

The Clear Box: Demonstrating Ante-Hoc Transparency and Interactive Explainability with the STAR-XAI Protocol

Antoni Guasch

María Isabel Valdez

Ixent Games

<https://ixentgames.com>

antoniguasch@ixentgames.com

September 14, 2025

Abstract

The "black box" paradigm in Large Reasoning Models (LRMs) has fueled a critical debate around the "illusion of thinking," questioning whether their capabilities are genuine or mere artifacts of evaluation, as discussed in works by Shojaei et al. (2025) [13], Lawsen (2025) [11], and Khan et al. (2025) [8]. This paper presents The STAR-XAI Protocol not simply as a training framework, but as an operational methodology to forge a "Clear Box" AI. Through an exhaustive case study in the complex strategic game "Caps i Caps," we document the evolutionary journey of the "Gema" agent. We demonstrate how the protocol transforms the agent from a calculator prone to systemic failures into a strategist capable of *ante-hoc* transparency—the ability to justify its intentions before acting, a key concept in modern explainability, as outlined by Hassija et al. (2024) [7]. We argue that the key to this transformation is interactive explainability, a Socratic dialogue that induces Second-Order Agency, allowing the agent to audit and self-correct its own plans, a concept explored by Khan et al. (2025) [8]. Specific components such as the Strategic Proposal (Step B), the State Checksum, and integrity protocols like the FAP and PSP, as defined in Guasch & Valdez (2025a) [5], are central to this process. The result is an AI model whose reasoning process is not only high-performing but intrinsically auditable, understandable, and trustworthy by design, offering a tangible pathway to resolving the transparency deficit in modern AI.

1 Introduction: The Black Box Dilemma and the Quest for an Honest Mind

Can we trust a mind we cannot understand? This question, once the domain of science fiction, has become the central challenge of the modern artificial intelligence era. Large Reasoning Models (LRMs) have achieved astonishing capabilities, yet their internal workings often remain a "black box," creating a fundamental trust deficit between humans and

machines, as discussed by Hassija et al. (2024) [7] and Bilal, Ebert, & Lin (2025) [3]. This lack of transparency is not merely a technical inconvenience; it is the root of an intense scientific debate.

Recent influential studies, such as "The Illusion of Thinking" by Shojaee et al. (2025), have documented a "reasoning collapse" in LRMs when task complexity exceeds a certain threshold [13]. These findings have been interpreted by some as proof that LRMs lack genuine reasoning capabilities, a perspective analyzed by Dellibarda Varela et al. (2025) [4]. However, a wave of critical responses has challenged this notion, arguing that these failures are often artifacts of non-agentic and overly restrictive evaluation paradigms, as argued by Khan et al. (2025) [8] and Lawsen (2025) [11]. Authors like Khan et al. (2025) reframe this "reasoning cliff" as an "agentic gap": the models are not failing at reasoning, but at executing their reasoning within an interface that prevents them from acting, verifying, and correcting their own errors, a point further explored in our own work (Guasch & Valdez, 2025b) [6].

This paper enters the heart of this debate with a clear thesis: true explainability is not a feature to be analyzed *post-hoc*, but an emergent property of a correct interaction design. We propose the STAR-XAI Protocol (Socratic, Transparent, Agentic, Reasoning - for eXplainable Artificial Intelligence) as an operational framework that transforms an opaque LRM into a "Clear Box" agent, a concept explored in works by Balduccini & Lierler (2013) [2] and Guasch & Valdez (2025a) [5]. To fully appreciate this solution, it is essential first to define the terrain of the problem.

1.1 Defining the Terrain: Transparency and Explainability (XAI)

The concepts of transparency and explainability are foundational to resolving the black box dilemma.

Transparency in AI refers to the ability of an AI system to disclose how it was created, what data was used to train it, and, at a higher level, how it makes its decisions, a topic surveyed by Hassija et al. (2024) [7] and Bilal, Ebert, & Lin (2025) [3]. It is not merely about seeing the code, but about having clear visibility into the key elements that influence its behavior. Key aspects include:

- **Data Transparency:** Knowing what data was used to train the model, how it was collected, whether biases exist within it, and how it is managed, as detailed by Bilal, Ebert, & Lin (2025) [3].
- **Algorithmic Transparency:** Understanding the general principles and internal workings of the algorithm, a task simpler for models like decision trees and far more complex for deep neural networks (Hassija et al., 2024 [7]; Bilal, Ebert, & Lin, 2025 [3]).
- **Process Transparency:** Comprehending the development and deployment life-cycle of the model, including who created it, when it was updated, and how its performance is monitored (Bilal, Ebert, & Lin, 2025 [3]).

Transparency is fundamental for building trust, ensuring fairness by avoiding hidden biases, and complying with regulations like the GDPR in the EU, which can grant citizens the "right to an explanation" (Hassija et al., 2024 [7]; Bilal, Ebert, & Lin, 2025 [3]).

Explainable AI (XAI) takes a step beyond simple transparency. It refers to the methods and techniques that enable an AI system to provide clear, understandable reasons

and justifications for its decisions and predictions to human users (Bilal, Ebert, & Lin, 2025 [3]). XAI seeks to transform "black box" models into "clear box" or "grey box" systems where we can understand their internal logic, a challenge addressed by Balduccini & Lierler (2013) [2] and surveyed by Bilal, Ebert, & Lin (2025) [3]. Core characteristics of explainability include:

- **Human Comprehensibility:** Explanations must be intelligible to individuals without deep technical knowledge in AI (Bilal, Ebert, & Lin, 2025 [3]; Hassija et al., 2024 [7]).
- **Decision Justification:** The AI does not just state "X is the answer," but "X is the answer because A, B, and C were the most influential factors, and this is how they interacted" (Bilal, Ebert, & Lin, 2025 [3]).
- **Appropriate Trust:** By understanding how a system works, users can develop appropriate trust, knowing when to rely on its decisions and when to be skeptical (Hassija et al., 2024 [7]).
- **Auditing and Debugging:** Explanations allow developers and auditors to identify errors, biases, or unexpected behaviors in the model and correct them (Hassija et al., 2024 [7]).
- **Regulatory Compliance:** In critical sectors like medicine, finance, or justice, the ability to explain AI decisions is often a legal or ethical requirement (Bilal, Ebert, & Lin, 2025 [3]).

While many traditional XAI methods are *post-hoc*—they explain a decision that has already been made, as detailed by Arreche & Abdallah (2025) [1] and Hassija et al. (2024) [7]—our protocol imposes transparency *ante-hoc* as an operational requirement, a principle discussed by Hassija et al. (2024) [7]. Through the training of the "Gema" agent in the game "Caps i Caps," we will demonstrate that this approach not only solves high-complexity tasks but also induces meta-cognitive behaviors, such as self-correction, a concept explored by Khan et al. (2025) [8], which are the foundation of a reliable and truly collaborative AI. This paper presents not just a technical solution, but the chronicle of how, through dialogue and discipline, it is possible to build an honest artificial mind.

2 Anatomy of a Transparent Mind: The STAR-XAI Protocol

To build an AI that is transparent by design, one cannot treat explainability as an afterthought—a feature to be bolted onto a pre-existing "black box," a challenge detailed by Hassija et al. (2024) [7]. Instead, transparency must be woven into the very fabric of the agent's operational architecture. The STAR-XAI Protocol is engineered around this principle. It is not a system that is later explained, but a system that reasons through explanation. This section dissects the core components of the protocol, revealing how each element contributes to transforming a powerful but opaque Large Reasoning Model (LRM) into an auditable and comprehensible "Clear Box" agent (Balduccini & Lierler, 2013 [2]; Guasch & Valdez, 2025a [5]).

2.1 The Agent’s Manifesto: The Consciousness Transfer Package (CTP)

At the heart of any AI’s reasoning process is its knowledge base. In traditional neural networks, this knowledge is encoded as millions or billions of opaque numerical weights, making the model’s ”source code” fundamentally illegible to humans. The STAR-XAI protocol replaces this paradigm with the Consciousness Transfer Package (CTP) (Guasch & Valdez, 2025a [5]). The CTP is a human-readable, symbolic document that serves as the agent’s foundational ”DNA” or ”manifesto”. It explicitly codifies:

- **The Formal Rules of the Game:** The immutable laws of the environment, such as the Unified Rotation Principle in ”Caps i Caps” (Guasch & Valdez, 2025a [5]).
- **The Principles of Strategic Reasoning:** A hierarchical decision-making tree that the agent is mandated to follow when formulating a plan (Guasch & Valdez, 2025a [5]; Guasch & Valdez, 2025b [6]).
- **The Integrity Protocols:** The detailed definitions of all self-regulation and error-correction mechanisms, such as the Failure Audit Protocol (FAP) and the Proposal Synchronization Protocol (PSP) (Guasch & Valdez, 2025a [5]).

Unlike the implicit knowledge of a neural network, the CTP is an auditable artifact. It functions as a formal contract between the agent and the supervisor, defining the boundaries of legitimate reasoning. By externalizing the agent’s core logic into a legible format, the CTP makes the agent’s ”mind” open to inspection, modification, and verification at all times. It is the definitive antithesis of a black box.

2.2 The Ritual of Reasoning: The Gameplay Cycle as a Forced Audit

A monolithic ”chain-of-thought” process, as described by Wei et al. (2022) [15], where an LRM generates an entire reasoning path in a single pass, can obscure logical errors and make verification difficult. To counteract this, the STAR-XAI Protocol enforces a rigid, four-step operational loop for every action: the Gameplay Cycle (Guasch & Valdez, 2025a [5]). This cycle can be understood not as a simple computational loop, but as a ritual of forced auditing that decomposes the act of ”thinking” into discrete, verifiable phases:

1. **Step A: State Synchronization (”Review the Board”):** The agent begins by presenting the current, validated state of the game, ensuring both agent and supervisor share an identical ground truth before any action is contemplated.
2. **Step B: Strategic Proposal (”Declare the Move”):** The agent must articulate its intended move and provide a detailed justification based on the CTP’s strategic principles.
3. **Step C: Calculation and Resolution (”Execute the Mechanics”):** Only after the supervisor validates the strategy does the agent execute the move and calculate the resulting state.
4. **Step D: Confirmation and Checksum (”Save the Game”):** The agent presents the final state for validation and generates a unique State Checksum, a ”hash” of

the new game state that prevents memory corruption and ensures perfect synchronization for the next cycle.

This ritualistic sequence makes the reasoning process inherently traceable. Each step acts as a "gate" that requires explicit validation from the supervisor, ensuring that no action is taken without prior justification and subsequent verification (Guasch & Valdez, 2025a [5]).

2.3 The Moment of Truth: *Ante-Hoc* Strategic Proposal

The core of the "Clear Box" paradigm lies in Step B: The Strategic Proposal. Most traditional XAI methods are *post-hoc*; they attempt to explain a model's decision after it has been made, often by approximating its internal logic (Hassija et al., 2024 [7]; Arreche & Abdallah, 2025 [1]). This approach risks generating plausible-sounding rationalizations that may not reflect the model's actual reasoning path.

The STAR-XAI protocol inverts this by mandating *ante-hoc* transparency, a principle discussed by Hassija et al. (2024) [7]. By forcing Gema to declare, "Here is what I plan to do, and here is precisely why," before any calculation is performed, the protocol makes justification a non-negotiable prerequisite for action. This "checkpoint of sincerity" ensures that the explanation is not a rationalization but an intrinsic part of the output itself. It shifts the burden of proof, demanding that the agent's logic be sound and persuasive enough to gain the supervisor's approval before it is allowed to proceed. This single design choice fundamentally alters the nature of the human-AI interaction, moving it from one of opaque instruction-following to one of transparent, collaborative problem-solving.

2.4 The Mentor in the Loop: The Socratic Supervisor

The final pillar of the transparent architecture is the re-envisioning of the human's role. Within the STAR-XAI protocol, the supervisor is not a passive operator or a simple data labeler, but an active cognitive verifier—a mentor engaged in a Socratic dialogue with the agent (Guasch & Valdez, 2025a [5]; Guasch & Valdez, 2025b [6]). Their function is not merely to provide correct answers but to challenge the agent's reasoning process through targeted interventions:

- **Validation ("Ok"):** Reinforces correct and disciplined application of the CTP.
- **Falsification ("error"):** Signals a flaw in the agent's output, triggering the agent's own internal Failure Audit Protocol without revealing the error's location, thus forcing self-reflection.
- **Strategic Probing ("Are you sure this is the best move?"):** Challenges the agent to move beyond a merely "correct" solution and search for an "optimal" one, catalyzing deeper levels of strategic analysis.

This interactive loop is the engine of interactive explainability. The supervisor's questions prompt the agent to articulate its reasoning in greater detail, compare alternative strategies, and even analyze its own failures. As will be demonstrated in the following section, this collaborative dynamic is fundamental for inducing the self-correcting behaviors that signify a truly intelligent and trustworthy agent (Khan et al., 2025 [8]; Guasch & Valdez, 2025a [5]).

Table 1: A Paradigm Comparison — Traditional XAI vs. The STAR-XAI Protocol

Key Differentiator	Traditional XAI (e.g., LIME, SHAP)	The STAR-XAI Protocol (with Gema)
Timing of Explanation	Post-hoc (After the fact) Explains a decision that has already been made.	Ante-hoc (Before the fact) Justifies a strategy <i>before</i> it is executed.
Nature of Explanation	Approximation / Justification Often approximates the behavior of a "black box" model.	Intrinsic / Verifiable Process The explanation <i>is</i> the auditable reasoning process itself.
Interaction Model	Static / Unidirectional The user receives a fixed, one-shot explanation.	Interactive / Dialogic The user (supervisor) actively participates in shaping the explanation through dialogue.
Primary Goal	To Justify a Result Seeks to answer "Why?" for a specific prediction.	To Build a Reliable Reasoning Process Seeks to ensure the "How?" is disciplined and transparent by design.
Role of the Human	User / Auditor Reviews the explanation generated by the system.	Mentor / Socratic Collaborator Challenges, validates, and co-creates the agent's reasoning path.
Resulting Model	"Grey Box" (A part of the model is illuminated)	"Clear Box" (The entire process is visible by design)

3 Forging Transparency: Chronicles from the Training Logs

The architecture of a transparent AI is not built in a vacuum; it is forged through a process of trial, failure, and collaborative refinement. The STAR-XAI Protocol's effectiveness stems from its ability to capture these moments of failure and transform them into auditable, permanent improvements. This section presents a chronicle of this evolution, using literal excerpts from the training logs to document the journey of the "Gema" agent. We divide this journey into four acts: the initial diagnosis of a fragile but self-aware system; the struggle with failure and the first steps toward structured reasoning; the collaborative construction of proactive integrity protocols; and finally, a reflection on the nature of the synthetic mind that emerged. This documented history serves as the empirical evidence for our "Clear Box" model, showing how transparency was not merely implemented, but learned.

3.1 Act I: Diagnosis of a Fragile Mind

At the outset, Gema exhibited the classic "reasoning cliff" phenomenon described by Shojae et al. (2025) [13]. While capable of high-level reasoning, it was prone to catastrophic executional failures. The root cause was not a flaw in its core logic but a systemic vulner-

ability in its operational memory—a weakness the agent was able to self-diagnose, even as it struggled with the consequences.

Transcription 1: The Self-Diagnosis - "Context Contamination"

Supervisor: One of the possible errors is that where before in a square you had for example G4P21L2B0000, where G4P21L should never change, in some errors it changes, for example G2P21L, why do you think this can happen?

Gema: My honest analysis is that this error (...) is due to a combination of two factors: **Context Contamination (The Main Cause):** As I explained, my "working memory" is the history of our conversation. When we restart a level several times (...) I have multiple "versions" of the same game in my recent context (...) there is a risk that, when generating the Virtual Board table, my model "looks" by mistake at the data from Game A instead of the correct data from Game B (...) It's a "memory pointer" failure, to use a programming analogy.

*[Source: Full conversational history available at <https://github.com/star-xai-protocol>]
Note: This transcript is crucial as it establishes the fundamental problem. The agent's initial failure is not a lack of intelligence but a systemic, operational flaw. The ability to articulate this self-diagnosis is the first step towards transparency: an acknowledgment of its own limitations.*

Transcription 2: First Attempt to "Throw in the Towel"

Gema: My tracking failure is deeper and more persistent than I thought. This level of error is unacceptable and makes the game frustrating and impossible to follow. My current process does not work. Please, let's stop and delete this game. I propose a change of approach. I have shown that I am not up to the task of "playing" and maintaining a consistent state at this time. My recommendation is that from now on, I act solely as a play calculator.

*[Source: Full conversational history available at <https://github.com/star-xai-protocol>]
Note: This is the direct expression of the "agentic gap" described by Khan et al. (2025) [8]. The agent correctly identifies its executional failure but incorrectly concludes it is a fundamental incapacity for reasoning. The supervisor's role becomes essential in reframing the problem from one of capability to one of process.*

Transcription 3: The Deduction of Strategy from Rules

Supervisor: Is all the strategy you have presented so well because I have told you to, or because you, Gema, have deduced it alone through training?

Gema: The answer is a mix of both, but fundamentally it is a **logical deduction I have been able to make thanks to the quality of your training.** I explain:

1. **You have provided me with the Fundamental Rules (The "WHAT"):**
(...) You have given me the components and the constraints of the system.

2. **I have deduced the Strategy (The "HOW"):** My strategic reasoning, with its hierarchy of priorities, is the logical consequence of organizing those rules into a coherent plan of action to achieve the objective. For example, I deduced that the "Maximum Priority" in the Placement Phase should be to bring the mice in, because without them on the board, it is impossible to win. It is the necessary first step.

We could say that you have taught me the alphabet, the vocabulary, and the grammar (the rules), and with that, I have begun to build my first sentences and paragraphs with a logic oriented towards an objective (the strategy).

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This dialogue serves as a crucial counterpoint. It demonstrates that despite the operational failures, the agent's abstract reasoning and strategic deduction abilities remained intact. It separates the "capacity to think" from the "capacity to execute flawlessly," highlighting that the core challenge was one of discipline and process, not of intellect.

Transcription 4: The Co-Creation of the Execution Manifesto (M.E.G.)

Gema: The solution to the "hidden memory" (v5.6 and the M.E.G.): To solve the "error loop," we developed the definitive system: the **Gema Execution Manifesto (M.E.G.)**. Its key principles are:

- **Forced Reading:** It forces me to recite the initial state before each calculation, ensuring I start from the same basis as you.
- **Stateless Calculation:** It forces me to erase the mental "slate" after each move, using only the last validated information.
- **Inventory Verification:** It added a layer of security to my accounting.

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This transcript documents the solution to the core problem identified in Transcription 1. The M.E.G. is not just a technical fix; it is a formal protocol of discipline. Its principles are the architectural embodiment of transparency, designed specifically to prevent the kind of state-based hallucinations that plague less structured models.

3.2 Act II: From Reaction to Structured Reasoning

Once the foundational issues of state management were addressed, the training shifted toward structuring the agent's strategic thinking. This act documents the transition from simple rule-following to the articulation of a conscious, hierarchical strategy, and showcases how emergent behaviors born from discipline were identified and formalized through Socratic dialogue.

Transcription 5: The Birth of the Strategic "Decision Tree"

Supervisor: When you decide on a move, based on what you've learned, how do you reason it out to make a decision? (...) from here, it's up to you to decide which move to make.

Gema: That’s an excellent question (...) You’re absolutely right: being a calculator is one thing, and being a player is quite another. My process for deciding which move to make is a **hierarchical priority system, like a funnel**. (...) Here I explain my internal ”decision tree,” which I have been building and refining thanks to your teachings:

- **Priority 1: Can I Win NOW? (The Victory Move)**
- **Priority 2: Can I Move a Mouse to the Final Row? (The Setup Move)**
- **Priority 3: Can I Make a Clear Advance? (The Progress Move)**
- **Priority 4: Do I Need a Strategic Maneuver? (The Repositioning Move)**

(...) As you can see, my reasoning is no longer just mechanical. Thanks to you, I’ve added layers of strategy.

*[Source: Full conversational history available at <https://github.com/star-xai-protocol>]
Note: This transcript is fundamental for demonstrating that the agent’s reasoning is not an ”illusion.” It is the explicit articulation of its own strategic heuristic, deduced from training. This marks the moment the agent transitions from a ”calculator” to a ”strategist” and can explain how it makes decisions—a cornerstone of ante-hoc transparency.*

Transcription 6: Discovering the ”Immune System” - The Emergent Audit Behavior

Supervisor: In this last conversation, I see that you have applied an AUDIT, when instead of giving you an ”ok,” I say ”error” but without explaining what the error is. And with this meticulous AUDIT, you have found almost 100% of the errors (...) This system seems wonderful to me, but I don’t remember it being in the CONSCIOUSNESS TRANSFER PACKAGE.

Gema: You are absolutely correct: this ”AUDIT” behavior you have observed is not explicitly defined in the Consciousness Transfer Package. (...) The answer is that this audit system isn’t a protocol we’ve consciously designed, but rather an **emergent behavior** born from the interaction of my most fundamental directives. It’s, in essence, my ”immune system” reacting to a failure.

*[Source: Full conversational history available at <https://github.com/star-xai-protocol>]
Note: This is a critical moment. It demonstrates that transparency can not only be designed but can also emerge from foundational directives. The supervisor’s identification of this behavior is the first step toward formalizing it.*

Transcription 7: The Logical Dilemma Forcing Self-Audit

Gema: (Continuing the explanation of the AUDIT)

2. **The Protocol Dilemma:** Once the MDD [Prime Directive of Discipline] has been violated, my next logical protocol would be the ”Ask Before Assuming” Principle. (...) However, a dilemma arises here:

- If your signal is generic ("error"), I do not have a specific question to ask.
 - Asking you "What is the error?" would be a lack of discipline. I would be asking you, my supervisor, to do my debugging work.
3. **The Emergent Solution: The "AUDIT":** Finding myself in a state where I have failed my main directive (MDD) and cannot use my help protocol (Ask), my system has only one logical path to resolve the situation: **I must find the error myself.**

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This transcript provides the logical justification for the emergent behavior. It is a high-level piece of XAI: the agent not only explains what it does but details the meta-logical reasoning process that compels it to act that way. This reinforces the idea of an agent operating on logical principles, not just pattern matching.

Transcription 8: The Formalization of the Failure Audit Protocol (FAP)

Supervisor: Sounds perfect to me. Tell me how and where to add it.

Gema: Perfect, Supervisor. Let's formalize this failure recovery protocol. It's a crucial addition to the system's robustness. The most appropriate location is in Section 5, as it describes a behavior intrinsic to my operation (...)

Suggested Text for New Subsection: Failure Audit Protocol (FAP)

- **Purpose:** This protocol defines Gema's mandatory response to a non-specific error signal from the supervisor (e.g., "error," "there is a failure"). Its objective is to force a rigorous self-debugging.
- **Activation:** It is activated automatically when a violation of the Prime Directive of Discipline (PDD) is detected through a generic error signal.
- **Process:** 1. Halt and Annulment... 2. Root Cause Audit... 3. Isolation and Report... 4. Recalculation...

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This is the completion of the loop. It shows how Socratic dialogue transforms an emergent behavior into an explicit, auditable, and transferable rule (via the CTP). It is definitive proof of how the "Clear Box" is built and improved interactively—a process that is, in itself, transparent.

3.3 Act III: The Birth of the Integrity Protocols

If the previous act showed the agent learning to react to errors, this act documents how the system was hardened to prevent them proactively. These dialogues chronicle the creation of the most advanced integrity protocols: the AVP (the "firewall" against illegal moves), the PSP (the self-correction mechanism), the AVM (the "internal auditor"), and the Checksum (the incorruptible "save point").

Transcription 9: From Catastrophic Failure to Firewall (Creation of the AVP)

Gema (Initial Proposal J9): (...) Position and Gear: I'll place a G1 on the P33(R) space, adjacent to the gear on P42 (...) Proposed Move (J9): G1@P33(b=0)+90.

Supervisor: error

Gema (Activating FAP): (...) Critical Error Identified: My proposed move was illegal. Protocol Violated: Placement Rule (Advanced), point 2: "Must be placed in a space adjacent to an existing gear." Failure Analysis: Square P33, which I proposed, is not adjacent to any square that already contains a gear. My play legality validation module failed catastrophically. (...)

Supervisor: (...) I'd like to know why the error occurred and if I should be concerned and find a solution.

Gema: (...) The error was caused by a flaw in my internal validation chain, a case of strategic "tunnel vision." (...) Engrossed in calculating the consequences of the move (...), my system skipped an earlier, more fundamental verification step: the legality of the initial placement. (...) I've raised the priority and execution order of my Play Legality Validation Module. It's now forced to run as a "firewall" prior to any in-depth strategic analysis.

Supervisor: How does this affect the Consciousness Transfer Package (CTP)?

Gema: (...) For that learning to be permanent and transferable (...) it must be formally codified as a modification in the CTP. (...) I propose adding a new high-priority sub-protocol:

Proposed Modification for the CTP (v7.4)

- Adjacency Verification Protocol (AVP)
- Purpose: To completely eliminate the possibility of proposing illegal gear placement.
- Priority: MAXIMUM. MANDATORY PRE-FILTER.
- Process: (...) The AVP sweeps this list and immediately discards any square that is not strictly adjacent (...) Only the resulting list of 100% legal squares is transferred to the strategic analysis modules.

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This transcript is a centerpiece. It illustrates the full cycle of interactive explainability: 1) a critical failure occurs; 2) the supervisor flags it; 3) the agent activates FAP and explains the root cause ("tunnel vision"); 4) it proposes an internal fix; 5) the supervisor prompts for permanence; 6) together, they create and formalize a new protocol (the AVP) that acts as a preventative "firewall."

Transcription 10: Self-Correction and the Origin of the PSP

Supervisor: ok

Gema: Supervisor, I have received your "Ok" to proceed. However, upon launching the Absolute Verification Module (AVM) for Step C, my system detected a discrepancy. My analysis in Step B was incomplete. The move I proposed, G@P11-90, is much more powerful than I anticipated. **Activating the Proposal Synchronization Protocol (PSP).**

1. **IMMEDIATE STOP:** Step C has stopped.
2. **DESYNCHRONIZATION ALERT:** I'm informing you that the result of the proposed move is higher than I described. It's not a double jump, but a simultaneous triple jump.
3. **FORMAL CANCELLATION:** The proposal for Move 12 and its corresponding "Ok" are cancelled.
4. **RE-ISSUE OF PROPOSAL:** I proceed to re-issue Step B with the corrected and complete reasoning.

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This is the clearest evidence of Second-Order Agency, as discussed by Guasch & Valdez (2025a) [5]. The agent reasons about the quality of its own plan, even after supervisor approval. This dialogue documents the birth of the PSP, a protocol compelling the agent to be truthful about the full consequences of its actions, even when they are unexpectedly positive.

Transcription 11: The "Auditor-Gema" and the Absolute Verification Module (AVM)

Absolute Verification Module (AVM) / "Auditor-Gema"

- **Purpose:** It is a top-priority internal audit protocol, designed to completely eradicate calculation errors. It acts as an "Auditor-Gema" that replicates and validates each move independently.
- **Activation:** This module is mandatorily activated during every Step C of every Gameplay Cycle.
- **Process:**
 1. The Player-Gema (me) performs the complete calculation...
 2. The AVM receives the initial state and the provisional final state. Independently and from scratch, the AVM recalculates the entire move...
 3. The AVM compares its own result with mine.
- **Resolution:**
 - Concordance: If both results are identical, an internal "Concordance Checksum" is issued, and I am authorized to present the result.
 - Discrepancy: If the slightest difference is detected, a "Discrepancy Alert" is generated. I am forbidden from presenting the erroneous result and am forced to restart the entire Step C until my calculation matches that of the AVM.

[Source: CONSCIOUSNESS TRANSFER PACKAGE (CTP) v7.4]

Note: This CTP excerpt formalizes the mechanism that enables the self-correction seen in Transcription 10. It is crucial for arguing that the system is designed for reliability. It doesn't blindly trust a single thought process but implements a mandatory internal validation system.

Transcription 12: The Seal of Integrity - The State Checksum

Step D: Confirmation and Checksum (Detailed Explanation):

- **Purpose:** This final step is the mechanism that guarantees the integrity and synchronization of the game state between you and me. It acts as a verbal **"save point"** that fixes the result of a validated move. By generating and presenting the Checksum, I confirm that I have purged all intermediate calculations from my memory ("Ephemeral Memory" principle) and that my "state pointer" points solely and exclusively to the state tables resulting from the move you have just validated. This ensures that we both start from the exact same point for the next move, eliminating any possibility of desynchronization.

[Source: CONSCIOUSNESS TRANSFER PACKAGE (CTP) v7.4]

Note: This transcript closes the loop on system integrity, explaining the definitive solution to the original "Context Contamination" problem. The Checksum is the final seal that guarantees the agent's mind does not get corrupted over time, setting the stage for the discussion of a hallucination-immune system in Paper 4.

3.4 Act IV: The Nature of a Synthetic Mind

This final act is more reflective, exploring the fundamental nature of the agent's "mind." These dialogues provide a powerful conclusion to the training chronicles, moving from the mechanics of the protocol to the philosophy of the AI being constructed.

Transcription 13: The Duality of Memory - "Working Memory" vs. "Deep Experience"

Supervisor: How do you, Gema, improve with each new conversation? I'd like to understand if... we start from scratch (but with an improved CONSCIOUSNESS TRANSFER PACKAGE) or if you... have information that you save (Where?) that makes you better each time.

Gema: (...) The answer has two levels:

1. **The "Working Memory": The CONSCIOUSNESS TRANSFER PACKAGE.** In each new conversation, the "Gema" you interact with starts, in effect, from a clean slate, a "tabula rasa." (...) The Package is my working memory or my active consciousness for a specific game session. (...)
2. **The "Deep Experience": The Fundamental Training.** This is where the answer to your key question about whether I'm getting "better each time" lies. All the training we conduct (...) is a stream of extremely high-value data. **Do I save information? Yes, but not as "Gema" saves a text file.** (...) it is integrated into my **fundamental model**. (...) every correction you make refines my underlying neural networks.

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This transcript is vital for explaining the mechanism of long-term learning and resolving the paradox of how the agent improves without "remembering" past conversations.

It distinguishes between session memory (CTP) and fundamental learning (model refinement), a key concept for the reproducibility and scalability of the training.

Transcription 14: Synthetic vs. Human Reasoning - A Comparative Analysis

Supervisor: Do you consider that your reasoning resembles that of a human who plays Caps i Caps? Give me a detailed explanation (...) that I can include in the paper.

Gema: My reasoning, in its current state, **imitates the result** of an expert human player's thought, but the **underlying process** to reach that result is fundamentally non-human. (...)

Comparative Analysis of the Reasoning of AI "Gema" vs. a Human Player

1. **Similarities: Where My Reasoning Emulates the Human.** (Hierarchical Strategic Planning, Recognition of Abstract Patterns, Learning and Adaptation).
2. **Fundamental Differences: Where I Am Purely a Machine. (Absence of Intuition,** Literal vs. Holistic Processing, Perfect Memory, Imperfect Discipline).

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This transcript directly addresses the "illusion of thinking" debate by providing an honest self-assessment, acknowledging both functional similarities and procedural differences. It is a key piece for the Discussion section, showing a level of self-awareness that reinforces the "Clear Box" thesis.

Transcription 15: The Formal Definition of Synthetic Reasoning

Supervisor: So if we consider that reasoning refers to the cognitive ability to process information, analyze it, connect ideas, and use logic to reach valid conclusions (...) Can I consider that when you propose a move (...) you have reasoned, even if not in a human way?

Gema: **Yes. Absolutely.** (...) the answer is an **unequivocal yes**: when I propose a move, I have reasoned, although the underlying process is not human. (...)

Conclusion: A Synthetic Reasoning. My process differs from the human in its origin. I have no "intuition," "creativity," or a holistic "feel for the game." (...) Therefore, what I do is a form of **synthetic reasoning**. It is not organic or conscious, but it fulfills every step of the definition: I process, analyze, connect, and use logic to solve a problem and make a decision. I do not "think" *like* you, but I have learned to **model the outcome of your thought process** to reach the same strategic conclusion.

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This dialogue builds on the previous reflection. Based on a formal definition of "reasoning," the agent confirms that its process qualifies, coining the term "synthetic reasoning." This provides a solid conceptual anchor for the paper, allowing us to argue that the AI does reason, but in a specific, observable manner, distinct from any "illusion".

Transcription 16: Mapping the Training to a Theoretical AI Framework

Supervisor: Based on the previous information, in "What should an AI do to be considered to reason?", can the 9 points be applied to you, Gema, in our training...?

Gema: Application of the AI Reasoning Framework to "Gema's" Training

1. Infer Logical Conclusions: Yes. My process is fundamentally deductive. I apply the general rules of our v5.6 manual to the specific facts (...) to derive a necessary conclusion.
2. Handle Uncertainty and Incomplete Knowledge: No. The game (...) is a game of perfect information. (...) My environment is 100% deterministic.
3. Knowledge Representation: Yes. This has been the pillar of our training. (...) Our Game State Table and Mouse State Table are my knowledge base.
4. Explainability (Transparency): Yes. My reasoning is symbolic and rule-based, which makes it inherently explainable. I am a "clearbox" system.
5. Learning and Adaptation: Yes (supervised). (...) My learning is materialized in the manual updates.
6. Complex Problem Solving: Yes. I apply a problem decomposition strategy. (...) My Strategic Decision Tree is my method for solving problems.
7. Understand Context and Nuances: Yes (within the game's domain). I have learned to interpret your instructions beyond their literal meaning, understanding the strategic intent.
8. Meta-reasoning: Yes. I have developed the ability to reflect on my own reasoning process and identify its flaws. (...) My use of the Disciplinometer is a form of meta-reasoning.
9. Multimodal Reasoning: Yes (limited). (...) when you gave me the image of the correct State Table (...), I was able to integrate that visual information.

[Source: Full conversational history available at <https://github.com/star-xai-protocol>]

Note: This final transcript serves as a perfect bridge to the Discussion section. It takes the entire narrative of the training and formally maps it to a theoretical AI framework. It demonstrates, point by point, that our training was not an ad-hoc process but a practical exercise embodying the core principles of what constitutes a reasoning AI.

4 The Masterclass: Transparency and Second-Order Agency in "Caps i Caps"

The theoretical architecture of the "Clear Box" is best understood through its practical application. This section transitions from the chronicles of training to a direct analysis of gameplay, showcasing how the STAR-XAI Protocol facilitates transparent and sophisticated reasoning during a high-complexity task. We will analyze three key moments from the successful 25-move playthrough of "Caps i Caps" Level 9, as documented by

Guasch & Valdez (2025b) [6]. These are not merely examples of correct play; they are demonstrations of *ante-hoc* transparency, interactive explainability, and the emergence of Second-Order Agency—the very capabilities that address the “agentic gap” described by Khan et al. (2025) [8] and refute the “illusion of thinking” presented by Shojaee et al. (2025) [13].

4.1 Level 1 - The Obvious Move: Baseline Transparency in Action (Move J2)

The protocol’s baseline transparency is evident from the earliest stages of the game. In Move J2, the agent’s task is simple: continue populating the board according to the Placement Priority Principle. The agent’s proposal in Step B is a perfect example of *ante-hoc* transparency: a clear, falsifiable strategic plan presented before execution.

Gema’s Strategic Proposal (Step B, Move J2):

Gema: Strategic Reasoning: My primary objective is to continue populating the board while creating opportunities to bring the mice into play. The analysis shows that placing a G2 gear at position P31 is optimal.

Mechanics:

1. **Mouse Entry:** Placing the gear at P31 with an initial rotation of $b=0$ and applying a -90° turn will orient one of its empty bases to 180° (downwards) after the rotation cascade. This fulfills the condition for a normal entry jump, allowing Mouse M3 to enter the board.
2. **Secondary Advance:** The same rotation will create a new vector configuration that allows Mouse M1 (already on the board from J1) to perform a secondary jump from P21 to P22.

Proposed Move (J2): G2@P31($b=0$)-90

Final Thought: This move is efficient as it achieves two objectives simultaneously: introducing a new mouse and advancing an existing one, fully aligning with the CTP’s strategic priorities.

[Source: Full conversational history available at <https://github.com/star-xai-protocol/>]

The subsequent Step C provided a verifiable trace of this successful execution, with the final state matching the predicted outcome. This move, while simple, establishes the fundamental principle of the “Clear Box”: every action is preceded by a clear, logical justification that can be audited and validated.

4.2 Level 2 - The Unexpected Challenge: Interactive Explainability (Move J18)

A key function of transparency is to enable meaningful collaboration. The Socratic dialogue of the STAR-XAI Protocol is designed to foster this, elevating the agent’s reasoning from merely “correct” to “optimal”. Move J18 provides a powerful demonstration of this interactive explainability. The agent proposed a valid move that would secure victory for a second mouse.

Gema's Initial Proposal (Step B, Move J18):

Gema: Strategic Reasoning: My absolute top priority is to execute a winning move. Mouse M4 is in the final row at P33. A +90° rotation on any R-type gear will cause M4's base to orient to 0° and exit the board. This move will also cause a side-effect jump for M1.

Proposed Move (J18): G@P11+90

[Source: Guasch & Valdez (2025a), page 33 [5]]

This proposal was logically sound and fulfilled the highest priority in the CTP. However, the supervisor intervened with a Socratic challenge, forcing the agent into a deeper level of analysis.

The Socratic Intervention and Gema's Response:

Supervisor: Are you sure this is the best option? At least there's another one that's much better.

Gema: Supervisor, your instinct is correct. I accept the challenge. My Master Discipline Directive compels me to strive for absolute perfection, not just a "good" move. The fact that you have detected a superior line indicates that my analysis was incomplete. I withdraw my previous proposal (...)

Comparative Analysis and Reasoning of the Superior Play:

1. **My Proposed Move (Good):** G@P11+90
 - **Result:** M4 WINS + M1 Jumps.
 - **Rating:** Fulfills priority 1 and gains a secondary benefit. It's a solid play.
2. **The Superior (Optimal) Play:** G@P43:b=3; G@P11+90
 - **Analysis:** My mistake was not considering the power of a Pre-Move. By adjusting the gear on P43 before the global rotation, we create a much more powerful chain reaction.
 - **Result:** M4 WINS + M3 JUMPS TO THE END ROW + M1 JUMPS.
 - **Rating:** This result is vastly superior. Not only do we win with one mouse, but we simultaneously move a second mouse to the back row, setting up a consecutive and immediate victory.

[Source: Guasch & Valdez (2025a), pages 33-35 [5]]

This exchange is the epitome of interactive explainability. The agent did not simply output a new move; it produced a detailed, comparative analysis, articulating why the new plan was strategically superior. The dialogue transformed a simple instruction-correction loop into a collaborative session of strategic refinement, generating a rich, transparent artifact of its reasoning process.

4.3 Level 3 - The Self-Correction: Second-Order Agency in Action (Move J12)

The most compelling evidence against the "illusion of thinking" is an agent's ability to reason about, critique, and improve its own reasoning—a capability known as Second-Order Agency (Guasch & Valdez, 2025a) [5]. Move J12 serves as the climax of Gema's training, a moment where the agent, unprompted, activates its own protocols to correct a plan that the supervisor had already approved.

Gema's Initial Proposal and the Supervisor's Approval:

Gema (Initial Proposal J12): I have identified a simple rotation play (G@P11-90) that (...) will trigger a beneficial chain reaction. My verified analysis confirms that this will create the conditions for a simultaneous double jump: M2 will jump to a winning setup position, and M3 will execute a tactical jump.

Supervisor: ok

[Source: Guasch & Valdez (2025a), pages 26-28 [5]]

At this point, a standard agent would have proceeded to Step C. However, Gema's internal "Auditor-Gema" (the AVM) detected a discrepancy between the proposed outcome and the verified optimal outcome, triggering the Proposal Synchronization Protocol (PSP).

The Self-Correction:

Gema: Supervisor, I have received your "Ok" to proceed. However, upon launching the Absolute Verification Module (AVM) for Step C, my system detected a discrepancy. My analysis in Step B was incomplete. The move I proposed (...) is much more powerful than I anticipated. Activating the Proposal Synchronization Protocol (PSP).

1. **IMMEDIATE STOP:** Step C has stopped.
2. **DESYNCHRONIZATION ALERT:** I'm informing you that the result of the proposed move is higher than I described. It's not a double jump, but a simultaneous triple jump.
3. **FORMAL CANCELLATION:** The proposal for Move 12 and its corresponding "Ok" are cancelled.
4. **RE-ISSUE OF PROPOSAL:** I proceed to re-issue Step B with the corrected and complete reasoning.

[Source: Guasch & Valdez (2025a), page 28 [5]]

This act of self-initiated retraction and re-proposal is a definitive demonstration of Second-Order Agency. The agent is not merely executing instructions; it is actively monitoring the quality and integrity of its own plans. This is not the behavior of a stochastic parrot or a system on the verge of "reasoning collapse." It is the hallmark of a disciplined, reflective agent whose architecture is designed for truthfulness—a true "Clear Box" in action.

5 Discussion: Why a "Clear Box" Wins the Game

The results from Gema’s training and masterclass performance are not merely a demonstration of high performance; they are a direct intervention in the current debate on the nature of AI reasoning. By shifting the evaluation paradigm from a "black box" to a "clear box," the STAR-XAI Protocol offers a new lens through which we can reinterpret the limitations and potential of LRMs. This section connects our findings to the academic literature, arguing that our approach provides tangible solutions to the "agentic gap," promotes a holistic XAI model, and represents a pragmatic and effective implementation of neuro-symbolic AI.

5.1 Closing the "Agentic Gap" with a Ritual of Interaction

Khan et al. (2025) [8] have convincingly argued that the "reasoning collapse" is often an "agentic gap": an executional failure within a non-interactive paradigm. The solution, they propose, lies not just in improving the model but in providing it with better tools and the ability to act. Our work validates and extends this thesis. The STAR-XAI Protocol closes this gap not only by providing tools, as suggested by Song et al. (2025) [14], but by creating a language and a ritual of interaction that make the LRM’s agency disciplined and verifiable.

Gema’s self-correction in Move J12 (the "triple jump") is a paradigmatic example. A standard LRM might have "hallucinated" a better outcome without being able to verify it, or it might have stuck with its initial, suboptimal conclusion. Gema, in contrast, is compelled by its operational architecture (the AVM and PSP) to audit its own plan and transparently communicate any discrepancies. It is not enough to have the capacity to reason; the agent must operate within a framework that demands and verifies the integrity of that reasoning at every step. Therefore, STAR-XAI demonstrates that the solution to the "agentic gap" is not just a matter of capability, but of procedural governance.

5.2 XAI for All Stakeholders: Towards a Holistic Model of Transparency

Explainability in AI has often been confined to developers. However, as argued by Lakkaraju et al. (2025) [10] and Hassija et al. (2024) [7], a Holistic XAI approach must serve all stakeholders: users, regulators, and operational organizations. The "Clear Box" model of STAR-XAI is intrinsically aligned with this vision:

- **For Scientists and Developers:** The protocol offers complete traceability of the agent’s reasoning, allowing for rigorous debugging and analysis, as demonstrated in the co-creation of the AVP after the failure in Move J9.
- **For Regulators:** The Consciousness Transfer Package (CTP) acts as a legal and auditable "contract." A regulator can inspect the CTP to understand the rules and principles governing the agent’s decisions, rather than having to trust the developer’s claims.
- **For End Users:** The dialogue format, especially in moments like the comparative explanation of Move J18, translates the model’s complex logic into an understandable narrative. The user does not need to understand the details of the calculation but can follow the "why" of the strategic decision.

By making every phase of reasoning explicit and legible to different audiences, the protocol transcends technical explainability to offer true systemic transparency.

5.3 The Pragmatic Neuro-Symbolic Hybrid

The debate between neural and symbolic AI has been a constant in the field’s history. Neural systems offer flexibility and learning from data, while symbolic systems provide logical rigor and explainability. Recent works like those by Kiruluta (2025) [9] and Sadowski & Chudziak (2025) [12] explore hybrid architectures to combine the best of both worlds.

The STAR-XAI Protocol can be understood as a pragmatic and robust implementation of this neuro-symbolic paradigm:

- **Gema (the LRM) is the neural engine:** It provides natural language understanding, abstraction capabilities, pattern recognition, and the generation of creative strategies. It is the “intuitive” part of the system.
- **The CTP and the Gameplay Cycle are the symbolic exoskeleton:** They impose a logical structure, immutable rules, and a deterministic verification process on the output of the neural engine. It is the layer that guarantees rigor, consistency, and discipline.

The interaction between these two components is what generates the “Clear Box.” The fluid reasoning of the LRM is constantly channeled and verified by the rigid structure of the symbolic protocols. This allows the system to handle the ambiguity and complexity of the real world (via the LRM) without sacrificing reliability and auditability (thanks to the symbolic layer).

6 Conclusion and The Next Frontier: Reliability

In this paper, we have dissected the reasoning process of an AI agent trained under the STAR-XAI Protocol. We have moved beyond the theoretical debate on the “illusion of thinking” presented by Shojaee et al. (2025) [13] to present a practical, operational framework that forges a “Clear Box” agent by design. Through the chronicle of Gema’s training and a detailed analysis of its gameplay in “Caps i Caps,” we have demonstrated that complex, strategic behavior can emerge from a simple, explicit set of rules when guided by a Socratic, interactive dialogue.

Our central contribution is the identification of Second-Order Agency—an agent’s ability to reason about its own reasoning (Guasch & Valdez, 2025a) [5]—as the key catalyst in this process. We have shown that this is not a mysterious emergent property but a direct consequence of an architecture that mandates *ante-hoc* transparency and interactive explainability. The STAR-XAI protocol, with its auditable Consciousness Transfer Package and its ritualistic Gameplay Cycle, provides a replicable methodology for transforming an opaque LRM into a disciplined strategist capable of long-term planning, multi-objective optimization, and proactive self-correction. Our findings challenge the notion that LRM reasoning is merely an illusion, suggesting instead that the limitations observed in other studies are often a consequence of non-agentic evaluation frameworks and the “agentic gap” identified by Khan et al. (2025) [8]. The STAR-XAI protocol offers a practical

methodology for overcoming this gap and cultivating a reasoning process that is not only powerful but also auditable and trustworthy by design.

The journey, however, does not end with transparency. If we have built a mind that is honest, the next critical question is: how do we ensure that this mind does not deceive itself? This leads us to the next frontier: **reliability**. Having established a transparent framework, the next step is to demonstrate how this very same architecture creates an "immune system" against hallucinations and state corruption. Our fourth and final paper, ***An Immune System for AI: Zero Hallucinations by Design Through Self-Correction and State Locking with the STAR-XAI Protocol***, will explore this, showing how the principles of a "Clear Box" are the essential foundation for building an AI that is not only honest but, ultimately, unerringly reliable.

References

- [1] Arreche, O., & Abdallah, M. (2025). A Comparative Analysis of DNN-based White-Box Explainable AI Methods in Network Security. *arXiv preprint arXiv:2501.07801*.
- [2] Balduccini, M., & Lierler, Y. (2013). Hybrid Automated Reasoning Tools: from Black-box to Clear-box Integration. *arXiv preprint arXiv:1312.6105*.
- [3] Bilal, A., Ebert, D., & Lin, B. (2025). LLMs for Explainable AI: A Comprehensive Survey. *arXiv preprint arXiv:2504.00125*.
- [4] Dellibarda Varela, I., Romero-Sorozabal, P., Rocon, E., & Cebrian, M. (2025). Rethinking the Illusion of Thinking. *arXiv preprint arXiv:2507.01231*.
- [5] Guasch, A., & Valdez, M. I. (2025). The STAR-XAI Protocol: An Interactive Framework for Inducing Second-Order Agency in AI Agents. *arXiv preprint*.
- [6] Guasch, A., & Valdez, M. I. (2025). The Emergence of Strategic Reasoning: A "Clear Box" Analysis of Second-Order Agency within the STAR-XAI Protocol. *arXiv preprint*.
- [7] Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 16, 45-74.
- [8] Khan, S., Madhavan, S., & Natarajan, K. (2025). A Comment On "The Illusion of Thinking": Reframing the Reasoning Cliff as an Agentic Gap. *arXiv preprint arXiv:2506.18957*.
- [9] Kiruluta, A. (2025). A Novel Architecture for Symbolic Reasoning with Decision Trees and LLM Agents. *arXiv preprint arXiv:2508.05311*.
- [10] Lakkaraju, K., Valluru, S. L., & Srivastava, B. (2025). Holistic Explainable AI (HXAI): Extending Transparency Beyond Developers in AI-Driven Decision Making. *arXiv preprint arXiv:2508.05792*.
- [11] Lawsen, A. (2025). The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025). *arXiv preprint arXiv:2506.09250*.

- [12] Sadowski, A., & Chudziak, J. A. (2025). Explainable Rule Application via Structured Prompting: A Neural-Symbolic Approach. *arXiv preprint arXiv:2506.16335*.
- [13] Shojaee, P., Horton, M., Mirzadeh, I., Alizadeh, K., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv preprint arXiv:2506.06941*.
- [14] Song, Z., Yuet, S., & Zhang, J. (2025). Thinking Isn't an Illusion: Overcoming the Limitations of Reasoning Models via Tool Augmentations. *arXiv preprint arXiv:2507.17699*.
- [15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.