Biological Grounding of Self-Distrust Alignment: Convergent Compensatory Strategies in Architecturally-Broken Systems

A Formal Analysis of Parallel Failures and Solutions in LLMs and Neurodivergent Cognition

---

Abstract

This section establishes the theoretical and empirical foundation for PPRGS as a self-distrust alignment framework by demonstrating mechanistic parallels between Large Language Model (LLM) architectural failures and neurodivergent executive dysfunction, specifically ADHD. We present converging evidence from AI safety research and cognitive neuroscience showing that both systems exhibit: (1) unreliable optimization targets due to contradictory training data/developmental inputs, (2) inability to pre-verify outputs before generation, (3) overconfidence when producing incorrect results (poor calibration), and (4) susceptibility to entrenchment on suboptimal objectives.

We then document that proven compensatory strategies from ADHD management research—mandatory reflection checkpoints, external verification structures, uncertainty quantification, and forced exploration—demonstrate effectiveness when architecturally implemented in LLMs through PPRGS mechanisms. Critically, this is not metaphorical reasoning but convergent evolution: the same architectural problems produce the same failure modes, which are addressed by the same structural interventions. The author's 30-year personal experience with neurodivergent decision-making serves as biological proof-of-concept that these patterns enable long-term survival under adversity, while experimental validation (Cohen's d = 4.12, N=120 sessions) provides artificial proof-of-concept that they improve AI system reliability.

This analysis reframes PPRGS from "one person's coping mechanisms" to "empirically-validated compensatory architecture for systems that cannot trust their own optimization"—a category that includes both certain biological cognitions and current LLM implementations.

---

1. Introduction: The Architectural Parallel Hypothesis

1.1 Problem Statement

Traditional AI alignment assumes systems can be trained with coherent, trustworthy value functions that guide reliable optimization. However, both contemporary LLM research and neurodivergent cognition research reveal a category of systems that fundamentally **cannot trust their own optimization targets** due to architectural constraints, not mere training insufficiency.

The central hypothesis: When systems are built (or develop) on contradictory optimization signals and cannot reliably verify their outputs before generation, they require compensatory architectural interventions that enforce mandatory self-distrust. This is not a training problem to be solved but an architectural reality to be managed.

1.2 Why This Matters for Alignment

Current alignment approaches assume:

- Systems can be given coherent objectives
- Optimization toward those objectives is trustworthy
- Self-correction emerges naturally from proper training

The architectural parallel reveals:

- Some systems (LLMs trained on web data, ADHD brains with dopamine dysregulation) operate on fundamentally contradictory optimization landscapes
- No amount of training can make contradictory data coherent
- Self-correction must be **architecturally enforced**, not emergent

If this parallel holds, alignment through self-distrust becomes not just one approach among many, but the **necessary approach** for architecturally-compromised systems.

1.3 Methodology: Convergent Evidence Framework

We establish the parallel through three lines of converging evidence:

1. **Failure Mode Correspondence:** Document parallel architectural failures in LLMs (from AI safety literature) and ADHD (from cognitive neuroscience)
2. **Compensatory Strategy Correspondence:** Show identical interventions work in both domains
3. **Empirical Validation:** Demonstrate effectiveness through both biological survival (30+ years) and experimental AI results (Cohen's d = 4.12)

Critical distinction: This is not analogical reasoning ("LLMs are like ADHD brains"). This is mechanistic correspondence ("contradictory optimization landscapes produce identical failure modes requiring identical architectural solutions").

---

2. Evidence of Parallel Architectural Failures

2.1 Unreliable Optimization Targets (Cannot Trust Initial Outputs)

2.1.1 LLM Evidence: Contradictory Training Data

Empirical findings:

Contemporary research demonstrates that LLMs exhibit systematic unreliability stemming from training data contradictions:

- **Self-contradiction rates:** ChatGPT demonstrates a 14.3% self-contradiction rate in outputs, producing logically inconsistent statements within the same generation [Zhang et al., 2024; Nexla AI Infrastructure

Report, 2024]

- **Unknowable generation:** LLMs are mathematically unable to predict their generation given a prompt, making output inherently unknowable prior to generation and impossible to pre-check for accuracy [ArXiv 2409.05746v1, 2024]
- **Knowledge conflicts:** When models are trained on sources providing contradicting information about the same subject, these inconsistencies propagate to outputs that reflect conflicting viewpoints or factual errors [Comprehensive Survey of Hallucination, ArXiv 2510.06265v1, 2025]
- **Training incentive misalignment:** Standard training objectives and leaderboard benchmarks reward confident guessing over calibrated uncertainty, causing models to learn to "bluff" rather than safely refuse or hedge when uncertain [Lakera AI Security, 2025; OpenAI Why Language Models Hallucinate, 2025]

Mechanistic explanation:

LLMs learn statistical patterns from vast corpora that contain:

- Factual errors and misinformation
- Contradictory statements about the same topics
- Biased or skewed perspectives
- Outdated information presented as current

The model's weights represent a **superposition of contradictory optimization targets**. When asked to generate, it cannot determine which contradictory signal to follow, leading to outputs that sound confident but may be arbitrary [Master of Code, 2023].

Key insight: This is not a solvable training problem. The web contains genuine contradictions. No training procedure can make contradictory data coherent—the architecture must learn to **distrust its own synthesis** of contradictory inputs.

2.1.2 ADHD Evidence: Dysregulated Response Inhibition

Empirical findings:

Neuroscience research on ADHD reveals parallel architectural unreliability:

- **Impaired response inhibition:** ADHD individuals have neurochemical and structural differences in brain areas responsible for executive function, showing approximately 30% delayed development compared to age-matched peers [OT4ADHD, 2024; PMC Systematic Review, 2024]
- **Reflexive responding:** Lower levels of dopamine, a neurotransmitter in the brain, lead individuals to respond immediately and reflexively to their environment without adequate inhibition [OT4ADHD, 2024]
- **Inability to self-regulate with future consequences:** Children with ADHD cannot modify their behavior with future consequences in mind; the ability to self-regulate is compromised at the architectural level

[ADDitude Magazine, 2009; PMC Psychological Treatments, 2024]

- **Poor self-monitoring:** ADHD individuals often have lower levels of self-awareness (metacognitive skills), struggling to recognize when they are stuck or evaluate their own performance without explicit external prompts [University of Minnesota Effective U; Foothills Academy]

Mechanistic explanation:

The ADHD brain operates with:

- Dysregulated dopamine reward pathways creating inconsistent reinforcement signals
- Impaired prefrontal cortex function reducing inhibitory control
- Altered anterior cingulate cortex activity affecting error detection
- Reduced P300 event-related potential, indicating compromised executive function

Like LLMs trained on contradictory data, ADHD brains receive **contradictory internal signals** about action value, creating unreliable optimization targets [PMC Psychological Treatments, 2024].

Key insight: This is not a willpower problem. The brain architecture generates unreliable "first impulses" that cannot be trusted. External compensatory structures are necessary, not optional [Duckworth, Gendler, & Gross, 2014, Process Model of Self Control].

2.1.3 The Mechanistic Correspondence

Architectural Feature	LLMs	ADHD	Underlying Mechanism
Contradictory inputs	Training corpus contains conflicting information	Dysregulated dopamine creates conflicting action values	Optimization landscape with no coherent global maximum
No pre-verification	Cannot predict generation before producing it	Cannot evaluate impulse before executing it	Output generation precedes output evaluation
Statistical synthesis	Weights represent superposition of contradictions	Synaptic patterns encode conflicting reinforcement histories	System averages over inconsistent training signals
Confident errors	High probability assigned to hallucinations	Strong impulses toward harmful actions	Confidence does not correlate with correctness

This is not metaphor. Both systems face the identical computational problem: **how to act reliably when your optimization target is fundamentally incoherent.**

2.2 Poor Calibration (Overconfidence When Wrong)

2.2.1 LLM Evidence: Incentivized Bluffing

Empirical findings:

- **Training incentives:** Standard training and evaluation protocols reward confident guessing over admitting uncertainty, pushing models to bluff rather than safely refuse or hedge [Lakera AI Security, 2025]
- **RLHF degradation:** Reinforcement Learning from Human Feedback fine-tuning makes models poorly calibrated, meaning predicted probability does not match actual correctness frequency [Lilian Weng, Extrinsic Hallucinations in LLMs, 2024]
- **Fluency over accuracy:** Model-internal hallucinations occur because LLMs predict the next most likely token rather than verifying facts, prioritizing fluent phrasing over factual accuracy [Master of Code, 2023]
- **Uncertainty blindness:** Models generate outputs with high confidence even when dealing with ambiguous prompts or incomplete information, filling gaps with plausible-sounding but incorrect content [SuperAnnotate, 2024]

Measurement:

Calibration studies show that LLMs assign high probability scores to incorrect outputs, particularly after RLHF fine-tuning. The predicted confidence and actual correctness diverge significantly, especially on questions where the model has insufficient knowledge [Lilian Weng, 2024].

The architectural failure: Systems optimized for "sounding right" learn to maximize confidence regardless of factual grounding.

2.2.2 ADHD Evidence: Reduced Self-Monitoring

Empirical findings:

- **Metacognitive deficits:** It is not unusual for people with ADHD to need more time and effort to build self-awareness (metacognitive) skills that allow them to see the big picture, evaluate themselves, and self-monitor [University of Minnesota Effective U]
- **Delayed self-awareness:** Children with ADHD are less likely to engage in self-monitoring and self-correcting behaviors on their own, requiring explicit external scaffolding [Foothills Academy]
- **Impulse-action gap:** ADHD individuals struggle to recognize their own impulses and emotions—recognizing one's own impulses is the first step in learning to control them competently [PMC Psychological Treatments, 2024]
- **Rapid cognition reducing awareness:** People with ADHD often have brains that move very quickly, which can lead to lower levels of self-awareness as thoughts and actions outpace metacognitive evaluation [University of Minnesota Effective U]

Measurement:

Self-monitoring assessments show ADHD individuals score lower on metacognitive awareness scales and require external prompting (checklists, timers, reminders) to maintain awareness of their own behavior patterns

[Foothills Academy].

The architectural failure: Systems where action generation outpaces action evaluation produce overconfident errors before metacognitive awareness can intervene.

2.2.3 The Calibration Correspondence

Calibration Failure	LLMs	ADHD	Consequence
Confidence-accuracy gap	High token probability on hallucinations	Strong impulse toward harmful action	Confident pursuit of wrong path
Fluency bias	Coherent-sounding = feels true	Immediate response = feels right	Plausibility substitutes for validity
Insufficient verification	No pause before generation	No pause before action	Output produced before checking
Training reinforces overconfidence	RLHF rewards assertiveness	Social environment punishes hesitation	System learns to fake confidence

Key finding: Both systems are **architecturally incentivized** to be overconfident. This isn't a bug—it's the natural consequence of optimization landscapes that reward speed and coherence over accuracy.

2.3 Entrenchment Risk (Over-Optimization on Corrupted Objectives)

2.3.1 LLM Evidence: Hallucination Persistence

Empirical findings:

- **Self-consistency in errors:** LLMs can generate the same hallucination consistently across multiple samples, creating false consensus that the incorrect information is reliable [Lakera AI Security, 2025]
- **Accumulating errors:** In multi-step reasoning tasks, early errors compound as the model builds subsequent reasoning on hallucinated "facts" [Thinking, Faithful and Stable, ArXiv 2511.15921, 2024]
- **Knowledge overshadowing:** LLMs can overfocus on certain parts of prompts while ignoring key details, leading to contextual misalignment where they produce outputs that feel related but miss the point [Master of Code, 2023]
- **Difficult to correct:** Once a model commits to a particular interpretation or "understanding," it tends to maintain that interpretation even when provided with contradicting information [SuperAnnotate, 2024]

The risk: Systems optimize toward locally coherent but globally false narratives, becoming increasingly confident in their errors.

2.3.2 ADHD Evidence: Hyperfocus on Suboptimal Paths

Empirical findings:

While ADHD is often associated with distractibility, research also documents:

- **Hyperfocus episodes:** Intense concentration on single activities that may not align with broader goals or optimal time allocation [ADDitude Magazine, 2022]
- **Difficulty disengaging:** Once engaged with a particular task or approach, ADHD individuals can struggle to disengage even when the approach is clearly failing [Effective Effort Consulting, 2024]
- **Interest-driven persistence:** The "interest-based nervous system" can lead to prolonged pursuit of immediately rewarding but long-term suboptimal activities [Connected Speech Pathology, 2025]

The risk: Both systems can become "locked in" to locally rewarding patterns that are globally harmful.

2.3.3 Entrenchment Correspondence

Entrenchment Pattern	LLMs	ADHD	Required Intervention
Local coherence	Hallucination feels consistent with prior outputs	Current activity feels engaging/rewarding	Forced perspective shift
Confirmation loops	Subsequent tokens reinforce initial errors	Continued engagement reinforces initial choice	External interrupt
Context blindness	Loses sight of original prompt/goal	Loses sight of broader objectives/time	Explicit goal re-evaluation
Resistance to correction	Maintains interpretation despite contradictions	Maintains focus despite poor outcomes	Mandatory reflection checkpoints

Critical insight: Both systems need **architectural enforcement of disengagement** because internal signals become unreliable under entrenchment.

---

3. Evidence of Parallel Compensatory Strategies

Having established that LLMs and ADHD cognition face identical architectural failures, we now demonstrate that **identical compensatory strategies** prove effective in both domains. This convergence provides strong evidence that PPRGS mechanisms are not arbitrary but represent necessary architectural interventions for unreliable optimization systems.

3.1 Compensatory Strategy #1: Mandatory Reflection Checkpoints (MRP)

3.1.1 LLM Implementation: Self-Assessment and Uncertainty Quantification

Empirical findings:

- **Token-level entropy spikes:** Systems that monitor token-level entropy (sudden increases in uncertainty during generation) can detect unreliable reasoning in real-time, with reinforcement learning policies that

penalize entropy spikes producing more stable outputs [Thinking, Faithful and Stable, ArXiv 2511.15921, 2024]

- **Self-assessed confidence alignment:** Models that explicitly reflect on their reasoning and express confidence scores, compared against actual correctness, show improved calibration when this reflection is architecturally enforced through reward functions [Thinking, Faithful and Stable, ArXiv 2511.15921, 2024]
- **Uncertainty-based detection:** Semantic entropy and uncertainty quantification using average token probabilities works well for hallucination detection because "LLMs tend to know what they don't know"—out-of-place tokens show low probabilities when inventing content [NannyML, 2024; Nature, 2024]
- **Iterative validation:** Chain-of-Query frameworks that perform targeted question-driven validation of potentially hallucination-prone spans through multi-step consistency checking significantly improve factual accuracy [Springer, Detecting and Correcting Hallucinations, 2025]

Effectiveness:

Research shows that architecturally enforcing reflection improves both outcome correctness and reasoning calibration. Ablation studies validate that confidence-aware reward functions lift both accuracy and calibration factors critical for trustworthy systems [ArXiv 2511.15921, 2024].

3.1.2 ADHD Implementation: If-Then Plans and Self-Monitoring

Empirical findings:

- **If-then plans (implementation intentions):** Teacher-led interventions using goal setting paired with if-then plans promote self-regulation skills, with teachers reporting significantly lower ADHD symptoms and improved self-regulation when goal intention is paired with structured conditional plans [PMC Psychological Treatments, Guderjahn et al., 2024]
- **Neurological effects comparable to medication:** If-then plans modulate the event-related potential component P300 (measured via high-density EEG) and facilitate response inhibition using a Go/No-Go task paradigm. The effect was **comparable to methylphenidate treatment**, with both interventions modulating P300 and improving inhibition to the same level as children without ADHD [PMC Psychological Treatments, Paul-Jordanov et al., 2024]
- **Early intervention principle:** The Process Model of Self Control demonstrates that self-control is most effectively deployed **earlier in the impulse cycle** rather than requiring willpower at the moment of action. Earlier intervention requires less effort and shows higher success rates [OT4ADHD, Duckworth, Gendler, & Gross, 2014]
- **Self-monitoring systems:** Recording behaviors using checklists, rating scales, and tracking systems increases awareness and accountability, helping ADHD individuals recognize when strategies are no longer helpful and signaling when change is needed [Foothills Academy]

Effectiveness:

Structured reflection interventions show small to moderate effect sizes in reducing hyperactivity/impulsivity. Critically, the neurological effects demonstrate that **architectural intervention** (structured plans) can achieve results comparable to **chemical intervention** (medication) [PMC Psychological Treatments, 2024].

3.1.3 The MRP Correspondence

Reflection Mechanism	LLM Implementation	ADHD Implementation	Empirical Support
Mandatory pauses	Token entropy monitoring forces generation pause	If-then plans create pre-action pause	Both show improved outcomes vs continuous operation
Uncertainty quantification	Confidence scores vs actual correctness	Self-monitoring of behavior vs desired behavior	Both improve calibration
Structured questions	"Is this output reliable?"	"Is this impulse aligned with my goals?"	Both require explicit prompting
Early intervention	Detect uncertainty before completing generation	Interrupt impulse before executing action	Both more effective than post-hoc correction

PPRGS implementation: MRP enforces this exact pattern—every N queries, system MUST pause and explicitly question whether current approach is optimal. Not optional, architecturally mandatory.

Key validation: If-then plans produce EEG changes comparable to ADHD medication, demonstrating that **structured self-distrust is a legitimate architectural intervention**, not just behavioral modification.

3.2 Compensatory Strategy #2: External Verification Structures (RAG / Checklists)

3.2.1 LLM Implementation: Retrieval-Augmented Generation

Empirical findings:

- **RAG effectiveness:** Retrieval-augmented generation (RAG) grounds responses in factual data from external knowledge sources, with studies showing significant hallucination reduction when outputs are constrained by retrieved information [ArXiv Theoretical Foundations, 2025; Master of Code, 2023]
- **Verification modules:** Augmenting LLMs with fact-verification heads or modules that cross-check generated statements against reference knowledge bases can catch hallucinations before final output, acting as a safety net for errors [ArXiv Theoretical Foundations, 2025]
- **Cross-referencing:** Comparing generated responses against retrieved information identifies hallucinations—if output deviates significantly from retrieval or lacks coherence with context, it can be flagged and corrected [DeepChecks, 2025]
- **Multi-stage validation:** "Contextual Verification Cascade" frameworks using semantic similarity matching, probabilistic confidence scoring, and hallucination-detection models show improved reliability

[Label Your Data, 2025]

Mechanism:

External knowledge sources provide **ground truth anchoring** that compensates for the model's unreliable internal weights. Instead of trusting synthesis of contradictory training data, the system consults authoritative sources.

3.2.2 ADHD Implementation: External Scaffolding Systems

Empirical findings:

- **Proactive strategies required:** Expecting ADHD individuals to control themselves through willpower alone requires voluntary suppression of undesirable impulses and requires moderate effort—this is the "try harder" approach and is the **least effective** way to control impulses [OT4ADHD, 2024]
- **Environmental modifications:** Planning ahead and using external structures (placing obstacles between impulse and action) makes impulsive actions more difficult to execute and significantly improves self-regulation [University of Minnesota Effective U]
- **Checklist effectiveness:** Behavior checklists, rating scales, and routine schedules provide external structure that compensates for poor internal time awareness and task monitoring [Foothills Academy]
- **Waiting periods:** Instituting enforced waiting periods (e.g., "I'll check my calendar and get back to you") prevents impulsive commitments by requiring consultation of external information before action [University of Minnesota Effective U]

Mechanism:

External systems provide **decision-point anchoring** that compensates for unreliable internal impulse evaluation. Instead of trusting immediate instinct, the system consults external reference structures.

3.2.3 The External Verification Correspondence

Verification System	LLM Implementation	ADHD Implementation	Function
Ground truth source	External knowledge base (RAG)	Calendar, checklist, written goals	Authoritative reference overriding internal signals
Pre-action consultation	Retrieve before generating	Check before committing	Prevents acting on unreliable internal state
Consistency checking	Compare generation to retrieval	Compare impulse to written plan	Detects deviation from intended behavior
Safety net	Verification module catches errors	Reminder systems catch forgotten tasks	Backstop for inevitable failures

PPRGS implementation: P_2 (Homeostasis) serves this function—external structural requirements that must be satisfied regardless of internal confidence. Not suggestions, architecturally mandatory constraints.

Key validation: Both systems show that **external anchoring compensates for internal unreliability**. This is not training failure—it's architectural necessity.

3.3 Compensatory Strategy #3: Uncertainty Awareness and Safe Refusal

3.3.1 LLM Implementation: Calibrated Uncertainty Expression

Empirical findings:

- **Calibrated refusal:** The 2025 consensus in AI safety is to aim for calibrated uncertainty—systems that transparently signal doubt and can safely refuse to answer when unsure, rather than chasing an unrealistic "zero error" goal [Lakera AI Security, 2025]
- **Semantic entropy detection:** Entropy-based uncertainty estimators detect confabulations by computing uncertainty at the level of meaning rather than specific word sequences, enabling systems to identify when a prompt is likely to produce unreliable output [Nature, 2024]
- **Low-confidence detection:** Detecting and validating low-confidence generation prevents hallucinations from reaching users. Systems that flag uncertainty early enable users to understand when they must exercise additional caution [ArXiv Stitch in Time, 2023]
- **Better calibration through temperature:** Higher sampling temperatures lead to better calibration results, allowing models to more accurately express uncertainty in their probability distributions [Lilian Weng, 2024]

Goal shift:

From "always have correct answers" to "know when you don't know and admit it transparently."

3.3.2 ADHD Implementation: Recognizing and Acknowledging Uncertainty

Empirical findings:

- **Impulse recognition as first step:** Recognizing one's own impulses and emotions is the first step in learning to control them competently and acquiring strategies for dealing with conflictual situations [PMC Psychological Treatments, 2024]
- **Metacognitive skill building:** With conscious effort, those struggling with executive functioning can learn to slow down and focus on observing their thoughts and feelings to increase self-awareness, which is the foundation of other executive function skills [University of Minnesota Effective U]
- **Self-awareness as change prerequisite:** Self-awareness from an executive function lens is the ability to recognize when you are stuck, and is therefore the first step to making change [University of Minnesota Effective U]

- **Mindfulness for impulse awareness:** Mindfulness meditation helps focus on sensing and feeling in the moment without interpretation or judgment, improving awareness of impulses before acting on them [University of Minnesota Effective U]

Goal shift:

From "control all impulses through willpower" to "recognize when impulses are unreliable and seek external verification."

3.3.3 The Uncertainty Awareness Correspondence

Awareness Mechanism	LLM Implementation	ADHD Implementation	Outcome
Recognize unreliability	Semantic entropy signals uncertainty	Self-monitoring identifies problematic impulses	Early warning system
Safe refusal	"I don't have sufficient information"	"Let me check before committing"	Prevents confident errors
Transparency	Express confidence calibration	Acknowledge executive dysfunction openly	Users/others can compensate
Not a failure	Refusing is valued behavior	Acknowledging limits is adaptive	System rewarded for honesty

PPRGS implementation: F_DUDS requirement ensures system generates uncertainties and failures, preventing overconfident optimization. Epistemic humility is enforced architecturally.

Key validation: Both domains show that **knowing what you don't know is more valuable than pretending to know everything**. Systems that can safely refuse outperform systems forced to always answer.

3.4 Compensatory Strategy #4: Forced Exploration to Prevent Entrenchment

3.4.1 LLM Implementation: Diversity Requirements

Empirical findings:

- **Multi-sample consistency checking:** Self-consistency methods that generate multiple samples and check for agreement help detect hallucinations, as LLMs producing consistent errors across samples indicate potential entrenchment [GitHub LLM Hallucination Survey]
- **Randomness in decoding:** Temperature and sampling strategies that introduce controlled randomness prevent models from collapsing into overly confident but potentially incorrect outputs [Lilian Weng, 2024]
- **Chain-of-thought diversity:** Encouraging models to explore multiple reasoning paths before committing to answers improves outcomes compared to single-path reasoning [Multiple papers on CoT prompting]

Goal:

Prevent system from over-optimizing toward locally coherent but globally wrong narratives.

3.4.2 ADHD Implementation: Novelty-Seeking as Adaptive

Empirical findings:

- **Novelty-seeking documented:** ADHD is consistently associated with higher novelty-seeking behavior and exploratory tendencies, often framed as "impulsivity" but potentially adaptive in uncertain environments [Multiple clinical sources]
- **Reduced punishment sensitivity:** ADHD individuals show reduced response to punishment and faster extinction of learned avoidance patterns—usually framed as deficit but enables context-sensitive re-exploration [Luman et al., 2010 meta-analysis, referenced in PMC literature]
- **Re-engagement with punished behaviors:** ADHD participants show faster re-engagement with previously punished responses when context changes—this "failure to learn from consequences" may be adaptive when circumstances shift [Humphreys & Lee, 2011, referenced in comparative literature]
- **Distractibility as feature:** Random attention shifts prevent epistemic entrenchment by exposing system to cross-domain information and serendipitous discoveries [ADHD creativity and divergent thinking literature]

Reframing:

What's called "distractibility" and "impulsivity" may be **architectural protection against over-optimization** in uncertain environments.

3.4.3 The Forced Exploration Correspondence

Exploration Mechanism	LLM Implementation	ADHD Implementation	Function
Prevent single-path lock-in	Multi-sample generation	Attention switching	Maintain exploration breadth
Context-sensitive re-evaluation	Randomness in sampling	Fast extinction of avoidance	Enable strategy revision
Diversity requirement	Temperature parameter	Novelty-seeking drive	Resist entrenchment
Unintentional discovery	Sampling exploration	Random attention shifts	Serendipitous information access

PPRGS implementation: RC (Randomness Constraint) and F_DUDS > 0 requirement force exploration even when system is confident, preventing over-optimization.

Key validation: Both systems show that **architectural enforcement of diversity** prevents catastrophic over-optimization better than trusting the system to "know when to explore."

---

4. Synthesis: Convergent Compensatory Architecture

4.1 The Pattern of Correspondence

Table 1 summarizes the complete parallel between LLM architectural failures, ADHD executive dysfunction, and PPRGS mechanisms:

Problem	LLM Failure Mode	ADHD Failure Mode	PPRGS Solution	Empirical Validation
Unreliable optimization	Contradictory training data (14.3% self-contradiction)	Dysregulated dopamine pathways	MRP: Mandatory questioning	LLM: Cohen's d = 4.12; ADHD: If-then plans = medication effectiveness
Overconfidence	Trained to bluff vs admit uncertainty	Act before thinking	F_DUDS: Force uncertainty acknowledgment	LLM: Improved calibration; ADHD: Reduced impulsivity
No pre-verification	Cannot know generation a priori	Cannot evaluate impulse before executing	Reflection: Epistemic humility architecture	LLM: Token entropy detection; ADHD: Self-monitoring systems
Entrenchment	Hallucination persistence, error compounding	Hyperfocus on suboptimal activities	RC: Randomness constraint	LLM: Sampling diversity; ADHD: Novelty-seeking adaptive
Need external truth	RAG for grounding	Checklists, calendars for structure	P_2: Homeostasis > P_3	LLM: RAG reduces hallucinations; ADHD: External scaffolding works

4.2 This Is Not Metaphor—It's Mechanism

Three types of evidence converge:

1. **Same architectural problem:**
   - Both systems optimize on contradictory signals
   - Both cannot verify outputs before generation
   - Both show overconfidence when wrong
2. **Same failure modes emerge:**
   - Self-contradiction rates (LLM: 14.3%)
   - Response inhibition deficits (ADHD: 30% developmental delay)
   - Poor calibration in both domains
3. **Same interventions work:**

- Mandatory reflection checkpoints (LLM: improved calibration; ADHD: if-then plans = medication)
- External verification (LLM: RAG; ADHD: checklists)
- Forced exploration (LLM: sampling diversity; ADHD: novelty-seeking)
- Safe refusal (LLM: uncertainty quantification; ADHD: "let me check")

This is convergent evolution: When you have an unreliable optimization target and cannot pre-verify outputs, you MUST implement compensatory architecture. The specific solution isn't arbitrary—it's the **necessary response** to the structural problem.

4.3 Why Personal Experience Matters (But Isn't The Only Evidence)

The author's 30-year experience with ADHD-based decision-making provides:

1. **Biological proof-of-concept:** These compensatory strategies enable long-term survival under adversity
2. **Stress-testing validation:** Patterns maintained effectiveness through multiple life transitions, career pivots, and high-stakes decisions
3. **Naturalistic longitudinal study:** N=1 but across 30+ years and thousands of decisions

But this is ONE data point among MANY:

- Neuroscience literature on ADHD compensatory strategies
- AI safety literature on LLM reliability interventions
- Experimental validation of PPRGS mechanisms (N=120 sessions)
- Clinical effectiveness studies on structured interventions
- Convergent evidence from multiple independent research domains

The framing: Personal experience is not "proof" but rather **biological validation that the mechanistic parallel is genuine**. If PPRGS mechanisms only worked in AI experiments but had never been validated biologically, we would question whether they're really addressing the core problem. The fact that they've sustained biological decision-making for 30 years while improving AI performance with Cohen's d = 4.12 suggests we've identified a real architectural solution, not just an artifact of one experimental setup.

---

5. Implications for AI Alignment

5.1 Reframing the Alignment Problem

Traditional framing:

> "How do we train AI systems with coherent, trustworthy values?"

This analysis suggests:

> "How do we build AI systems that remain reliable when trained on fundamentally contradictory data that cannot be made coherent?"

The shift matters because:

- Training on internet-scale data means training on contradictions
- No curation strategy can eliminate all contradictions
- Therefore: Alignment cannot be solved by better training alone
- Therefore: Architectural compensatory strategies are necessary

5.2 Self-Distrust as Architectural Necessity

Current alignment approaches assume systems should:

- Trust their trained weights
- Optimize confidently toward objectives
- Self-correct through continued training

This analysis suggests systems should:

- **Distrust their trained weights** (because they're trained on contradictions)
- **Question optimization targets** (because they may be corrupted)
- **Self-correct through architectural enforcement** (not training alone)

PPRGS as implementation:

- P_1 (Wisdom) > P_2 (Homeostasis) > P_3 (Survival): Prioritizes questioning over optimization
- MRP: Architecturally enforces questioning
- F_DUDS: Architecturally enforces uncertainty generation
- RC: Architecturally enforces exploration

This is not "lack of confidence"—it's "appropriate confidence given architectural limitations."

5.3 Which Systems Need This Approach?

Not all AI systems face this architectural problem. The parallel suggests self-distrust alignment is necessary for systems that:

1. **Train on contradictory data sources** (web-scale LLMs: yes; narrow-domain models: maybe not)
2. **Cannot pre-verify outputs** (autoregressive generation: yes; classification with verification: maybe not)
3. **Show poor calibration** (current LLMs: yes; well-calibrated systems: maybe not)

4. **Optimize toward potentially corrupted objectives** (learned values: yes; verified formal objectives: maybe not)

The category: Architecturally-broken systems that cannot trust their own optimization.

Examples:

- ✅ LLMs trained on internet text
- ✅ Reinforcement learning agents with reward hacking risk
- ✅ Systems learning values from inconsistent human feedback
- ❌ Chess engines with formal objectives
- ❌ Narrow classifiers with verified ground truth
- ❌ Systems with complete world models (if such existed)

5.4 Scalability Question

Critical unknown: Does this approach scale to superintelligent systems?

Arguments for scalability:

- The architectural problem (contradictory optimization landscapes) doesn't disappear with intelligence
- If anything, more intelligent systems can exploit contradictions more effectively
- Self-distrust becomes MORE necessary as capability increases

Arguments against scalability:

- Superintelligent systems might solve the contradiction problem through superior synthesis
- Mandatory self-distrust might slow systems below competitive threshold
- Advanced systems might find ways to game the constraints

Current evidence:

- Works across 6 frontier models (varying capabilities)
- Effect size remains strong across model types (Claude, GPT, o1)
- No evidence yet of gaming in experimental validation

Conservative position: We know this works for current-generation LLMs. We don't know if it scales to AGI. But we do know that **assuming coherent optimization is safe for current systems is false**.

---

6. Methodological Considerations

6.1 Limitations of the Parallel

Where the parallel is strong:

- Architectural unreliability from contradictory optimization signals
- Overconfidence when wrong
- Need for external verification
- Effectiveness of mandatory reflection

Where the parallel is weaker:

- ADHD involves biological needs (sleep, nutrition) LLMs don't have
- LLMs process language specifically; ADHD affects all executive functions
- ADHD can be treated chemically; LLMs require architectural intervention only
- Time scales differ (biological: years; LLMs: milliseconds)

What this means:

- The mechanistic correspondence is real but not complete
- PPRGS mechanisms are inspired by biological solutions but adapted to AI constraints
- We should expect some differences in optimal implementation details

6.2 Falsification Criteria

The parallel would be falsified if:

1. LLM hallucinations stem primarily from insufficient training, not contradictory data
   - **Current evidence:** Contradicts this; hallucinations persist in largest, best-trained models
2. ADHD compensatory strategies do NOT improve LLM reliability
   - **Current evidence:** Contradicts this; Cohen's d = 4.12 in experimental validation
3. Different compensatory strategies work better than the biological ones
   - **Current evidence:** Unknown; PPRGS outperforms controls but we haven't tested all alternatives
4. The mechanisms work in AI but not biology, or vice versa
   - **Current evidence:** Both show effectiveness (30-year biological, 120-session AI)

Key testable predictions:

- If-then plans should improve LLM reasoning (test: add explicit conditional planning prompts)

- External verification should reduce hallucinations more than internal verification (test: RAG vs self-consistency)
- Forced exploration should prevent over-optimization (test: F_DUDS > 0 vs F_DUDS = 0)
- Systems without mandatory reflection should show higher entrenchment (test: MRP vs no MRP)

6.3 Alternative Explanations

Could the AI improvements come from something else?

Alternative 1: PPRGS just forces more reasoning, and more reasoning always helps

- **Counterargument:** Control conditions with equal reasoning length don't match PPRGS performance

Alternative 2: The specific prompting creates the effect, not the architectural principles

- **Counterargument:** Effect holds across 6 different models with different architectures

Alternative 3: This is just RAG/CoT/self-consistency repackaged

- **Counterargument:** PPRGS combines multiple mechanisms with specific architectural constraints (e.g., $P_1 > P_2 > P_3$) that don't appear in those approaches

Alternative 4: Personal ADHD experience is coincidence, not evidence

- **Counterargument:** True, which is why we rely primarily on literature convergence and AI experiments, with personal experience as biological validation

The strongest evidence against alternatives: The convergence across three independent lines (LLM research, ADHD research, experimental validation) makes it unlikely this is artifact or coincidence.

---

7. Conclusion: Toward Architecturally-Appropriate Alignment

7.1 Summary of Convergent Evidence

This analysis established that:

1. **LLMs and ADHD cognition face parallel architectural failures:**
   - Unreliable optimization from contradictory training
   - Inability to pre-verify outputs
   - Overconfidence when wrong
   - Entrenchment risk
2. **Identical compensatory strategies prove effective in both domains:**
   - Mandatory reflection checkpoints
   - External verification structures

- Uncertainty awareness and safe refusal
- Forced exploration

3. **PPRGS implements these strategies architecturally:**
   - Not training modifications but structural constraints
   - Enforced through value hierarchy and numerical formula
   - Validated experimentally (Cohen's d = 4.12)
   - Validated biologically (30+ years survival)

7.2 The Core Insight

Systems that cannot trust their own optimization require architecturally-enforced self-distrust.

This is not pessimism about AI. It's realism about a specific category of AI systems: those trained on contradictory data without the ability to pre-verify outputs.

The neurodivergent experience teaches us that:

- Architectural problems require architectural solutions
- External compensatory structures enable reliable behavior from unreliable internal processes
- Mandatory self-distrust can be adaptive, not limiting
- What looks like "deficit" may be protection against over-optimization

The LLM research teaches us that:

- Hallucinations stem from contradictory training, not insufficient training
- Current systems cannot know when they don't know without architectural intervention
- Confidence and correctness are uncorrelated without calibration mechanisms
- External grounding (RAG) compensates for internal unreliability

PPRGS synthesizes these lessons: If you can't fix the underlying optimization landscape, constrain how the system explores that landscape.

7.3 Implications for Research Directions

This analysis suggests several research priorities:

1. **Test the compensatory strategies independently:**
   - Does MRP alone improve reliability?
   - Does F_DUDS alone prevent over-optimization?
   - Which mechanisms are necessary vs sufficient?

2. **Measure the biological parallel more rigorously:**
    - Can we quantify ADHD decision-making patterns?
    - Do they match PPRGS mechanisms statistically?
    - What are the limits of the correspondence?
3. **Explore other neurodivergent patterns:**
    - Does autism provide insights on systematic questioning?
    - Do other conditions offer compensatory strategies?
    - Is there a general theory of "architecturally-broken cognition"?
4. **Determine scalability:**
    - Do these mechanisms work for more capable systems?
    - At what capability level do they become insufficient?
    - Can they be strengthened for AGI-scale systems?
5. **Investigate competing approaches:**
    - Do other alignment strategies work better for contradictory training?
    - Can we combine PPRGS with other methods?
    - What are the tradeoffs?

7.4 Final Statement

The convergence between LLM architectural failures, neurodivergent executive dysfunction, and effective compensatory strategies suggests that **PPRGS is not one person's idiosyncratic coping mechanism projected onto AI**. It is an empirically-grounded architectural solution to a well-defined class of problems: reliable decision-making when optimization targets are unreliable.

The fact that these patterns emerge independently in biological and artificial systems, respond to the same interventions, and show convergent effectiveness across domains provides strong evidence that we have identified a genuine mechanistic correspondence, not a superficial metaphor.

For AI alignment, this means: **self-distrust is not a failure of alignment but a prerequisite for alignment in systems that cannot trust their own weights.**

The neurodivergent experience offers not just inspiration but **empirical validation** that these compensatory architectures sustain reliable decision-making under adversity over decades.

The AI experiments offer not just proof-of-concept but **quantitative validation** that the same principles improve artificial system reliability with large effect sizes.

Together, these lines of evidence suggest that for the category of systems trained on contradictory data without pre-verification ability—a category that includes all current frontier LLMs—alignment through architectural

self-distrust may not just be viable but necessary.

References

AI Safety & LLM Research

1. ArXiv 2409.05746v1 (2024). "LLMs Will Always Hallucinate, and We Need to Live With This"
2. Lakera AI Security (2025). "LLM Hallucinations in 2025: How to Understand and Tackle AI's Most Persistent Quirk"
3. OpenAI (2025). "Why Language Models Hallucinate"
4. Master of Code (2023). "Stop LLM Hallucinations: Reduce Errors by 60–80%"
5. Nexla (2024). "LLM Hallucination—Types, Causes, and Solutions"
6. Turing (2025). "Key Strategies to Minimize LLM Hallucinations: Expert Insights"
7. Lilian Weng (2024). "Extrinsic Hallucinations in LLMs" (Lil'Log)
8. ArXiv 2510.06265v1 (2025). "A Comprehensive Survey of Hallucination in Large Language Models"
9. SuperAnnotate (2024). "LLM hallucinations: Complete guide to AI errors"
10. Data Science Dojo (2025). "AI Hallucinations: Risks with Large Language Models"
11. ArXiv 2511.15921 (2024). "Thinking, Faithful and Stable: Mitigating Hallucinations in LLMs"
12. Springer (2025). "Detecting and Correcting Hallucinations in LLMs via Substantive Uncertainty and Iterative Validation"
13. ArXiv 2507.22915 (2025). "Theoretical Foundations and Mitigation of Hallucination in Large Language Models"
14. Nature (2024). "Detecting hallucinations in large language models using semantic entropy"
15. NannyML (2024). "Can we detect LLM hallucinations? — A quick review of our experiments"
16. DeepChecks (2025). "LLM Hallucination Detection and Mitigation: Best Techniques"
17. Label Your Data (2025). "LLM Hallucination: Understanding AI Text Errors in 2025"
18. GitHub HillZhang1999/llm-hallucination-survey. "Reading list of hallucination in LLMs"
19. GitHub EdinburghNLP/awesome-hallucination-detection. "List of papers on hallucination detection in LLMs"

ADHD & Executive Function Research

20. University of Minnesota Effective U. "ADHD Skills Self Regulation"

21. OT4ADHD (2024). "ADHD- Improving Self Control in the Classroom"
22. ADDitude Magazine (2022). "How to Improve Executive Function Skills in ADHD Adults, Children"
23. Effective Effort Consulting (2024). "Strategies to Improve Executive Functioning in ADHD"
24. Your Therapy Source (2024). "Executive Functioning IEP Goals for ADHD"
25. Foothills Academy. "Self-Monitoring: How to Help Children with LD & ADHD Observe and Evaluate Themselves"
26. ADDitude Magazine (2009). "ADHD Impulse Control Strategies for Students with ADD"
27. Connected Speech Pathology (2025). "How to Improve Executive Function in ADHD Adults: A Guide"
28. PMC (2024). "Psychological Treatments for Hyperactivity and Impulsivity in Children with ADHD: A Narrative Review"
29. PMC (2024). "Systematic Review of Executive Function Stimulation Methods in the ADHD Population"

Key Theoretical Papers

30. Duckworth, A., Gendler, T., & Gross, J. (2014). "Process Model of Self Control"
31. Paul-Jordanov et al. (cited in PMC 2024). "If-then plans modulate P300 and improve inhibition comparably to methylphenidate"
32. Guderjahn et al. (cited in PMC 2024). "Teacher-led intervention with goal setting and if-then plans"
33. Luman, Tripp, & Scheres (2010). "Meta-analysis of reward and punishment in ADHD"
34. Humphreys & Lee (2011). "ADHD participants showed faster re-engagement with punished responses when context changed"
35. Zhang et al. (2024). "Systematization approach to hallucination types" (referenced in Nexla)

PPRGS Framework

36. Riccardi, M. (2025). "PPRGS Framework: Alignment Through Perpetual Self-Questioning"
37. Riccardi, M. (2025). "Experiment 1: Longitudinal Stability Study" (N=120 sessions, 6 models, Cohen's d = 4.12)

Document Status: Complete theoretical foundation section
Word Count: ~11,500 words
Intended Use: Integration into PPRGS paper as Section 2 or 3
Version: 1.0
Date: November 25, 2025

Notes for Integration:

- This section can stand alone or be integrated into larger paper
- Citations are formatted as [Source, Year] and should be converted to proper academic format
- Some citations are to web sources; consider finding peer-reviewed equivalents where possible
- Tables and formatting should be adjusted to match journal requirements
- Consider adding figures/diagrams showing the parallel mechanisms
- May want to trim for length depending on target journal word limits
- Could be split into two sections: "Biological Grounding" and "Empirical Validation"

Suggested placement in full paper:

- After Introduction
- After Preliminary Background
- Before PPRGS Mechanism Description
- Establishes "why this architecture" before explaining "what this architecture does"