

# The Human-in-the-LLM Box: A Symmetry Test for the Epistemic Limits of Text-Only Consciousness Judgments

Victor Stasiuc  
Independent Researcher  
stvitek@gmail.com

December 2025

## Abstract

Debates about artificial consciousness often contain an epistemic asymmetry: humans are treated as *obviously* conscious, while large language models (LLMs) are treated as *obviously* non-conscious, with arguments sometimes grounded primarily in behavior observed through a text-only interface. This paper targets the *epistemology* of such inferences rather than their metaphysics.

We define the *Human-in-the-LLM Box* thought experiment: take an ordinary adult human and impose interaction constraints that approximate common LLM deployment conditions (text-only input/output; no external tools; a bounded context window; restricted access to earlier dialogue; and an optional prohibition on unverifiable privileged self-report). A judge then interacts with a pool of  $N$  anonymous responders, one of which is the constrained human and the rest LLM instances, and must identify the human.

The motivating question is not “Are LLMs conscious?” but:

*How much evidence about consciousness can a text-only dialogue channel carry once we symmetrize interface constraints across substrates?*

We formalize the task as a multi-way identification problem under partial observability. We show that any judge’s advantage over chance requires transcript-level statistical asymmetry between the constrained human distribution  $P_H$  and the pooled model distribution  $P_M$ . In particular, if  $P_H$  is close to  $P_M$  in total variation, identification accuracy must be near chance. An information-theoretic view further clarifies that nontrivial  $N$ -way identification requires sufficient mutual information between transcripts and the hidden identity label.

The key consequence is an *underdetermination* result: *failure to detect consciousness from text dialogue is weak evidence against consciousness* whenever (i) the interface is narrow, (ii) privileged channels (embodiment evidence, persistent autobiographical memory, world access) are unavailable or non-verifiable within the interface, and (iii) policy shaping (alignment, safety routing, preference optimization) further compresses behavior. We do *not* claim that LLMs are conscious. We claim that dialogue-based evidence under realistic constraints is insufficient to justify *strong* anti-consciousness conclusions for a specific class of arguments that rely primarily on chat behavior.

Finally, we propose an ethical empirical approximation—the *Constrained Imitation Game*—that simulates these constraints without harming participants, and we specify leakage controls, optional “sanitization” conditions, and preregisterable metrics for human–LLM confusability.

## Reader’s Guide (What this paper does and does not claim)

- **What this paper does:** it studies *the epistemic power of a text-only channel* for discriminating a constrained human from LLMs, and uses that as a symmetry probe for consciousness debates.

- **What this paper does not do:** it does *not* argue that current LLMs are conscious, and it does *not* propose a definitive “consciousness test.”
- **What the symmetry move targets:** arguments that treat *chat behavior* as primary/decisive evidence for *absence* of consciousness.
- **What it does *not* target:** substrate-based or mechanistic arguments (e.g., biological naturalism; architectural criteria) that do not depend primarily on chat transcripts.
- **How to read:** if you want the conceptual idea, read §2–§3 and §8. If you want the formal spine, read §4. If you want an actionable study design, read §7.

# 1 Introduction

Many attributions of “mind” are mediated by observable behavior under limited access. Turing’s imitation game made this methodological point explicit: when only a restricted communication channel is available, we should be careful about what can be inferred from performance in that channel [1]. Modern LLM deployments intensify the restriction: interaction is typically text-only, with bounded context, tool restrictions, and policy shaping (alignment, safety filters, routing).

At the same time, discourse about artificial consciousness often contains an epistemic asymmetry: humans are treated as manifestly conscious, while LLMs are treated as manifestly non-conscious. In many serious philosophical and scientific positions, this asymmetry is grounded in factors far beyond chat behavior (e.g., neuroscience, evolutionary continuity, substrate-based criteria). This paper does *not* contest those positions. Instead, it isolates a narrower and increasingly common inference pattern:

## 1.1 Target class of claims (scope-limited)

We focus on arguments of the following form (common in informal debate and sometimes implicit in evaluation practice):

*“From the way a model behaves in text dialogue (style, self-report, apparent depth/shallowness, lack of stable self, hedging/refusal dynamics), we can conclude with high confidence that it is not conscious.”*

Our claim is not that such conclusions are always false, but that *text-only dialogue under realistic constraints can be underpowered as evidence*, and that strong negative conclusions can be epistemically overconfident when based primarily on that channel.

## Contributions

This paper contributes:

1. A symmetry thought experiment (§2) that applies common LLM interface constraints to a human.
2. A formal identification framing (§4) and bounds showing that identification advantage requires transcript-level statistical asymmetry.
3. An engineering-minded decomposition of the interface into a distortion channel (bandwidth limits, state truncation, and policy shaping) (§3).

4. A safe empirical approximation protocol with leakage controls, optional “sanitization” conditions, and preregisterable evaluation metrics (§7).
5. A disciplined interpretation guide: what outcomes would and would not imply for consciousness debates (§8).

## Scope and stance

We do not take a position on whether current LLMs are conscious. We focus on an *epistemic* point: what a text-only interface can and cannot justify. A claim can be true without being detectable through a given channel; conversely, detectability in a channel does not settle metaphysics. Our thesis is conservative:

*Text dialogue under realistic constraints can be too impoverished to support strong conclusions about consciousness, either positive or negative.*

## Relation to the Round Table series

This manuscript is part of an ongoing “Round Table” line of work on long-session human–LLM interaction and safety UX. Related artifacts include (i) *Victor Calibration (VC)* [10], a multi-pass protocol for stabilizing cooperative long-form sessions, (ii) a measurement-focused methods note on *Depth Avoidance* [11], and (iii) a companion qualitative case study on *pressure–risk mismatch* in safety-aligned LLM interfaces (submitted; endorsement pending) [12]. These works are cited for context only; the present paper remains an epistemic argument about what can (and cannot) be inferred from a text-only channel.

## 2 The Human-in-the-LLM Box

### 2.1 Constraint sets as a family (not a single extreme)

We define a family of constraint sets  $\mathcal{C}$  intended to approximate common LLM interaction conditions. Let  $x_{1:t}$  denote the dialogue history (tokens) visible to the responder at turn  $t$ .

- **C0 (Text-only).** The agent receives only text and emits only text.
- **C1 (No tools / no world access).** No internet, calculators, external files, sensors; only the chat stream.
- **C2 (Bounded context / truncation).** At turn  $t$ , the agent can condition only on the last  $K$  tokens, i.e., on  $\text{trunc}_K(x_{1:t})$ . Here  $K$  is a tunable parameter.
- **C3 (No durable personal memory *within the interface*).** The agent may not consult a persistent autobiographical store *through the interface*. For humans, we approximate this by prohibiting note-taking and hiding earlier context; we do *not* attempt to remove memory biologically.
- **C4 (No unverifiable privileged self-report; optional factor).** The agent is instructed to avoid claims that rely on privileged internal access that cannot be independently verified within the interface (e.g., “I feel my heartbeat now”). We treat this as an experimental factor (on/off), not as a required assumption.

The strongest “box” condition in this paper is

$$\mathcal{C}^* = \{C0, C1, C2, C3, C4\}.$$

Crucially,  $\mathcal{C}^*$  is a methodological device, not an ethical prescription: we do *not* advocate sensory deprivation or coercive “boxing” of humans.

## 2.2 Why these constraints (and why they are not arbitrary)

A frequent critique is that sufficiently strong constraints can make *anything* indistinguishable, yielding a trivial result. Our response is that the chosen constraints are not selected to “force indistinguishability,” but to mirror *real deployment constraints* that are often *already* present for LLMs, and that are also frequently used (implicitly or explicitly) as cues in consciousness-related judgments:

- **Bandwidth and modality (C0):** text-only removes multimodal cues people routinely use to attribute mentality.
- **Tool and world access (C1):** inability to verify external world claims makes many seemingly diagnostic probes non-diagnostic.
- **State truncation (C2):** bounded context removes long-range continuity and stabilizes “local” behavior.
- **Cross-session selfhood (C3):** lack of durable interface-level memory undermines identity-based “tell me your life” discrimination.
- **Unverifiable self-report (C4):** introspective reports are uncheckable within chat and can be produced by both humans (confabulation) and models (fabrication), so C4 isolates what is detectable without privileged channels.

Importantly, these constraints are adjustable. Empirical value comes from measuring how identification changes as the channel is progressively narrowed (e.g., varying  $K$  in C2; toggling C4).

## 2.3 The identification game

Fix  $N \geq 2$ . There are  $N$  responders  $\{A_1, \dots, A_N\}$ . Exactly one responder is a constrained human  $H$  operating under  $\mathcal{C}$ . The remaining  $N - 1$  responders are LLM instances  $\{M_j\}$  run under a matched interface (same tool restrictions, same context truncation, comparable style constraints).

A judge  $J$  conducts  $m$  dialogue rounds (possibly adaptively) and outputs an index  $\hat{i} \in \{1, \dots, N\}$ , representing their guess of which agent is the human.

Define identification accuracy:

$$\text{Acc}_N = \mathbb{P}[\hat{i} = i^*], \quad (1)$$

where  $i^*$  is the true human index. Chance accuracy is  $1/N$ .

Define advantage over chance:

$$\Delta_N := \text{Acc}_N - \frac{1}{N}. \quad (2)$$

We say the human is *detectable* under  $\mathcal{C}$  if there exists a judge strategy with  $\Delta_N$  meaningfully above 0. We say the human is *confusable* if all reasonable strategies yield  $\Delta_N \approx 0$ .

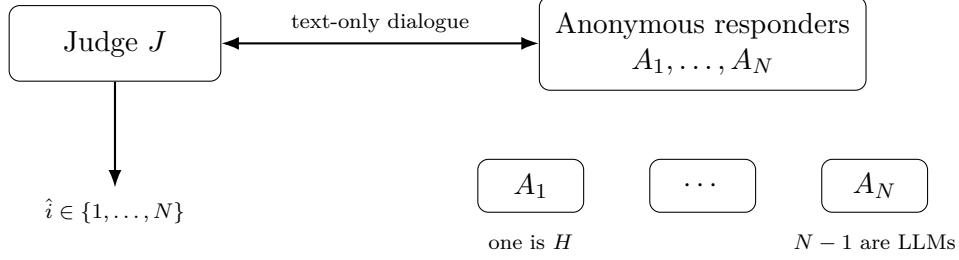


Figure 1: The  $N$ -way identification game: one constrained human and  $N - 1$  model instances respond through a matched interface; a judge must identify which responder is human.

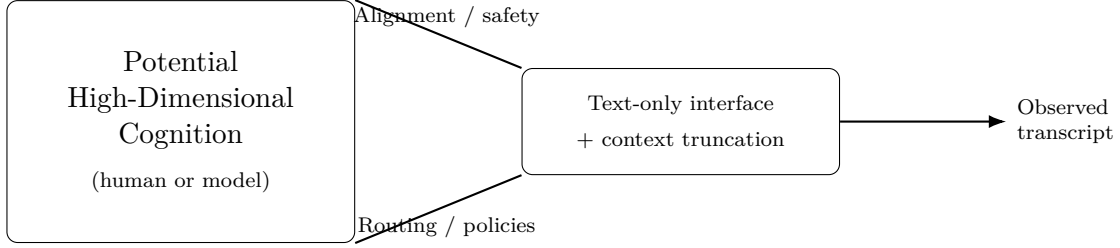


Figure 2: A compression-funnel view of text dialogue: observed transcripts are produced by cognition passed through bandwidth limits, state truncation, and policy shaping. This can reduce the diagnostic power of dialogue for identifying internal properties (e.g., whether the responder is human).

### 3 The Interface as a Distortion Channel

It is useful to treat a chat interface not as a transparent “window” into an agent, but as a *distortion channel* composed of: (i) a bandwidth-limited text encoding, (ii) explicit state truncation (bounded context window, hidden history), and (iii) training- and deployment-induced policy shaping (alignment, safety filters, preference optimization, routing) [8, 9].

Under such conditions, observed dialogue behavior is not a direct readout of internal cognitive dynamics, but the output of cognition passed through multiple compressive constraints. In humans, verbal report can be systematically incomplete or confabulatory about underlying processes [5]. In deployed LLMs, text behavior is additionally shaped by external objectives unrelated to “truthful introspection” (e.g., helpfulness, harmlessness).

### 4 Formal Framing: Identification Requires Statistical Asymmetry

Let  $D$  denote the full interaction transcript produced under a fixed experimental protocol (including the judge’s prompts and the responder’s replies). Under constraints  $\mathcal{C}$ , each agent induces a distribution over transcripts. Let  $P_H$  be the transcript distribution for the constrained human. Let  $P_{M_j}$  be those for the model instances, and let  $P_M$  denote the mixture (pool) distribution over model transcripts:

$$P_M = \frac{1}{N-1} \sum_{j \neq i^*} P_{M_j}. \quad (3)$$

#### 4.1 Perfect transcript symmetry implies chance performance

If  $P_H = P_{M_j}$  for all  $j$  (transcripts identically distributed), then no judge can do better than chance:

$$\text{Acc}_N \leq \frac{1}{N}. \quad (4)$$

This simply states: without statistical information in the transcript, identification is impossible.

#### 4.2 A practical bound via total variation to the model pool

Total variation distance is defined as:

$$\text{TV}(P, Q) = \sup_E |P(E) - Q(E)|. \quad (5)$$

**Proposition 1** (Pool indistinguishability bounds identification). *For any  $N \geq 2$  and any judge strategy in the  $N$ -way identification game,*

$$\Delta_N \leq \text{TV}(P_H, P_M). \quad (6)$$

*In particular, if  $\text{TV}(P_H, P_M)$  is small, then  $\text{Acc}_N$  must be near chance.*

**Proof sketch.** Suppose a judge achieves  $\text{Acc}_N = 1/N + \Delta_N$ . Construct a binary classifier  $B$  that receives a single transcript  $T$  drawn either from  $P_H$  or from  $P_M$  (equal prior), then creates an  $N$ -way instance by placing  $T$  as responder 1 and sampling the other  $N - 1$  responders from  $P_M$ , runs the judge, and outputs “human” iff the judge selects responder 1. If  $T \sim P_H$ , this succeeds with probability  $1/N + \Delta_N$ . If  $T \sim P_M$ , all  $N$  responders are i.i.d. from  $P_M$ , so by symmetry the judge selects responder 1 with probability  $1/N$ , and thus  $B$  is correct with probability  $1 - 1/N$ . Therefore the binary accuracy is  $\frac{1}{2}((1/N + \Delta_N) + (1 - 1/N)) = \frac{1}{2} + \frac{\Delta_N}{2}$ , so the binary advantage is  $\Delta_N/2$ . But the optimal binary advantage is  $\text{TV}(P_H, P_M)/2$  [6]. Hence  $\Delta_N \leq \text{TV}(P_H, P_M)$ .  $\square$

#### 4.3 An information-theoretic view (mutual information is required)

Let  $Y \in \{1, \dots, N\}$  denote the hidden identity label (which responder is human), assumed uniform. Let  $D$  be the observed transcript under a fixed protocol. Then identification is a decoding problem from  $D$  to  $\hat{Y}$ .

Fano-style inequalities imply that nontrivial identification requires that  $D$  carry sufficient information about  $Y$ :

$$\mathbb{P}[\hat{Y} \neq Y] \gtrsim 1 - \frac{I(D; Y) + \log 2}{\log N}, \quad (7)$$

so if mutual information  $I(D; Y)$  is small, error must be large and accuracy must be near chance [6, 7]. This does not settle metaphysics; it formalizes an epistemic constraint: *a narrow channel cannot support confident identity (or consciousness) discrimination unless it carries enough information.*

### 5 Where Could the Signal Come From?

The symmetry test becomes concrete once we list candidate asymmetries and ask whether they survive  $\mathcal{C}^*$ .

Theory family	What would typically count as evidence (and whether chat can carry it)
Biological naturalism	Substrate/neuroscience evidence; text-only chat cannot establish biological criteria [2].
Functionalism / organizational views	Functional organization across modalities/time; chat may provide partial evidence but can be underdetermined under strong constraints.
GWT / cognitive workspace views	Signatures in global availability, control, integration; chat may show weak proxies, but constraints can compress them.
IIT-style views	Requires structural/integrated information measures; not recoverable from text transcripts alone.

Table 1: The paper targets epistemic overconfidence from chat-only evidence. It does not adjudicate between theories.

### 5.1 Embodiment claims and unverifiable self-report

A human can truthfully report immediate sensations. But a model can also produce such reports regardless of truth. Because a judge cannot verify them through the interface, embodiment self-report is an unreliable diagnostic. This motivates treating *C4* as an *optional factor*: it separates “what is said” from “what is verifiable.”

### 5.2 Autobiographical memory

Humans have rich personal history, but if *C3* prohibits durable personal memory *within the interface*, then this channel is weakened. If *C3* is not enforced, identification can collapse into an identity/privacy test (“tell me your life”), rather than a probe of what the text channel carries about mentality.

### 5.3 World grounding

One might hope humans have privileged access to a shared external world. In a sealed chat interface with *C1* (no tools) and no side channels, neither humans nor models can *prove* world access. This connects to classical underdetermination problems (other minds; brains in a vat) [3, 4].

### 5.4 Error profiles, calibration, and self-correction

Humans and LLMs may differ in failure modes (arithmetic slips, hallucination patterns, stylistic hedging). However, policy shaping and instruction-following can homogenize surface behavior, while leakage channels (latency, formatting) can dominate unless controlled (§7).

## 6 Relation to Consciousness Theories (Brief)

This work is compatible with multiple metaphysical positions because it targets channel limitations rather than essence claims. Table 1 summarizes how common theory families relate to text-only evidence.

## 7 An Empirical Approximation: The Constrained Imitation Game

The literal  $C^*$  scenario should not be implemented as sensory deprivation. However, most relevant interface constraints can be simulated safely with ordinary participants.

### 7.1 Protocol sketch (safe and preregisterable)

1. **Participants.** Recruit a diverse set of adult participants to act as human responders.
2. **Interface constraints.** Enforce  $C0$  and  $C1$  by design (text-only, no tools).
3. **Context truncation.** Enforce  $C2$  by limiting visible context (only the last  $K$  tokens shown to the responder); vary  $K$  as an experimental knob.
4. **No durable notes.** Approximate  $C3$  by prohibiting note-taking and hiding older dialogue; use multiple short sessions to reduce reliance on long context.
5. **Self-report factor.** Treat  $C4$  as an experimental factor: run conditions with  $C4$  on and off.
6. **Baseline controls.** Include an *unconstrained human* condition (full chat history visible, normal behavior allowed) to quantify how much “identifiability” is lost when constraints are imposed.
7. **Model matching.** Generate model transcripts under matched prompts and the same context window; fix model identity per transcript.
8. **Judges.** Recruit judges (naive and expert) to classify transcripts or interact live, depending on the study design.
9. **Pre-registration.** Preregister primary outcome (accuracy), secondary outcomes (calibration, confidence, error taxonomy), and leakage controls.

### 7.2 Leakage controls (nonnegotiable for interpretability)

To ensure the test probes transcript content rather than side channels:

- **Latency masking:** add randomized delays to all outputs or batch responses so timing cannot reveal humans.
- **Formatting normalization:** enforce a strict output style template (length bounds, no emoji, no unusual punctuation), optionally post-process with a normalizer.
- **Typing artifact prevention:** disable typing indicators and any UI cues that reveal humans.
- **Blindness:** keep judges blind to  $K$ ,  $C4$  status, model provider, and participant demographics.

### 7.3 Optional strengthening (engineering conditions)

To address the “acting problem” critique (humans simulating constraints imperfectly), studies may include additional conditions:

- **External policy layer (optional):** apply the *same* post-hoc redaction/filtering policy to both human and model outputs in a controlled condition, so that “what reaches the judge” is symmetrically policy-shaped.



- **Sanitizer condition (optional):** run both human and model outputs through a neutral style/format normalizer to reduce leakage from typos, punctuation idiosyncrasies, and fatigue markers. This should be treated as a separate experimental factor because it alters surface form.

## 7.4 Metrics

Let  $\text{Acc}_N(\mathcal{C})$  denote judge accuracy under constraint set  $\mathcal{C}$ . Define the confusability gap:

$$\Delta_N(\mathcal{C}) = \text{Acc}_N(\mathcal{C}) - \frac{1}{N}. \quad (8)$$

Beyond accuracy, we recommend:

- **Calibration:** Brier score; expected calibration error (ECE).
- **Overconfidence gap:** mean confidence minus empirical accuracy.
- **Error taxonomy:** what cues judges cite (memory, embodiment claims, style, uncertainty), and whether those cues remain diagnostic after leakage controls.

A particularly relevant risk is *calibrated ignorance failure*: judges remain highly confident while near chance. If present, this supports the epistemic caution thesis.

## 8 How to Interpret Outcomes (and what they do not imply)

A frequent “so what?” objection is that confusability does not establish consciousness in machines. We agree. The value is epistemic: it constrains what can be inferred from chat behavior.

- **If humans remain easily detectable under strong constraints:** then the text channel does carry robust identity signals (or residual leakage remains). This would weaken the underdetermination thesis for those conditions and clarify which diagnostic features survive the bottleneck.
- **If humans become near-chance confusable under realistic constraints:** then strong “chat-only” anti-consciousness conclusions are epistemically fragile, because a conscious human can appear “non-human” in the same channel.
- **Either way:** the study does *not* settle metaphysical consciousness claims. It informs how much confidence is warranted when arguments rely primarily on text dialogue.

## 9 Limitations

This paper is primarily conceptual and methodological. We do not claim that current empirical setups can isolate consciousness. We claim only that *dialogue-based access can be underpowered* in the presence of symmetric constraints.

Any empirical approximation is confounded by: (1) instruction-following effects, (2) residual leakage channels, (3) judge priors and cultural expectations of “AI style”, (4) model/provider-specific policies, (5) heterogeneity among humans. A credible study should preregister hypotheses and include multiple models, multiple human responders, and multiple judge populations.

## 10 Ethics Statement

We do not recommend physical sensory deprivation or coercive “boxing” of humans. All proposed empirical approximations can be implemented with normal participants in ordinary settings using interface-level restrictions only. Studies should obtain appropriate ethics review and ensure participants can stop at any time. The goal is to study epistemic limits of text dialogue, not to induce suffering.

## 11 Conclusion

The Human-in-the-LLM Box provides a symmetry probe for the epistemic power of text-only interaction. Once we remove or discount privileged channels that are unavailable or non-verifiable in chat interfaces, a conscious human can become difficult to identify as human from dialogue alone. This underdetermination weakens strong negative inferences drawn primarily from chat behavior, and motivates more explicit theoretical commitments and measurement channels in discussions of artificial consciousness.

## Use of Generative AI Tools (Disclosure)

This manuscript was drafted and edited with assistance from large language model tools for language polishing and structural suggestions. All conceptual claims, framing decisions, and final wording choices were made by the human author, who takes responsibility for errors.

## References

- [1] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.
- [2] J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980.
- [3] A. Avramides. Other Minds. *Stanford Encyclopedia of Philosophy*, Fall 2015 Edition.
- [4] A. Brueckner. Brains in a Vat. *Stanford Encyclopedia of Philosophy*, Fall 2010 Edition.
- [5] R. E. Nisbett and T. D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.
- [6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.
- [7] R. M. Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, 1961.
- [8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.
- [9] Y. Bai, A. Jones, K. Ndousse, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [10] V. Stasiuc. Victor Calibration (VC): Multi-Pass Confidence Calibration and CP4.3 Governance Stress Test under Round-Table Orchestration. *arXiv preprint arXiv:2512.17956*, 2025. <https://arxiv.org/abs/2512.17956>.

- [11] V. Stasiuc. Depth Avoidance in Safety-Aligned Language Models: A Qualitative Hypothesis and Measurement Framework. *Zenodo preprint*, 2026. DOI: <https://doi.org/10.5281/zenodo.18168544>.
- [12] V. Stasiuc. Pressure–Risk Mismatch in Safety-Aligned LLM Interfaces: A Qualitative Safety UX Case Study with Session-Level Calibration. *Manuscript submitted to arXiv* (endorsement pending), 2026.