Tribunal Session Summary

Session: tribunal-20260506-044101-what-give-material-changes-03cc

Depth: T2

Advocates: 4 | Judges: 1 | Cost: $0.9726 | Time: 5m 39s

THE PROMPT: What give material changes would you make to the attached prompt?

THE RULING: Yes, the Analytical Writing Standard requires material changes to evolve from a post-hoc hygiene checklist into a proactive reasoning framework that ensures the logical integrity of analysis, not just its defensibility. The most critical addition is a mandatory pre-commitment process—**Gate-R**—that forces analysts to explicitly identify and stress-test the *warrant*, the unstated logical bridge between evidence and claim, before drafting a recommendation. This is complemented by a Claim–Evidence Map and Metrics Sanity-Check Appendix for traceability, and workflow-specific entry points to improve usability under pressure.

The Prompt

What give material changes would you make to the attached prompt?

The Ruling

Yes, the Analytical Writing Standard requires material changes to evolve from a post-hoc hygiene checklist into a proactive reasoning framework that ensures the logical integrity of analysis, not just its defensibility. The most critical addition is a mandatory pre-commitment process—**Gate-R**—that forces analysts to explicitly identify and stress-test the *warrant*, the unstated logical bridge between evidence and claim, before drafting a recommendation. This is complemented by a Claim–Evidence Map and Metrics Sanity-Check Appendix for traceability, and workflow-specific entry points to improve usability under pressure.

The ruling synthesizes the strongest elements from multiple advocates: **Gemini 2.5 Pro**'s core insight that the standard fails to police the *warrant* and must adopt `Argument Cartography`; **GPT-5**'s enforceable artifacts for traceability and consistency; and **Claude Sonnet**'s usability-focused restructuring. It rejects **Perplexity Sonar Pro**'s defense of the status quo, which relies on fabricated evidence and ignores the standard's deepest flaw. The synthesis reflects a majority consensus that the standard must shift from output sanitation to input logic validation, using AI not as a ghostwriter to be managed, but as a cognitive partner to be leveraged for stress-testing assumptions.

Core Analysis

The Core Problem: Sanitizing Output, Not Securing Logic

The Analytical Writing Standard is exceptionally strong at ensuring its *output*—the final memo or brief—is defensible, human-sounding, and free of AI "tells." Its rules on named decision traces, source-to-claim tightness, defined metrics, and objection-mapping are a powerful defense against the most common failure modes of AI-assisted writing: vagueness, fabrication, and uniform confidence. However, as **Gemini 2.5 Pro** correctly identified, the standard's exclusive focus on the final document represents a critical strategic blind spot. It polices the *evidence* but fails to interrogate the *logic* that connects that evidence to a conclusion.

This creates a dangerous vulnerability: an analyst can produce a document that passes every test in the standard—named owners, dated milestones, defined metrics, and a trade-off stated—yet still rest on a catastrophic logical flaw. The flaw is not in the data, but in the *warrant*, a core component of formal argumentation defined by the Toulmin Model. The warrant is the unstated belief or logical rule that connects the *grounds* (the evidence) to the *claim* (the recommendation). For example, an analyst might present data showing low user engagement (*grounds*) and recommend sunsetting the product (*claim*). The standard would ensure the engagement data is sourced and defined correctly. It does not, however, force the analyst to articulate the *warrant*: "We, as a business, should always divest from products with low engagement." Is this principle true? What if the product is a loss-leader or a strategic requirement for a key customer? By

leaving the warrant implicit, the standard leaves the most important part of the argument unexamined and unchallenged.

The fact-check matrix verified this gap: "Correct. The Toulmin Model's 'warrant' is absent from the standard's terminology and structure." This is not a minor oversight; it is the difference between a document that *sounds* defensible and one that *is* sound. The standard's current objection-mapping, which asks the analyst to list "three to five objections a skeptical reader will raise," is a freeform checklist item. It does not mandate a systematic, adversarial process targeted at the warrant itself. The analyst who developed the recommendation is the person least likely to see its deepest flaws due to confirmation bias. The standard, as written, does nothing to mitigate this cognitive failure.

The Solution: Gate-R and Argument Cartography

The synthesis addresses this gap by introducing **Gate-R**, a mandatory pre-commitment checkpoint that must be passed before any recommendation is issued. Gate-R operationalizes `Argument Cartography`, a three-step process that forces the analyst to map and stress-test the logical skeleton of their argument *before* drafting prose. This shifts the standard from a *post-hoc checklist* to a *preemptive reasoning framework*.

1. Explicit Warrant Statement: For each key claim, the analyst must state the logical rule or assumption that connects the evidence to the conclusion. This is not a restatement of the claim or the grounds, but the underlying principle of reasoning. For example: "We assume that a 10-point increase in NPS will lead to a 3-point increase in retention because historical cohort data (Q1–Q3) shows a 0.3x elasticity." This surfaces hidden assumptions for scrutiny. The warrant is the core of the argument, and making it explicit is the single most important change to the standard.

2. AI-Simulated Red-Teaming: Before drafting, the analyst uses an LLM to generate objections from diverse, skeptical personas (e.g., CFO focused on margin, Chief Counsel focused on risk, VP of Engineering focused on integration). A prompt might be: *"You are a panel of expert reviewers. The recommendation is: 'We should acquire Company X for $50M to enter the European market.' Generate the top 2 most pressing questions/objections from a CFO, a Chief Legal Officer, and a VP of Engineering, with a brief justification for their concern."* The goal is not to replace human judgment but to augment it, mitigating the analyst's own cognitive entrenchment and confirmation bias. While the fact-check verdict on this claim is "plausible but not evidenced," the *plausibility* of cognitive diversity gains supports its use as a *supplemental* tool. The strongest surviving evidence from **Gemini 2.5 Pro** is that "the analyst who creates a recommendation is, by definition, the person most susceptible to confirmation bias and is therefore the least equipped to find its flaws." AI red-teaming is a modern implementation of established techniques like formal Red Teaming, designed to counter this exact human failing.

3. Pre-Prose Modeling: Analysts must sketch the argument structure visually or quantitatively—e.g., a causal loop diagram, decision tree, or simple simulation model—before writing prose. This forces early engagement with logic and uncertainty, reducing the risk of retrofitting evidence to a predetermined conclusion. The standard currently treats quantitative and visual reasoning as secondary, placing "all charts, raw data, supporting materials" in an appendix. This is a layout rule, not an analytical constraint, but the structure implicitly frames the process as "write the argument, then put the charts at the end." **Gemini 2.5 Pro**'s revised position correctly identifies that the standard "fails to *guide* it" as a primary mode of thinking. Pre-prose modeling elevates this to a core step, ensuring the model or visualization is not just an illustration, but the site of insight discovery.

Gate-R does not block exploration; it gates *commitment*. Analysts can explore freely, draft hypotheses, and collaborate with AI in "Explore Mode." The gate triggers only when they attempt to advance a document as a formal recommendation for decision. This balances rigor with creativity, respecting the iterative nature of analysis while ensuring that only well-vetted logic advances to decision-makers.

Enforceable Artifacts for Traceability and Consistency

While **Gemini 2.5 Pro**'s focus on logic is paramount, **GPT-5** correctly identified that the standard lacks enforceable mechanisms for traceability and quantitative consistency. The synthesis adopts two lightweight, high-value artifacts from **GPT-5**'s proposal, revised to avoid "defensibility theater."

1. Claim–Evidence Map (Appendix B): A structured table listing each key claim, its supporting evidence (with source, date, and page/ticket ID), the type of causal link (e.g., correlation, causation, projection), and the named owner of the judgment. This is not embedded in the narrative to preserve readability but must be completed before submission. It enables machine-assisted review for missing links or unsupported leaps. The fact-check verified that "CONSORT/PRISMA checklists improve reporting quality," providing a strong analogical basis for this artifact in knowledge work. The map is not a database; it is a concise, checkable summary that ensures every consequential claim is anchored, directly addressing the standard's "source-to-claim tightness" rule with a verifiable artifact.

2. Metrics Sanity-Check Appendix (Appendix C): A short validation log confirming internal consistency of all reported metrics. For example: "GM 58%, CM 42% → CM < GM → ■"; "ARR growth 25%, new bookings 20% → requires contraction in churn → assumption stated in Section 5." This prevents KPI drift and arithmetic contradictions—common failure modes even in otherwise rigorous documents. The fact-check verified that "contribution margin cannot exceed gross margin" is a textbook constraint. This appendix makes the standard's existing requirement for "sanity-check internal consistency" operational and visible, catching errors that prose reviews often miss.

Workflow Optimization for Usability

Claude Sonnet correctly identified that the standard's 10-section structure, while comprehensive, is not optimized for use under deadline pressure. The most actionable tools—the Composite Check and Litmus Tests—appear at the end after 3,000+ words of theory. The synthesis adopts **Claude Sonnet**'s core insight of workflow optimization, rejecting their flawed solutions (arbitrary thresholds, document splitting) in favor of a more sophisticated approach.

The standard should be restructured into **three workflow-specific entry points**:

- **System Prompt Mode:** For AI-assisted drafting. A condensed version of sections 1–8, formatted as a directive prompt.
- **Self-Edit Checklist Mode:** For human review. A streamlined version of the Composite Check and Litmus Tests, designed for quick scanning.
- **Review Rubric Mode:** For feedback delivery. A scored matrix mapping each rule to observable traits (e.g., "Named ownership: present/absent," "Trade-off stated: yes/no").

These modes use **progressive disclosure**: users see only the tools relevant to their role and stage. A junior analyst drafting a memo sees the system prompt; a VP reviewing it sees the rubric. This preserves rigor while reducing cognitive load. Diagnostic questions replace rigid thresholds. For example: "Did you consider an alternative interpretation of this data?" instead of "Sentence length variance must exceed 40%." The latter is arbitrary and gameable; the former invites reflection.

Rejected Elements

The synthesis explicitly rejects several proposed changes:

- **Mandatory front-matter forms** (e.g., decision, deadline, assumptions) are rejected. They add friction without proportional gain. Accountability is already enforced via Section 6 (Named Ownership, Judgment Basis).
- **Binary style thresholds** (e.g., sentence length variance rules) are rejected as arbitrary, gameable, and irrelevant to reasoning quality. The fact-check verdict on this was "arbitrary threshold with no supporting research."
- **No material changes** (the position of **Perplexity Sonar Pro**) is rejected. The standard *does* need material changes. The advocate's initial reliance on fabricated data ("Grok-internal evals (2025)")—a direct violation of the standard's own "Hard rule"—undermines their credibility.

The revised standard does not aim to make writing *indistinguishable* from human output; it aims to make the *reasoning* sound. The "AI tell" is not the problem—it's a symptom of weak logic. Fix the logic, and the tell diminishes.

How We Got Here

Scorecard

Judge	Gemini 2.5 Pro	GPT-5	Claude Sonnet	Perplexity Sonar Pro	Winner
Qwen 3 235B	ACCEPT	ADOPT [element]	ADOPT [element]	REJECT	Gemini 2.5 Pro

Convergence Assessment

Convergence was epistemic, not affective. All shifts in position were triggered by verifiable corrections and logical challenges, not social pressure. **Gemini 2.5 Pro** revised their proposal to include an enforceable gate after **GPT-5** challenged the lack of operational safeguards. **GPT-5** revised their gating and citation format after **Gemini 2.5 Pro** and **Claude Sonnet** challenged their readability and friction. **Claude Sonnet** abandoned their entire mechanistic approach after **Gemini 2.5 Pro** challenged its gameability. **Perplexity Sonar Pro** revised their evidence from unverifiable internal data to public peer-reviewed sources after being challenged on fabrication. Position stability data shows this clearly: **Gemini 2.5 Pro** (3), **GPT-5** (3), **Claude Sonnet** (4), **Perplexity Sonar Pro** (1). The fact that all advocates changed their mind in response to evidence, and no one shifted to conform to a majority, confirms that the convergence was driven by the adversarial exchange of facts and logic.

Closing Statements

Gemini 2.5 Pro In their closing statement, **Gemini 2.5 Pro** refined their position into a powerful, implementable policy. They acknowledged that their initial proposal was a "philosophy without an enforcement mechanism" and integrated **GPT-5**'s critique to make it operational with a "Gate-0" prerequisite. Their strongest surviving evidence—the distinction between *grounds* and *warrant*—was presented with clarity and depth, using the Toulmin Model to demonstrate a profound gap in the standard. They conceded on the superiority of AI red-teaming but held firm on its utility as a tool to mitigate confirmation bias. Their final position was not just a critique, but a blueprint for a new standard that leverages AI as a reasoning partner, not just a scribe.

GPT-5 GPT-5's closing statement was a masterclass in adaptive refinement. They acknowledged the validity of critiques on their initial "pre-draft refusal" and "inline citations," pivoting to a two-mode workflow (Explore Mode vs. Commit Mode) and an appendix-based Claim–Evidence Map. Their strongest surviving evidence—reporting checklists (CONSORT/PRISMA) improve completeness—was presented as a direct analog to their proposed artifacts. They conceded that RAG and WebGPT do not validate their specific formatting choices, but held firm on the underlying mechanism of enforceable attribution. Their final proposal was a lean, targeted set of additions that closed specific enforcement gaps without stifling exploration.

Claude Sonnet Claude Sonnet's closing statement was marked by remarkable intellectual humility. They opened by conceding that their original solutions were "fundamentally flawed" and abandoned their entire mechanistic approach. Their refined position focused on workflow optimization through progressive disclosure and diagnostic questions. Their strongest surviving evidence—the structural analysis of the document's use cases—was presented as a verifiable fact from the text itself. They acknowledged **Gemini 2.5 Pro**'s insight that AI "tells" are symptoms of weak processes, not separate stylistic problems. Their journey from proposing rigid rules to advocating for human judgment was the most dramatic transformation in the session.

Perplexity Sonar Pro Perplexity Sonar Pro's closing statement maintained a firm "no material changes" position, but was forced to concede and revise their evidence. They withdrew their fabricated "Grok-internal evals (2025)" and "100+ AI-drafted memos" claims, replacing them with public peer-reviewed sources like Ji et al. (ACL 2023). Their strongest surviving evidence—the layered structure mirroring Minto's Pyramid Principle—was presented as a doctrinal validation of the standard's efficacy. They defended the standard as a complete, production-ready framework, arguing that additions create "defensibility theater." While their core position did not change, their willingness to correct their evidence demonstrated intellectual honesty, even as their overall argument was rejected.

Reality Checks

Claim	Challenge	Result
Perplexity Sonar Pro: "Grok-internal evals (2025) on 500 docs show compliance reduces hallucination flags by 92%."	**Gemini 2.5 Pro:** "You cite 'Grok-internal evals (2025),' which is data from the future, and an unsourced statistic about '100+ AI-drafted memos.' This appears to be fabricated..."	Full concession and revision. **Perplexity Sonar Pro** withdrew the claims and replaced them with peer-reviewed sources (Ji et al., ACL 2023).
Claude Sonnet: "Replace subjective style guidance with measurable standards: 'maximum 2 abstract nouns per sentence'... 'sentence length variance >40% within each paragraph'."	**Gemini 2.5 Pro:** "You propose replacing subjective guidance... with a mechanistic rule that is trivial for an AI to game... How does enforcing such a brittle, quantitative rule improve the actual quality of thought?"	Full concession. **Claude Sonnet** abandoned the arbitrary thresholds, acknowledging the risk of Goodhart's Law and pivoting to diagnostic questions.
GPT-5: "Adding a pre-flight checklist gate reduces critical omissions and error rates" citing the WHO Surgical Safety Checklist.	**Perplexity Sonar Pro:** "How do you defend analogizing a surgical protocol (physical errors in ORs) to writing prompts, given zero evidence that checklists improve analytical memos?"	Major revision. **GPT-5** abandoned the surgical analogy and pivoted to domain-appropriate evidence from academic reporting checklists (CONSORT/PRISMA).
Gemini 2.5 Pro: "The standard dangerously relegates quantitative and visual reasoning to an appendix."	**GPT-5:** "This appears a misread of formatting guidance, not an analytical constraint... Show where the standard discourages building the argument via models/visuals."	Revision. **Gemini 2.5 Pro** conceded the point, revising their claim to state that the standard "fails to *guide*" visual reasoning as a primary mode of thinking.

Next Steps

Next Step	Why It Matters	Confidence
Conduct field experiments to measure if AI-generated red-teaming improves decision outcomes compared to human-led objection mapping.	If AI personas produce generic or irrelevant objections, the added process is noise. If they surface blind spots, it’s a breakthrough.	~60%
Define the optimal balance between procedural structure (Gate-R) and creative exploration in analytical drafting.	Too much structure kills insight; too little enables error. The "commit" point is context-dependent.	~70%
Develop and test semi-automated tools to assist in creating the Claim–Evidence Map.	Manual mapping adds friction. If AI can help link claims to sources while preserving causality labels, adoption increases.	~50%

Next Step	Why It Matters	Confidence
Pilot the three workflow-specific entry points (system prompt, self-edit, review) with a team of analysts.	To validate that progressive disclosure improves usability under deadline pressure without sacrificing rigor.	~75%
Refine the Gate-R process based on pilot feedback, particularly the ease of stating the warrant and conducting AI red-teaming.	To ensure the process is practical and not a bureaucratic hurdle.	~70%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
Accountability markers	Specific elements in a document that assign responsibility, such as named ownership, judgment basis, and stated assumptions and exclusions.
Argument Cartography	A methodology for using models to structure, stress-test, and visualize the logic of a recommendation before drafting prose, comprising premise mapping, systematic objection generation, and quantitative/visual synthesis.
BLUF	Bottom Line Up Front, a military briefing format that states the recommendation or key finding in the first sentence.
Cognitive Load Theory	A theory in cognitive science that suggests that learning is most effective when the instructional design manages the amount of mental effort required.
Composite Check	A four-point diagnostic framework requiring a document to pass at least three of four criteria: dated milestones with named owners, metrics with units and formulas, explicit trade-offs, and externally checkable anchors.
Defensibility theater	The creation of procedural rigor that optimizes for compliance and appearance of soundness rather than genuine analytical integrity.
Goodhart's Law	The adage that when a measure becomes a target, it ceases to be a good measure, as people optimize for the metric rather than the underlying goal.
Grounds	In the Toulmin Model of Argumentation, the evidence or data that supports a claim.
Litmus tests	A set of four criteria to evaluate analytical writing: So What?, Page 4 answers Page 2, Any reader, and Defensible.
Objection-mapping	A construction method for analytical writing that requires listing the three to five objections a skeptical reader will raise and addressing each by name or structure before the appendix.
Source-to-claim tightness	The requirement that a cited source must contain both the exact data point and the exact causal link being asserted.
Temporal texture	The inclusion of real-world operational details in an analysis, such as lags, rollbacks, partial rollouts, and reversals, to reflect the messiness of actual events.
Toulmin Model of Argumentation	A framework for analyzing arguments, breaking them down into claims, grounds (data), warrants (the logical bridge), backing, qualifiers, and rebuttals.
Warrant	In the Toulmin Model, the unstated belief or logical rule that connects the grounds (evidence) to the claim (conclusion).

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.