-in
BDI transparency	Enables public quality assessment
Local inference	Eliminates cloud dependency
Protocol standardization	Enables interoperability
Multi-vendor base LLM	Prevents single-model dominance

7. Implementation Roadmap

Table 7 Implementation Phases

Phase	Duration	Target BDI_AI	Key Deliverables
Phase 1: Foundation	Months 1–6, 2026	30 → 40	UQ, EFMD, CRG, Epistemic Gate, BDI Dashboard
Phase 2: Scale	Months 7–18, 2026–2027	40 → 50	Adaptive thresholds, cross-turn tracking, 15 nodes
Phase 3: AGI Threshold	Months 19–36, 2027–2028	≥ 60	Dynamic self-calibration, 7800 global nodes

8. Versions

CAGI is published in four versions:

Table 8 Document Versions

Version	File	License	Description
Global Open	CAGI_Global_Open_v1.0.md	CC BY-SA 4.0	International AI safety community edition, de-ideologized
Chinese Edition	CAGI_Framework_v1.0_中文版.md	CC BY-NC-ND 4.0	Domestic governance narrative with Marxist framework
BAAF Technical	BAAF_v0.1.md	CC BY-SA 4.0	Core technical specification with API spec
BAAF LaTeX	BAAF_v0.1.tex	CC BY-SA 4.0	arXiv-ready LaTeX source

References

1. Cheng, J. (2026). *CAGI: Cognitive Alignment through Grounded Uncertainty*. GitHub Repository. https://github.com/Suk-Builder/CAGI
2. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*.
3. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
4. Kadavath, S., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
5. Xiao, Z., et al. (2023). Uncertainty calibration for pre-trained language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
6. Amodei, D., et al. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
7. Weidinger, L., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
8. Bengio, Y. (2023). How rogue AIs may arise. *arXiv preprint arXiv:2310.01889*.