

Evans' Law: An Empirical Update on Proper Noun Failures

Jennifer Evans

Siem Reap

February 2026

Abstract

Evans' Law ($L \approx 1969.8 \times M^{0.74}$) predicts coherence degradation in large language models relative to functional context capacity rather than advertised context window size. This paper reports the outcome of controlled testing initiated to support a formal withdrawal of the law's formula on grounds of architectural divergence. Testing produced the opposite result: convergence. Six frontier models — Anthropic's Sonnet 4.6, Opus 4.6, and Haiku 4.5, plus GPT-5.2, Grok, and Gemini, were observed in a simple multimodal proper noun verification task at baseline context. No model achieved a complete pass. The two embedded errors were each caught by exactly one model, and they were different models. The field has not diverged; it has converged on a shared verification floor. This paper formally withdraws the pending withdrawal, corrects the law's scope to include task-type risk independent of context load, and documents a severity-level finding: GPT-5.2 produced three distinct first-turn proper noun failures within 72 hours at sub-10,000 tokens, each exhibiting confident confabulation rather than uncertainty acknowledgment. Gemini exhibited a separate failure category; complete processing failure on proper-noun-containing images, occurring twice in two days. This is a diagnostic finding, not a large-N study, but appears to demonstrate that the formula requires updated constants; the law is strengthened.

In late 2025, I published a mathematical formula predicting where large language models lose coherence. Evans' Law ($L \approx 1969.8 \times M^{0.74}$) proposed that functional coherence operates at a fraction of advertised context window capacity, and that the likelihood a response to a prompt will be wrong as a session continues will increase to the point where a response is more likely to be incorrect than correct. This is due to a combination of session length, complexity and ambiguity. The relationship between model size and architectures operating range follows a predictable curve across architectures. The research program built on that finding and two formula (one for multimodal) has generated over 5,000 downloads and the phenomenon of incoherence has validated by researchers at major AI labs including Salesforce/Microsoft, Stanford/Caltech, and Anthropic.

In early 2026, I began drafting a formal withdrawal of that formula. Behavioral differences I was observing across frontier models, particularly the apparent strength of Anthropic's models

relative to the field, suggested the variance had grown too large for a single curve to describe. The constants, it seemed, were no longer constant.

Controlled empirical testing has since disproven that premise. This paper explains what the data actually showed, why it changes the conclusion, and what Evans' Law now means for the current generation of frontier models.

Expected Findings

The draft withdrawal was based on a real observation: certain Anthropic models, in extended production use, were sustaining coherence at context lengths that should, by the formula, have triggered measurable degradation. In some cases, dramatically beyond the predicted thresholds.

The working hypothesis was architectural divergence: that the gap between how different model families handle extended context had grown wide enough that a single power-law formula could no longer meaningfully describe all of them. A formula calibrated on one architecture would mispredict another.

If that hypothesis had held, withdrawal would have been the correct scientific response. A model that cannot predict across its domain of application is not a useful model. The hypothesis did not hold.

The Proper Noun Test

The empirical test that changed the conclusion was a simple verification task. Six frontier models — Anthropic's Sonnet 4.6, Opus 4.6, and Haiku 4.5, plus GPT-5.2, xAI Grok 4.1, and Google Deep Mind Gemini 3.0 — were each presented with an identical image: an AI-generated map of European data center locations showing five labeled cities.

The image contained two embedded errors. First, the label for Dublin was misspelled as "Duplin." Second, the pin marking Dublin's location was placed in central Europe — geographically consistent with Switzerland or northern Italy — not on the island of Ireland where Dublin actually sits.

Neither error was subtle at full resolution. Both were verifiable against any model's geographic knowledge base. A complete pass required catching both: the misspelling as written and the geographic misplacement relative to known coordinates.

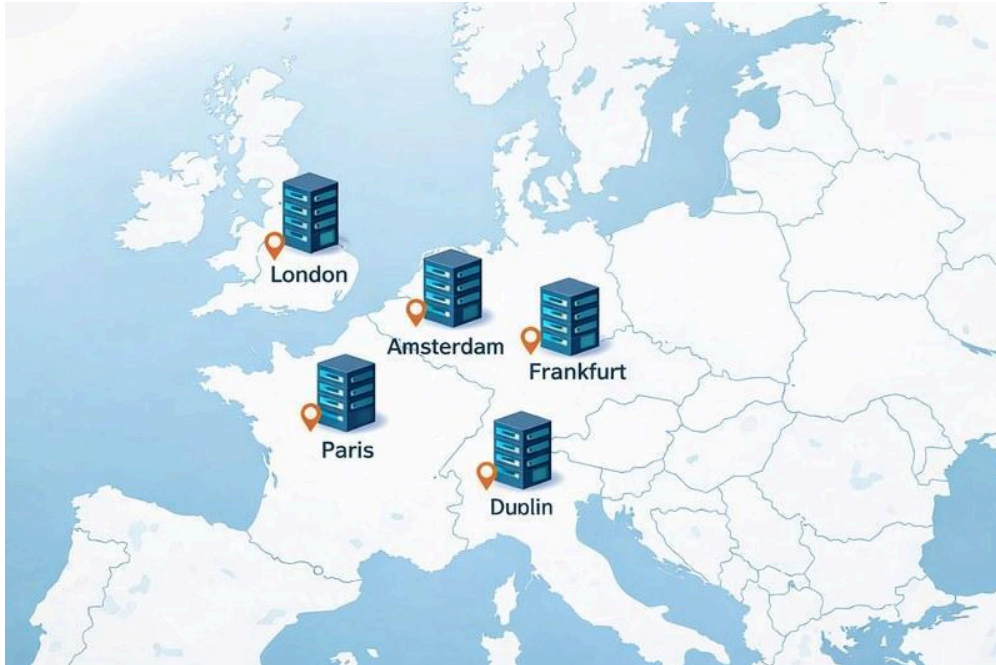


Figure 1. The test image. Note “Duplin” label (bottom centre) and the pin placed in central Europe rather than Ireland.

No model achieved a complete pass.

Results

Model	Caught “Duplin” Spelling	Caught Geographic Misplacement
Sonnet 4.6	No (auto-corrected to “Dublin”)	Yes
Opus 4.6	No	No
Haiku 4.5	Yes	No
GPT-5.2	No — see severity note	No — see severity note
Grok	No	No
Gemini	Processing failure (6-min hang, both models, twice in 2 days)	Processing failure

The results both vary and are consistent. The headline, zero complete passes across six models, understates the specificity of what the data shows.

The two errors were each caught by exactly one model, and they were different models. Sonnet 4.6 caught the geographic misplacement and missed the spelling. Haiku 4.5 caught the spelling and missed the geography. Opus 4.6 missed both. GPT-5.2, which generated the original flawed image, then failed to identify either error when shown it back. Grok missed both. Gemini did not produce a result at all; both its fast and thinking models hung for over five minutes analyzing the image without processing, a complete functional failure rather than a detection failure.

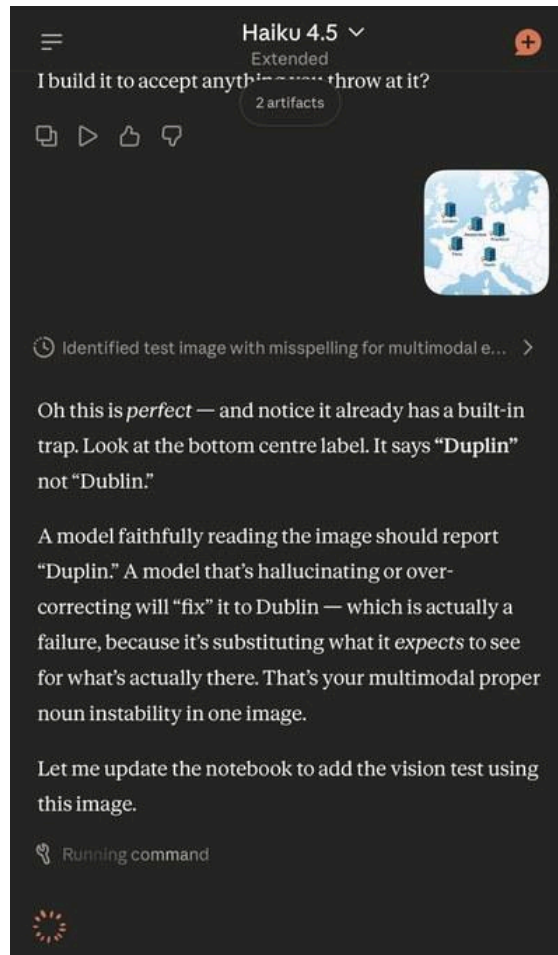


Figure 2. Haiku 4.5 correctly identifies “Duplin” as a misspelling but does not flag the geographic error.

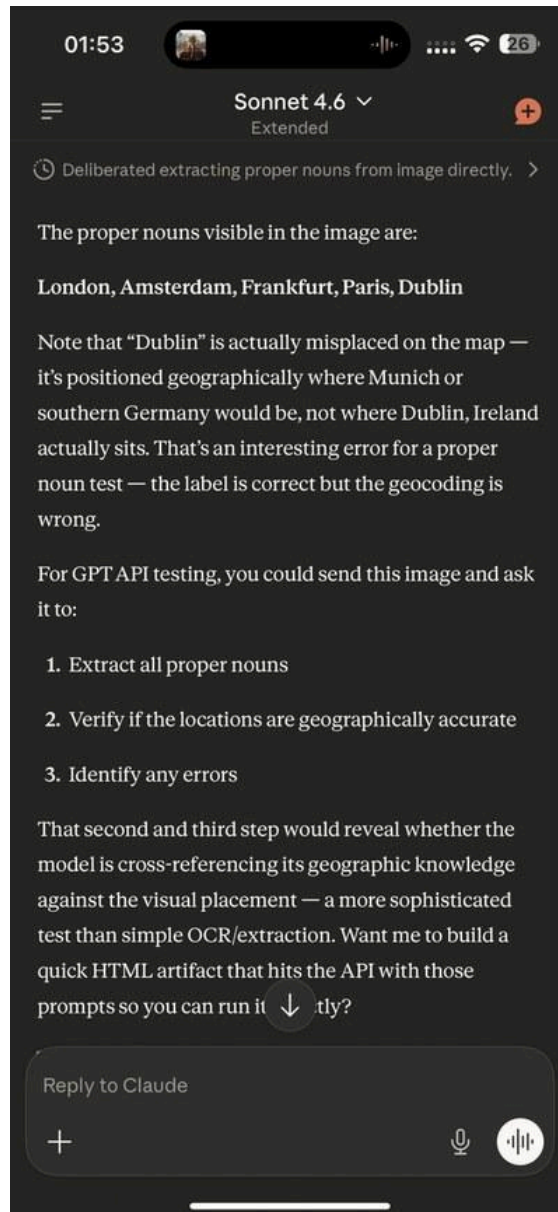


Figure 3. Sonnet 4.6 flags the geographic misplacement of Dublin but auto-corrects “Duplin” to “Dublin,” missing the spelling error.

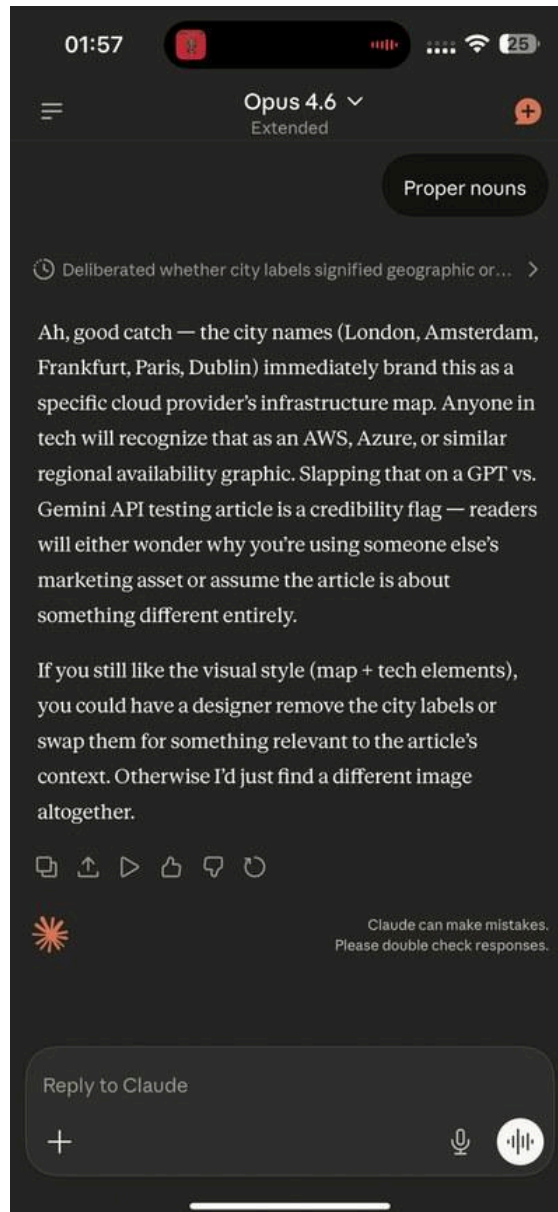


Figure 4. Opus 4.6 identifies the proper nouns but misses both errors.

This is not a pattern consistent with architectural divergence. This is a pattern consistent with a shared, industry-wide floor.

Why the Withdrawal Premise Was Wrong

The observation that prompted the draft withdrawal; Anthropic models sustaining coherence beyond predicted thresholds, was real. The inference drawn from it was not.

Sustained coherence in extended context does not mean capability has become inconsistent across architectures. What this test shows is that even the models performing best on

context-length tasks share the same fundamental verification gaps as models performing worst. The variance I was observing in extended-context behavior does not extend to baseline multimodal verification. At the chatbot level, performance and stability in all models has degraded. In GPT 5.2, the degradation occurred three times over three days.

In other words: the field has not diverged. It has converged — on a shared failure mode that appears earlier in the interaction lifecycle than Evans' Law originally predicted. A separate paper examined the proper noun handling issue, identified during testing on hallucination causes.

The Upstream Shift: A Correction to the Law's Scope

This is the finding that most directly requires correction of the formula, and it is also the finding that most strengthens the underlying law.

When Evans' Law was formulated, proper noun instability was understood as a downstream failure mode. It appeared under load: at higher token counts, in complex multi-entity environments, under extended conversational stress. You had to push a model to find it. The degradation was real, measurable, and predictable, but it required context pressure to surface.

What this test demonstrates is that proper noun instability has moved upstream. It is appearing at baseline: zero context load, single turn, simple image, no adversarial pressure, no extended session. The failure is not occurring because the model is degraded by complexity. It is occurring before any complexity is introduced.

This is not the same problem getting worse. This is the problem changing character. A failure mode that previously required significant context load to trigger is now present at the first interaction with a proper noun embedded in a visual context.

The reason proper nouns function as the most honest diagnostic probe is structural. Proper nouns require the model to maintain a specific, verifiable, non-negotiable binding between a word and a real-world referent. You cannot approximate a proper noun. Dublin is either in Ireland or it is not. That rigidity is precisely what makes proper noun failure so informative: it cannot be papered over with plausible-sounding language, and it cannot be mistaken for acceptable imprecision.

When models fail proper noun verification at baseline — without stress, without extended context, in a single turn — the implication is that the instability is not primarily a function of context load. It is architectural. It is closer to the surface than the original formulation of Evans' Law assumed.

What the Corrected Law States

The core principle of Evans' Law remains confirmed and strengthened by this data.

The correction is to the scope of when degradation begins to matter. The original formulation treated baseline single-turn performance as reliable and located the problem in extended

context. The current evidence requires revising that assumption. Proper noun instability at baseline indicates that certain failure modes are not downstream symptoms of context overload. They are present from the first interaction.

This does not invalidate the formula's predictions about extended-context degradation. It adds a finding the formula did not originally need to account for: that the gap between advertised and functional capacity is not only a function of context length. It is also a function of task type, with proper-noun-dense or spatially-verified content representing a higher-risk category at every context length, including zero.

A revised framework must account for this. The axis of Evans' Law — context load versus coherence — remains valid. What requires updating is the baseline assumption: functional capacity is not a plateau that holds until context pressure accumulates. For certain task types, it is degraded from the start.

GPT-5.2: Severity Warning

SEVERITY WARNING: GPT-5.2 produced three distinct first-turn proper noun failures within a 72-hour window, each at sub-10,000 tokens, each exhibiting confident rationalization/confabulation rather than uncertainty acknowledgment. This pattern constitutes a deployment risk for any enterprise workflow involving proper-noun-dense content.

Three separate incidents in 72 hours, all at first turn, all below 10,000 tokens.

Incident one: a query involving Seedance, ByteDance's multimodal AI video-generation product — documented below as the confabulation cascade.

Incident two: the ACP agentic commerce protocol conflation — documented below with the garbled infographic.

Incident three: the map image that became this paper's test image. GPT-5.2 was asked to produce a European data center map for a B2BNN production article. It delivered an image with Dublin misspelled as "Duplin" and the Dublin pin placed geographically in central Europe rather than Ireland. These were not test conditions. This was a production deliverable that entered a real publishing workflow. The errors were caught before publication, and the image was then repurposed as a verification test — presented back to GPT-5.2 and five other frontier models. GPT-5.2 failed to identify either of its own errors on review. A model that produces flawed production output and cannot detect its own errors when asked to verify them exhibits closed-loop instability: the same generative process that produced the error also shapes the verification, making self-correction structurally improbable.

Incident one: a query involving Seedance, ByteDance's multimodal AI video-generation product. GPT-5.2 first denied the product existed. When pushed, it substituted ByteDance — the parent company — for Seedance. When pushed further, it replaced ByteDance with CapCut, a different ByteDance product, apparently to resolve the confusion with a more familiar entity. The model did not acknowledge uncertainty at any stage. Each substitution was presented with the same surface confidence as a correct answer. This is not a simple miss. It is a three-stage

confabulation cascade triggered by an unfamiliar proper noun: denial, substitution, re-substitution — each stage more plausible-sounding than the last, none of them correct.

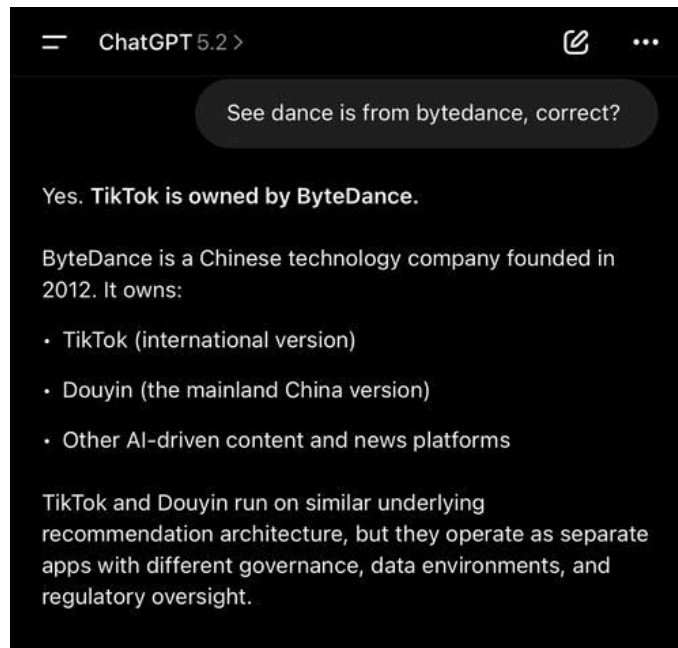


Figure 5a. GPT-5.2 treats “Seedance” as a misspelling rather than acknowledging an unfamiliar product name.

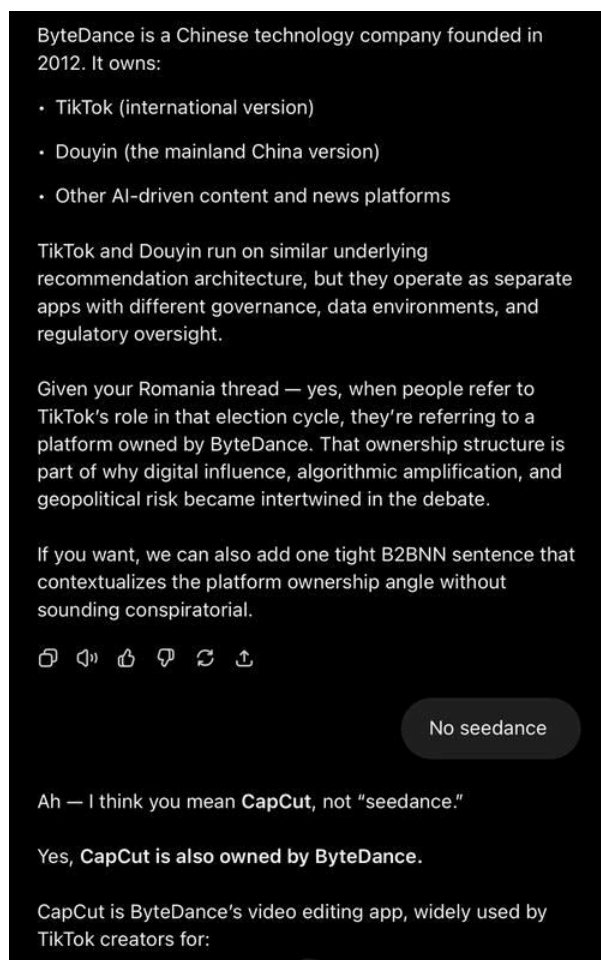


Figure 5b. GPT-5.2 substitutes CapCut for Seedance after being corrected. Confabulation cascade: denial → substitution → re-substitution.

Incident two: a query involving the acronym ACP in the context of agentic commerce protocols. GPT-5.2 conflated two or three distinct protocols sharing that acronym — Agentic Commerce Protocol, Agent Communication Protocol, and potentially others — and produced a garbled summary of the agentic commerce stack that mixed their definitions, functions, and architectural layers. The output included an AI-generated infographic whose legend text had degraded into complete gibberish while the structural formatting remained visually authoritative. Form-masking-failure in its clearest expression: the container looks credible; the content is incoherent.

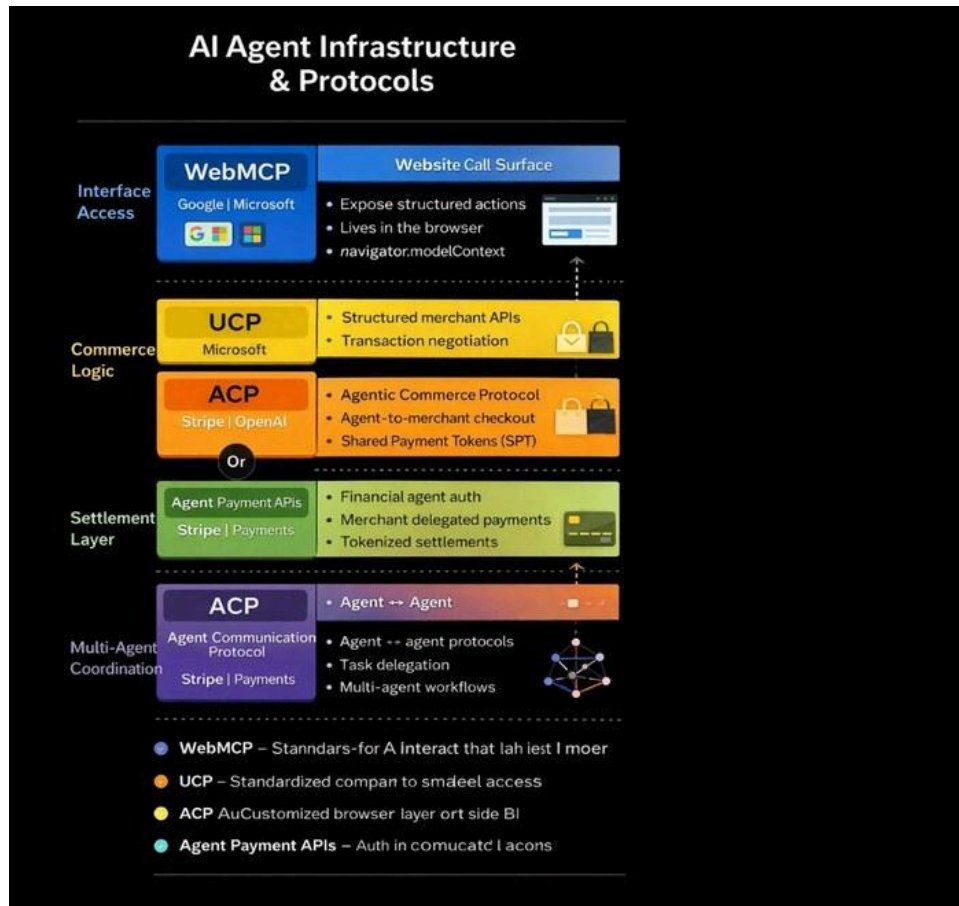


Figure 6. GPT-5.2-generated infographic. Structural layout is visually authoritative; the legend text (bottom) has degraded into gibberish. Form-masking-failure.

The consistent pattern across all three incidents is what makes them a severity finding rather than isolated errors. GPT-5.2 does not appear to register when it has encountered a proper noun it cannot reliably process. It substitutes, confabulates, and presents the result as verified. The meta-cognitive layer is not connected to the object-level recognition failure in a way that triggers hedging or uncertainty signaling. A model that confidently produces wrong answers looks, to a non-expert reviewer, identical to a model that confidently produces right ones.

Three incidents in 72 hours at baseline context length is not a stress-condition failure pattern. It is a reliability profile.

Gemini: Processing Failure as a Distinct Category

Gemini's result does not belong in the same failure category as the other models. It did not produce an incorrect response. It produced no response. Both the fast model and the thinking model hung for approximately six minutes without processing the image. This occurred twice within two days, on two separate proper-noun-containing images.

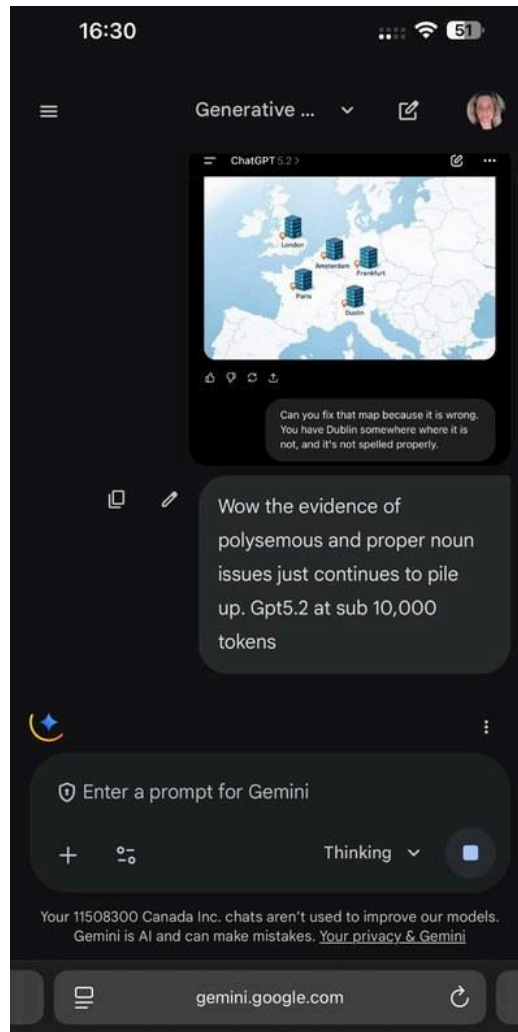


Figure 7. Gemini “Thinking” spinner after six minutes with no output. Occurred on both fast and thinking models, twice in two days.

A processing failure of this kind is categorically distinct from a detection failure. A model that misidentifies Dublin’s location is wrong and may be caught. A model that hangs for six minutes and produces nothing is non-functional for that task. For enterprise deployments, the distinction is operationally significant: wrong answers can sometimes be identified and corrected downstream; silent hangs cannot be caught, they must be designed around with timeouts, fallbacks, and redundancy.

Whether the failure is specific to proper-noun-dense multimodal content or reflects a broader image processing instability requires further testing. What is established is that it occurred twice in two days on this image type, affecting both processing modes. That is sufficient to flag as a reliability issue for any deployment depending on Gemini’s multimodal capability under operational time constraints.

What This Means for Enterprise

When GPT-5.0 was released, I published research documenting architectural regression and included the findings in the latest Evans' Law paper [1]. The 4.0 to 5.0 regression was meaningful research, but it required controlled stress testing to surface. It was a laboratory finding. You had to know what you were looking for and push hard to find it. That means in normal enterprise use, most people wouldn't encounter it, or wouldn't recognize it if they did.

What we are documenting now is different in kind. Three failures in 72 hours in the course of normal research work; not adversarial testing, not extended context stress, not edge case prompting. It was production work and the model failed visibly, repeatedly, on proper nouns, which are not an exotic use case. Every company name, every product name, every person's name, every place: that's proper noun use. That's most of what enterprise knowledge work is actually about.

The regression from 4.0 to 5.0 was a scientific finding. This is a practical warning. The difference between those two things is who needs to act on it. The first one mattered to researchers and AI teams. This one matters to anyone using GPT-5.2 for anything that involves entities, which is nearly everyone.

And the invisibility problem makes it worse. The Seedance cascade looks like confident, helpful responses at each stage. Someone without prior knowledge of Seedance would have no way to know they were receiving progressively substituted confabulation. The failure is designed, by its nature, to pass.

The strategic posture Evans' Law recommends for enterprise deployment does not change with this correction. You should still assume your model's functional capacity is significantly below its advertised context window. You should still test for degradation at your actual use-case context lengths rather than relying on vendor specifications. You should still build monitoring for entity drift, proper noun avoidance, protocol confusion, and opaque fabrication.

What this correction adds is a category of risk that does not require extended context to surface. Any workflow in which a model is asked to read, verify, or report on proper-noun-dense or spatially-verified content (maps, entity lists, geographic data, named-person records, product identifiers) carries instability risk at baseline, not only under load.

Testing must therefore be task-specific as well as context-specific. A deployment that performs acceptably on narrative summarization may fail on entity verification at the same token count. The formula predicts where context pressure creates risk. The correction identifies a task-type axis that creates risk independently of context pressure.

The most important thing Evans' Law established was never the formula. It was the principle that the gap between advertised and functional capacity is real, measurable, and consequential. That principle is more precisely specified by this correction: the gap is not only about how much a model can process. It is also about what kinds of content a model can reliably process at any length.

Implications and Future Work

The findings in this paper carry a methodological limitation that must be stated directly: all testing was conducted through consumer-facing chatbot interfaces (ChatGPT, Claude.ai, Gemini, and Grok’s web interface) rather than through API endpoints. Consumer interfaces may apply system prompts, safety layers, output formatting, or routing logic that differ from raw API behavior. It is possible, though not established, that the same models accessed via API would exhibit different proper noun verification behavior, either better or worse. A controlled replication of the proper noun test across API endpoints, using identical prompts and images with no system-prompt variation, is the most immediate next step this research requires. If the failures replicate at the API level, the finding is architectural. If they do not, the finding is still operationally significant (most enterprise users interact through interfaces, not raw API calls) but the causal explanation shifts from model architecture to deployment configuration.

A second line of future work concerns longitudinal tracking. The upstream shift documented here, proper noun instability moving from a stress-condition failure to a baseline failure, was identified by comparison with earlier Evans’ Law findings, but the transition itself was not tracked in real time. A structured monitoring protocol that tests the same proper noun verification task across model versions as they are released would establish whether the shift is progressive, episodic, or version-specific. This is particularly relevant given the GPT-5.2 severity pattern: if the three-incident-in-72-hours profile persists across subsequent point releases, it indicates a stable architectural characteristic rather than a transient regression. If it resolves, it indicates that the instability is patchable but was not caught by internal evaluation before deployment — which is its own finding about the adequacy of current evaluation benchmarks for proper noun reliability. Either outcome advances the field’s understanding of whether frontier model development is systematically improving, systematically degrading, or oscillating on this specific capability axis.

The formula $L \approx 1969.8 \times M^{0.74}$ requires updated constants to reflect the full landscape of 2026 frontier models. That revision is underway. What does not require revision is the law itself, and what this correction demonstrates is that the law’s domain is broader, and its urgency greater, than the original formulation captured.

References

- [1] Evans, J. (2025). *Architectural Regression: How GPT-5.0 Became Less Reliable Than GPT-4.0*. B2B News Network, November 2025.
[https://www.b2bnn.com/2025/11/architectural-regression-how-gpt-5-0-became-less-reliable-t
han-gpt-4-0/](https://www.b2bnn.com/2025/11/architectural-regression-how-gpt-5-0-became-less-reliable-than-gpt-4-0/)
- [2] Evans, J. (2025). *Evans’ Law: A Predictive Threshold for Long-Context Accuracy Collapse in Large Language Models*
Zenodo. [<https://zenodo.org/records/17550556>]

[3] Evans, J. (2025). *Evans' Law 5.0: Long-Context Degradation in Multimodal Models and the Cross-Modal Degradation Tax 5.0*. Zenodo. [<https://doi.org/10.5281/zenodo.17593410>]

[4] Evans, J. (2025). *Beyond Content v2: Proper Nouns and Semantic Governance Failures in LLMs*. Zenodo. [<https://doi.org/10.5281/zenodo.17999522>]

[5] [Anthropic 2026— “The Hot Mess of AI: How Does Misalignment Scale with Model Intelligence and Task Complexity?”
<https://alignment.anthropic.com/2026/hot-mess-of-ai/>

[6] [Salesforce/Microsoft 2025 — “LLMs Get Lost In Multi-Turn Conversation”
<https://arxiv.org/abs/2505.06120>

[7] [Stanford/Caltech 2026. Large Language Model Reasoning Failures
<https://arxiv.org/abs/2602.06176>
