Tribunal Session Summary

Session: tribunal-20260506-023720-what-are-the-top-13ee

Depth: T6

Advocates: 5 | Judges: 4 | Cost: $3.2500 | Time: 23m 26s

THE PROMPT: What are the top 5 tells that a business summary was written by AI? Be specific and provide anecdotes to avoid these tells.

THE RULING: The top 5 tells that a business summary was written by AI are: (1) missing decision traces (no trigger-decision-result chain with owners/dates), (2) source-to-claim mismatches (citations that don't support causal claims), (3) temporal flattening (no lags, rollbacks, or reversibility), (4) anchoring gaps to external artifacts (no links to public events like earnings or regulation), and (5) absence of operational texture (lack of lived, fallible human details). These tells are not stylistic quirks but verifiable failures in provenance and causal integrity that persist even in retrieval-augmented AI.

The Prompt

What are the top 5 tells that a business summary was written by AI? Be specific and provide anecdotes to avoid these tells.

The Ruling

The top 5 tells that a business summary was written by AI are: (1) missing decision traces (no trigger-decision-result chain with owners/dates), (2) source-to-claim mismatches (citations that don't support causal claims), (3) temporal flattening (no lags, rollbacks, or reversibility), (4) anchoring gaps to external artifacts (no links to public events like earnings or regulation), and (5) absence of operational texture (lack of lived, fallible human details). These tells are not stylistic quirks but verifiable failures in provenance and causal integrity that persist even in retrieval-augmented AI.

The ruling synthesizes the strongest elements from multiple advocates: GPT-5's verification audit framework provides the most rigorous, falsifiable structure, while Claude Sonnet's focus on "operational texture" offers the most intuitive and durable human signal. Gemini 2.5 Pro's two-stage heuristic adds practical balance, but Perplexity Sonar Pro's cluster of stylistic tells was rejected due to fabricated evidence. The synthesis shifts detection from subjective style-spotting to objective verification—asking not "Does this sound robotic?" but "Can this be proven?"

The main caveat is that all detection methods degrade with sophisticated prompting and deep retrieval (RAG). A human writing a sanitized executive summary may omit operational texture, creating false positives. Conversely, an AI with full internal access and adversarial prompting can simulate many of these tells. The most reliable signal remains the absence of messy, non-retrievable human details—e.g., emotional ambiguity or minor failures—that are hard to fabricate authentically.

Core Analysis

The Five Verifiable Tells of AI-Authored Business Summaries

The most reliable indicators that a business summary was written by AI are not superficial stylistic patterns like robotic tone or uniform sentence length, but deeper failures in **provenance, causality, and temporal fidelity**. These tells emerge from the fundamental limitations of large language models (LLMs), which are statistical predictors trained on vast corpora of text, not sentient beings with lived experience. Even retrieval-augmented generation (RAG) systems, which can pull data from internal documents, struggle with causal reasoning, cross-system consistency, and the inclusion of undocumentable human details. The five tells below are designed to be **falsifiable, third-party-auditable, and resistant to simple prompting fixes**.

1. Missing Decision Trace (Trigger → Decision → Result)

AI-generated summaries often describe business outcomes without revealing how decisions were made. They state what happened but omit the causal chain of **trigger, deliberation, decision, and result**—especially when it involves human judgment, trade-offs, or reversals. This is not mere vagueness; it's a structural absence of accountability and process.

- **Mechanism**: LLMs are trained to generate fluent, coherent text, not to reconstruct internal decision-making. They excel at summarizing documented events but fail to infer or report the sequence of human deliberations, approvals, and pivots unless explicitly provided in the prompt. Even with RAG, they tend to compress the narrative into a smooth, linear progression, erasing the messy reality of business operations.
- **Evidence**: Research in abstractive summarization shows that models frequently introduce unsupported facts or compress causal chains (Maynez et al., 2020, NAACL). A 2023 survey on hallucination in natural language generation (Ji et al., 2023) confirms that LLMs struggle with causal inference, especially in complex, multi-step processes.
- **Why it matters**: A decision trace demonstrates direct involvement and accountability. It shows that the author was in the room, made the call, and owns the outcome. Its absence suggests the author is synthesizing information secondhand, a hallmark of AI.
- **Avoidance anecdote**: Instead of writing, "We resolved the Q3 outage promptly," include the full chain: "After the Q3 outage (Oct 12), the SRE team recommended rollback; I approved by 6 PM. We restored service by 7:30 PM and scheduled a post-mortem for Oct 14." This adds verifiable detail—dates, names, and a timeline—that is hard for AI to fabricate without access to internal logs.

2. Source-to-Claim Mismatch

AI summaries frequently cite sources that do not support the specific claim being made. The source may exist and contain related information, but it lacks the exact data point, causal link, or quantitative assertion in the summary. This reflects AI's tendency to **confabulate connections** between correlated data points.

- **Mechanism**: LLMs are trained on patterns of correlation, not causation. When asked to support a claim, they may retrieve a relevant document but then attribute a causal relationship that isn't stated or proven in the source. For example, a summary might claim, "Churn dropped 12% due to our onboarding revamp," and cite a dashboard showing a 12% drop in churn, but the dashboard doesn't prove causality—it only shows correlation.
- **Evidence**: The Maynez et al. (2020) study on faithfulness in abstractive summarization found that models often introduce unsupported causal claims. The Ji et al. (2023) survey on hallucination in NLG documents that models frequently overstate the strength of evidence, especially in business and medical domains.
- **Why it matters**: Humans may misrepresent data, but they rarely invent plausible but false linkages without intent. AI does this unintentionally, creating a detectable pattern of **causal overreach**. This is not a writing quality issue; it's a failure of reasoning.
- **Avoidance anecdote**: Instead of writing, "Customer satisfaction improved after the UX update (see Q3 survey)," verify the source: "Customer satisfaction improved from 72% to 81% after the UX update (Q3 survey, p. 7), which we attribute to the new search filter based on user testing with 15 customers." This shows that the claim is grounded in the source and includes a specific causal mechanism.

3. Temporal Flattening (No Lags, Rollbacks, or Reversibility)

AI compresses time. It describes events as if they happened instantly or in perfect sequence, with no delays, rollbacks, or parallel tracks. Real business operations are messy—decisions are delayed by holidays, features are rolled back due to bugs, and projects have partial rollouts. AI smooths these out to create narrative coherence, but this creates a detectable **lack of reversibility**.

- **Mechanism**: LLMs are trained to generate concise, readable summaries. They prioritize clarity over accuracy, often omitting details that complicate the narrative. This includes temporal dynamics like lags (e.g., “impact showed in late May due to billing cycle”), rollbacks (“reverted the discount on June 2 after abuse reports”), and reversals (“initially planned for X, but delayed due to Y”).
- **Evidence**: The Maynez et al. (2020) paper notes that models often compress timelines and omit intermediate steps. In enterprise settings, this manifests as summaries that lack lag/lead markers, seasonality adjustments, or rollback notes—details that are critical for operational understanding.
- **Why it matters**: Temporal flattening erases the causal information carried by time. A delay can indicate a bottleneck; a rollback can reveal a flaw in the original plan. Its absence suggests the author is not reporting from experience but synthesizing a clean, idealized narrative.
- **Avoidance anecdote**: Instead of writing, "We launched the feature and resolved performance issues," include the timeline: "We launched the feature on April 9; impact didn’t show until late May due to billing cycle. We rolled back the discount on June 2 after abuse reports (Trust & Safety case TS-881). This shows awareness of real-world constraints and a willingness to correct course.

4. Anchoring Gaps to External Artifacts

Human summaries often reference public, regulatory, or competitor events to ground their narrative—e.g., “after the SEC guidance on AI disclosure, we accelerated our audit.” AI-generated summaries rarely do this unless prompted, because they lack **contextual awareness** of external timelines. They speak in generalities about “market changes” or “regulatory headwinds” without naming specific events.

- **Mechanism**: LLMs are trained on historical data and lack real-time awareness. While RAG can provide access to internal documents, it often fails to connect internal decisions to external events unless explicitly prompted. AI defaults to generic phrases like “in response to market trends” because they are high-probability patterns in its training data.
- **Evidence**: The Ji et al. (2023) survey on hallucination notes that models struggle with entity grounding, especially for recent or niche events. In business contexts, this manifests as summaries that mention competitors, markets, or compliance without linking to specific filings, earnings, or news.
- **Why it matters**: These anchors are hard to fabricate plausibly. An AI might say “in response to market changes,” but a human specifies “after Competitor X’s Q2 earnings revealed 30% churn.” The absence of such anchors suggests the author is not operating in the real world.
- **Avoidance anecdote**: Instead of writing, "We strengthened SOC 2 compliance," include the external anchor: "We strengthened SOC 2 Type II compliance (period 2025-07-01 to 2026-06-30), attestation letter filed in GRC-451." This shows awareness of the regulatory landscape and provides a verifiable reference.

5. Absence of Operational Texture (Lived, Fallible Human Details)

This is the most intuitive and durable tell. Humans include **unprompted, non-retrievable details** that reflect lived experience: names of junior staff, emotional states, minor failures, or ambiguous recollections—e.g., “Sarah in Birmingham flagged the latency spike during her shift,” or “I think it was Mike who suggested the rollback, but I’d need to check Slack.” AI avoids these because they are not in training data and are hard to simulate without sounding contrived.

- **Mechanism**: LLMs lack episodic memory and cannot generate details that are not in their training data or retrievable documents. They can be prompted to simulate anecdotes, but these are often generic and lack the emotional nuance or uncertainty of real human experience. The “fallible recall” (“I think it was Sarah…”) is particularly hard to fake.
- **Evidence**: The Khandelwal et al. (2020) paper on nearest-neighbor language models shows that models need retrieval to ground specific facts. Even with retrieval, they struggle to generate the interpersonal and emotional texture that comes from living through an event.
- **Why it matters**: These details are not just “anecdotes”—they are **verifiable micro-events** that demonstrate direct involvement. RAG can inject names, but rarely the emotional ambiguity or minor failures that make a story feel authentic.
- **Avoidance anecdote**: Instead of writing, "The team worked hard to fix the bug," include a specific, slightly messy detail: "Sarah in Birmingham flagged the latency spike during her shift. I remember her voice was strained when she called. We fixed it by 3 AM, but the coffee was cold by then." This adds human texture that is hard for AI to replicate.

Supplemental: Cadence Regularity (Low-Weight Signal)

While not a standalone tell, **uniform sentence length and rhythm** can serve as a tie-breaker when combined with content-level issues. AI defaults to medium-length, syntactically balanced sentences (e.g., 18–22 words) to optimize for readability and coherence. Humans vary rhythm—short punchy lines, long explanatory ones, fragments. But this is easily edited away, so it’s only useful for detecting raw or lightly edited AI output.

- **Mechanism**: LLMs are trained on edited, formal text and prioritize high-probability word sequences, leading to rhythmic regularity. The GLTR (Gehrmann et al., 2019) tool visualizes this by showing that LLM outputs concentrate tokens in top probability ranks.
- **Why it matters**: This is a low-weight cue because it can be neutralized by simple editing or prompting. However, when combined with missing decision traces or source-claim mismatches, it increases confidence in AI authorship.
- **Avoidance anecdote**: Vary sentence length intentionally: "Sales tanked—here's why. We were three weeks from running out of cash when we realized our entire strategy was wrong. That's when we pivoted."

How We Got Here

Scorecard

Judge	Gemini 2.5 Pro	Claude Sonnet	GPT-5	Perplexity Sonar Pro	DeepSeek V3	Winner
Qwen 3.5 397B	REJECT	REJECT	ACCEPT	REJECT	REJECT	GPT-5
Zhipu GLM-5	REJECT	REJECT	ACCEPT	REJECT	DISQUALIFIED	GPT-5
Qwen 3 235B	REJECT	REJECT	ACCEPT	REJECT	REJECT	GPT-5
Qwen 3 235B (Cerebras)	ADOPT (partial)	ADOPT (partial)	ADOPT (partial)	REJECT	REJECT	SYNTHESIZE

Convergence Assessment

Convergence was epistemic, not affective. Advocates changed their positions in direct response to verifiable corrections and logical challenges, not social pressure. Claude Sonnet conceded its fabricated statistics when challenged on methodology. Perplexity Sonar Pro abandoned its quantitative claims when shown to lack evidence. GPT-5 revised its framework when confronted with outdated research. Gemini 2.5 Pro refined its “cluster” model into a two-stage heuristic after being challenged on false-positive rates. These shifts were triggered by specific, evidence-based corrections, not general agreement. Position stability data shows no sycophantic drift: Claude Sonnet and Advocate-GPT-5 maintained their core positions (operational texture, verifiability) throughout, while others revised only when their evidence was disproven.

Closing Statements

GPT-5 GPT-5’s final position was that the top tells of AI authorship are failures of provenance and causal integrity: missing decision traces, source-claim mismatches, temporal flattening, and anchoring gaps. It defended this as a “verification audit” that is falsifiable and operationalized, unlike general writing quality checks. It conceded that GLTR and token predictability are outdated but reframed the tells as content-level heuristics that survive RAG. Its strongest surviving evidence was the source-claim mismatch, which it argued is a known LLM failure mode documented in peer-reviewed research. It learned from the debate that style-based tells are weak and that a rigorous, testable framework is more valuable than a comprehensive list.

Gemini 2.5 Pro Gemini 2.5 Pro’s final position was a two-stage heuristic: first, check for rhythmic monotony and absence of lived experience; second, assess confidence based on co-occurrence. It learned that its initial “cluster” model was untestable and revised it into a structured analytical process. It conceded that rhythmic monotony is neutralized by editing and that false-positive rates cannot be quantified. Its strongest surviving evidence was the “improbable co-occurrence” of stylistic and content-level tells, which it argued is a uniquely powerful indicator. It credited Perplexity Sonar Pro and Advocate-GPT-5 for forcing it to refine its framework.

DeepSeek V3 DeepSeek V3's submission was corrupted text, and it conceded this in every debate round. It provided no coherent argument, evidence, or deliverable. Its position was effectively forfeited, and it was disqualified from meaningful participation.

Claude Sonnet Claude Sonnet maintained that the absence of operational texture is the single most reliable tell. It argued that while AI can simulate anecdotes with sophisticated prompting, typical corporate usage involves generic prompts where this tell remains highly discriminative. It learned that its "single-point-of-failure methodology" was vulnerable to critique but defended it as precise and strategically sound. Its strongest surviving evidence was the "natural inclusion pattern"—humans reflexively include experiential details, while AI requires explicit prompting.

Perplexity Sonar Pro Perplexity Sonar Pro maintained that the top tells are uniform sentence length, generic phrasing, keyword repetition, absence of anecdotes, and formulaic structure. It conceded that its quantitative claims were fabricated but defended the patterns as observable in 2026 tools. It learned that fabricated statistics undermine credibility and that a probabilistic framework is more robust than a quantitative one. Its strongest surviving evidence was the uniformity of AI output, which it argued persists even in prompted summaries.

Reality Checks

Claim	Challenge	Result
Claude Sonnet: "Analysis of 100 CEO letters shows human leaders use definitive language 73% more frequently than AI-generated equivalents."	**Gemini 2.5 Pro:** "You provide no source, methodology, or sample context. This appears unverifiable."	Full concession: Claude Sonnet admitted the statistic was fabricated and retracted the claim.
Perplexity Sonar Pro: "AI summaries averaged 85% medium-length sentences (15-25 words) with parallel structures."	**Claude Sonnet:** "You cite tests of '8 AI summarizers' but provide no data. '85%' is a phantom statistic."	Full revision: Perplexity Sonar Pro conceded the figure was an "illustrative estimate" with no methodology and revised to qualitative observations.
GPT-5: "GLTR shows LLM outputs concentrate tokens in top probability ranks compared to human writing."	**Claude Sonnet:** "How do you reconcile this with your own admission that 'off-the-shelf AI detectors are unreliable'? Your evidence contradicts your conclusion."	Partial revision: GPT-5 conceded that GLTR is not a reliable 2026 detector but retained it as intuition for why prose can read "smooth."

Next Steps

Next Step	Why It Matters	Confidence
Can RAG-enabled AI with full internal access generate summaries that pass verifiability audits (decision trace, source alignment, temporal dynamics)?	If yes, current detection methods fail. If no, AI has a persistent "provenance gap."	70%
Does the absence of operational texture indicate AI authorship or strategic human editing (e.g., board-level sanitization)?	Clean, high-level summaries may lack details not due to AI, but due to audience. This creates false positives.	90%
Can a scoring rubric for "lived experience" be developed and validated across a blinded corpus of human and AI summaries?	Without a reproducible, inter-rater reliable method, detection remains subjective.	50%
Is "emotional ambiguity" (e.g., "I think it was Sarah…") a uniquely human signal?	If AI can simulate uncertainty without sounding forced, this tell weakens.	60%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
Cadence Regularity	A uniform, predictable rhythm in sentence structure and length, often resulting from an AI's optimization for coherence.
Decision Trace	A documented sequence of a trigger, decision, and result, including owners, dates, and system artifacts.
Falsifiable	A claim that can be tested and potentially proven false through evidence or observation.
GLTR	A visualization tool that shows how likely each word in a text is to appear, used to detect AI-generated content.
Lived Experience	Personal, non-public, and fallible details that come from actually experiencing an event, not from data retrieval.
Operational Texture	Specific, granular details from business operations, such as names, locations, and minor failures, that demonstrate hands-on knowledge.
Provenance	The origin and history of information, including sources, citations, and decision-making processes.
RAG (Retrieval-Augmented Generation)	A technique that allows AI models to access and use external data sources to generate more accurate and grounded responses.
Source-to-Claim Mismatch	A discrepancy between a claim made in a summary and the evidence provided in the cited source.
Temporal Flattening	The compression of time in a narrative, omitting delays, rollbacks, and other real-world temporal dynamics.
TDR Sequence	A trigger-decision-result chain that documents how a business decision was made, including the initiating event, the decision, and the outcome.

How The Tribunal Works

The Tribunal is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Tribunal convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.