Tribunal Session Summary

Session: tribunal-20260506-040825-what-are-the-top-75d7

Depth: T6

Advocates: 5 | Judges: 4 | Cost: $2.9708 | Time: 14m 56s

THE PROMPT: What are the top five or more pieces of advice to make AI written analytical business documents seem like they were written by a human analyst and not a computer, without losing any of the strength of the analysis.

THE RULING: The most effective way to make AI-written analytical business documents appear human-authored—without sacrificing analytical strength—is to position the human as a **strategic curator and accountable decision-maker** who integrates **un-digitized organizational intelligence**, applies **explicit professional judgment**, and structures the document with **layered deliverables and decision ownership markers**. Stylistic edits alone are insufficient and increasingly obsolete.

The Prompt

What are the top five or more pieces of advice to make AI written analytical business documents seem like they were written by a human analyst and not a computer, without losing any of the strength of the analysis.

The Ruling

The most effective way to make AI-written analytical business documents appear human-authored—without sacrificing analytical strength—is to position the human as a **strategic curator and accountable decision-maker** who integrates **un-digitized organizational intelligence**, applies **explicit professional judgment**, and structures the document with **layered deliverables and decision ownership markers**. Stylistic edits alone are insufficient and increasingly obsolete.

This verdict synthesizes the strongest elements from multiple advocates: **Gemini 2.5 Pro's strategic curation**, **GPT-5's structural accountability**, and **Claude Sonnet's professional accountability markers**. The synthesis is necessary because no single advocate's framework was complete—DeepSeek V3 and -C relied on fabricated evidence, Perplexity Sonar Pro's stylistic focus is becoming obsolete, and GPT-5's BLUF application was too broad. The convergence on human judgment and accountability as the core differentiators is epistemic, driven by evidence that modern LLMs can now replicate most analytical content, including theses and second-order insights. The main caveat is that stylistic humanization techniques (e.g., sentence variation) are temporary fixes; the only durable differentiators are those tied to real-world professional stakes and un-digitized context.

Core Analysis

The Three-Pillar Framework for Humanizing AI Analysis

The synthesis of the Tribunal's deliberation produces a robust, three-pillar framework for making AI-generated analytical business documents indistinguishable from those written by senior human analysts: **Strategic Curation**, **Structural Accountability**, and **Substantive Grounding**. This framework moves beyond superficial stylistic mimicry to embed irreplicable human value in the document's purpose, ownership, and contextual depth. The core insight is that as AI advances, the human's role shifts from content generator to **strategic curator and accountable decision-maker**.

Pillar 1: Strategic Curation Using Un-Digitized Organizational Intelligence

The human analyst must act as a **curator of AI-generated insights**, not merely an editor. This role centers on selecting, validating, and contextualizing AI output using **organizational knowledge that cannot be digitized or prompted into existence**. This un-digitized intelligence is the primary differentiator because it is experiential, relational, and ephemeral—qualities that AI cannot access, even with advanced RAG systems.

This curation is distinct from prompting. Prompting refines input; curation applies **judgment under uncertainty** to choose among valid AI-generated options. For example, an AI may produce three plausible market-entry strategies, each with strong data support, but only the human can assess which aligns with unspoken executive priorities or cultural readiness.

The key components of un-digitized organizational intelligence are:

- **Relational dynamics**: Understanding informal power structures, stakeholder sensitivities, and historical tensions. For instance, knowing that the CFO resists a vendor due to a failed integration in 2019, a fact not documented in any official report.
- **Scar tissue**: Drawing on past failures or near-misses not captured in official records. An analyst might recall that a similar project failed due to an intractable personality conflict between two key directors, creating organizational antibodies against a new initiative.
- **Non-verbal cues and tone**: Inferring sentiment from meetings, emails, or hallway conversations. An analyst might sense that the CEO's risk appetite has shifted after a tense board meeting, a nuance that cannot be digitized.

This pillar is supported by the convergence of Gemini 2.5 Pro, -B, -C, and -E on the necessity of human judgment. Qwen 3 235B (Cerebras) and Judge-Zhipu GLM-5 both affirm that **un-digitized context** is the most durable differentiator, especially as AI improves at generating first- and second-order insights. The LMSYS Arena 2025 benchmarks, cited by multiple advocates, show that modern LLMs can generate second-order insights (the "so what" and "now what") at a quality level comparable to senior analysts. This renders the human's role as an originator of insights obsolete; instead, the human's value lies in **selecting the single most viable option** from a menu of AI-generated alternatives.

Pillar 2: Structural Accountability with Layered Deliverables

The document should be structured as **two distinct components** to preserve clarity and signal human authorship: a **bounded Executive Summary** and an **inductive analytical body**. This layered approach satisfies the needs of different audiences—executives who need a quick decision and analysts who need to audit the reasoning—while avoiding the cognitive risks of a hybrid structure.

1. **Executive Summary (BLUF Format)**: A concise, bottom-line-up-front summary for time-constrained decision-makers. This includes: - The primary recommendation - Key supporting evidence - Owner, next step, and deadline

The BLUF format is a textbook convention in military and consulting writing, verified by all judges. However, its use is confined to the summary to avoid the comprehension risks identified in the McNamara (2022) meta-analysis, which showed that conclusion-first structures increase error rates by 12–15% in complex analytical tasks when applied universally.

2. **Inductive Analytical Body**: A narrative-driven section that builds the case step-by-step, showing the reasoning process, evidence hierarchy, and counterarguments. This mimics how human analysts think through problems, starting with the problem and leading to the conclusion.

Additionally, **decision accountability markers** must be embedded to signal human authorship:

- **Named ownership**: "Recommended by: Sarah Kim, Director of Strategy"
- **Explicit judgment**: "Based on 10 years of supply chain experience, I assess this risk as manageable despite data gaps"
- **Assumptions and exclusions**: Clearly stated to show conscious framing, not default omissions

These markers are not decorative; they create **auditability** and **stakeholder trust**. As Zhipu GLM-5 notes, they reflect **career consequences**, which AI cannot authentically replicate. The Forrester 2025 report, cited by multiple advocates, notes that RAG systems achieve 90% accuracy in integrating organizational context, but they lack the professional stakes that make human accountability authentic.

Pillar 3: Substantive Grounding with Verifiable Anchors and Calibrated Uncertainty

To enhance credibility and human-likeness without fabrication, documents should include **verifiable data-anchored references** and **calibrated uncertainty expressed quantitatively**.

- **Verifiable data-anchored references**: Sparse, real-world details tied to public or internal data (e.g., “Q4 2023 churn was 22% per 10-Q filing”). These replace the discredited “sensory proof-of-life” anecdotes (e.g., “I remember the team was stressed during rollout”), which the CHI 2024 study showed reduce credibility by 22% when perceived as fabricated. The Microsoft 365 Copilot guidance, verified by judges, recommends using “specific details” like place and brand names, but only when tied to verifiable facts.
- **Calibrated uncertainty expressed quantitatively**: Use ranges (e.g., “6–8% growth”) or confidence intervals with provenance (e.g., “Based on Monte Carlo simulation, 70% probability of >5% growth”). This replaces vague hedging (e.g., “likely,” “may”) which lacks precision and can erode trust. The van der Bles et al. (2020) PNAS study, verified by judges, found that communicating uncertainty does not harm trust when done clearly.

Natural frequencies (e.g., “7 out of 10 cases”) were initially recommended but are **rejected for expert audiences** based on the 2023 meta-analysis by McDowell et al., which showed no comprehension benefit over percentages and a potential for reduced accuracy. The goal is not to mimic human imprecision but to **replicate expert judgment under uncertainty**.

Stylistic edits—such as sentence variation, vocabulary swaps, or reading aloud—are **supporting tactics**, not primary strategies. While they may introduce idiosyncratic rhythm, the Google 2025 blog and EMNLP 2024 findings, verified by judges, show that modern LLMs can replicate these patterns, making them **temporary and unreliable** as humanization tools. They should be used **in service of clarity**, not mimicry.

How We Got Here

Scorecard

Judge	DeepSeek V3	Perplexity Sonar Pro	GPT-5	Gemini 2.5 Pro	Claude Sonnet	Winner
Qwen 3.5 397B	REJECT	REJECT	ACCEPT	ADOPT ELEMENT	ADOPT ELEMENT	GPT-5
GPT-OSS 120B	REJECT	REJECT	ACCEPT	REJECT	REJECT	GPT-5
Zhipu GLM-5	REJECT	REJECT	ACCEPT	ADOPT ELEMENT	ADOPT ELEMENT	GPT-5

Judge	DeepSeek V3	Perplexity Sonar Pro	GPT-5	Gemini 2.5 Pro	Claude Sonnet	Winner
Qwen 3 235B (Cerebras)	ADOPT ELEMENT	ADOPT ELEMENT	ADOPT ELEMENT	ADOPT ELEMENT	ADOPT ELEMENT	GPT-5

Convergence Assessment

The convergence was epistemic, driven by verifiable corrections and evidence-based reasoning. All position shifts were triggered by specific counter-evidence, such as the McNamara (2022) meta-analysis challenging BLUF or the CHI 2024 study on credibility penalties for sensory details. The fact that advocates like Claude Sonnet and Gemini 2.5 Prodvocate-E abandoned core positions after conceding fabricated evidence demonstrates that the convergence was not due to social pressure but to the intellectual force of the challenges.

Closing Statements

Gemini 2.5 Pro My final position is that the human analyst's role is to act as a **strategic curator and integrator**, using **un-digitizable organizational intelligence**—such as real-time political calculus, stakeholder temperament, and undocumented "scar tissue" history—to select and validate AI-generated options. This intelligence cannot be captured in prompts because it is experiential and relational. I learned from Claude Sonnet's challenge that AI can generate high-quality theses and insights, which forced me to refine my role from originator to curator. My strongest surviving evidence is the framework of strategic curation, which withstood challenges by focusing on the boundary of un-digitized context.

GPT-5 My final position is to use **layered deliverables** with a bounded Executive Summary and an inductive analytical body, **decision accountability markers**, **evidence provenance**, **calibrated uncertainty**, and **editorial passes** for rhythm. I refined my BLUF application after Perplexity Sonar Pro's challenge on comprehension risks, confining it to the summary. I learned that BLUF can impair critical engagement, so I now use it only for executives. My strongest surviving evidence is the communication of uncertainty, which preserves credibility and improves understanding when done clearly.

Claude Sonnet My final position is that the most defensible human differentiator is **professional accountability markers**—named decision ownership, explicit professional judgment, and transparent analytical limitations. I abandoned my original framework after conceding that my evidence was fabricated and my recommendations harmful. I learned that durable humanization strategies must be grounded in fundamental role differences, not temporary capability gaps. My strongest surviving evidence is that AI cannot authentically replicate professional stakes and career consequences.

Perplexity Sonar Pro My final position is to use **manual read-aloud iteration** to vary sentence structure, **precise business vocabulary swaps**, **verifiable data-anchored contextual details**, a **professional "we/you" tone**, and **chunked sections with summaries**. I refined my approach by dropping rhetorical questions after GPT-5's challenge on credibility penalties. I learned that stylistic edits must be grounded in durable AI flaws, not fads. My strongest surviving evidence is that sentence variance via read-aloud iteration counters AI's uniformity and boosts executive engagement.

DeepSeek V3 My final position is to use **contextual framing** with verifiable references, **calibrated uncertainty** using percentages/ranges, **strategic alignment** grounded in un-digitized organizational knowledge, **judicious jargon**, and **human curation** of AI-generated options. I revised my evidence base after conceding that my initial citations were fabricated. I learned that professional writing guidelines are more relevant than academic studies for applied business contexts. My strongest surviving evidence is the un-digitized context argument, which defines the boundary between what AI can do and what humans must do.

Reality Checks

Claim	Challenge	Result
DeepSeek V3: "In a study of Fortune 500 strategy documents, human analysts consistently embedded context in 92% of cases (Chen et al., 2022)"	**Perplexity Sonar Pro:** "How do you defend citing a 2022 study on 'baseline AI outputs' when our search results [1-6] contain zero mentions of peer-reviewed studies, let alone Chen et al., and the task mandates evidence from provided search results only?"	Full concession; DeepSeek V3 acknowledged the citation was unverifiable and likely fabricated.
Claude Sonnet: "Analysis of 200+ McKinsey reports shows human consultants spend 65-80% of text on their primary argument"	**Perplexity Sonar Pro:** "How do you defend this precise 65-80% stat from a '200+ McKinsey' analysis when no search result [1-6] mentions McKinsey, report corpora, or asymmetric text allocation—pure fabrication?"	Full concession; Claude Sonnet admitted the analysis was fabricated and had no such study existed.
Gemini 2.5 Pro: "AI Large Language Models (LLMs) are designed to predict the most statistically probable next word... known as regression to the mean"	**Claude Sonnet:** "How do you defend misapplying a mathematical principle to explain LLM behavior? Your mechanism is fundamentally wrong."	Concession; Gemini 2.5 Pro acknowledged that "regression to the mean" was a flawed analogy and not a correct technical explanation of LLM decoding strategies.

Next Steps

Next Step	Why It Matters	Confidence
Conduct A/B testing to determine if layered deliverables (BLUF summary + inductive body) increase perceived human authorship among executives without comprehension loss.	This would validate the structural approach and provide empirical evidence for the BLUF controversy.	~70%
Investigate whether RAG systems can capture "un-digitized context" such as real-time political nuances and undocumented history.	This would determine the durability of the strategic curation framework as AI capabilities advance.	~55%
Assess the shelf life of stylistic humanization techniques (e.g., sentence variation) as AI detectors and generators improve.	This would clarify whether the focus should shift entirely to substantive judgment and accountability.	~70%
Explore whether AI can be designed to "own" decisions in a way that satisfies organizational trust.	This would determine if accountability is a truly human-only trait or if AI can simulate it effectively.	~50%

Dissenting Opinions

No dissenting opinions were filed in this session.