Hallucinations by Design: A Cross-Model Assessment of How Major AI Consulting Firms Advise Clients to Manage Generative AI Errors

Methodological Overview

This concluding report was developed through a multi-stage comparative analysis using four independent large language models - GPT, Perplexity, Copilot, and Gemini - each provided with an identical, tightly scoped prompt focused exclusively on how major organizations conceptualize and control AI hallucinations. Each model produced a standalone assessment of Deloitte, PwC, KPMG, and Microsoft based solely on publicly available product documentation, governance frameworks, and architectural descriptions. These four reports were then treated as independent analytical inputs and reviewed side-by-side to identify areas of full agreement, partial alignment, and material divergence. The final conclusions presented here do not introduce new interpretation or synthesis beyond what is supported across the four model outputs; rather, they reflect a cross-model consensus assessment that weights findings according to consistency and recurrence across all analyses.

Model-Specific Research Approaches (Summary)

Although all four models responded to the same prompt and scope constraints, each employed a distinct research and extraction approach. These differences are summarized below to provide transparency into how the underlying assessments were produced.

GPT

GPT conducted a structured document-based analysis grounded in publicly available firm materials, including AI governance frameworks, responsible AI documentation, product descriptions, and architectural explanations. The analysis focused on extracting explicit and implicit assumptions about hallucinations, mapping them to lifecycle control points (before, during, or after generation), and classifying the implied architectural paradigm. GPT did not rely on live web search in the final synthesis stage; instead, it treated cited materials within the provided reports as authoritative and constrained conclusions strictly to what those materials supported.

Copilot

Copilot applied a cross-report synthesis methodology rather than introducing independent source expansion. It treated the four model-generated firm assessments as fixed analytical inputs and performed normalization across them using a shared taxonomy (hallucination framing, inevitability vs. preventability, lifecycle control point, and architectural paradigm). Copilot emphasized evidence-weighted consensus scoring, explicitly distinguishing strong, moderate, and mixed agreement, and avoided resolving disagreements through interpretation.

No new facts or architectural inferences were introduced beyond what appeared in the reports themselves.

Perplexity

Perplexity combined targeted web research with document-based extraction. It issued focused queries against firm websites, white papers, technical blogs, and (for Microsoft) platform documentation related specifically to hallucinations, grounding, RAG, and responsible AI controls. It then cross-referenced these findings with the provided compiled report to ensure alignment. In the consensus phase, Perplexity treated the compiled document as ground truth and restricted synthesis to points explicitly present across the four model analyses, flagging consensus levels based on recurrence rather than emphasis.

Gemini

Gemini employed a forensic, hypothesis-testing approach aimed at identifying underlying architectural assumptions rather than surface-level governance claims. It analyzed firm frameworks, platform documentation, and (for Microsoft) technical tooling descriptions to determine when intervention occurs relative to model execution. Gemini explicitly tested whether any organization demonstrated pre-generation execution denial or hard inference gating and reported negative findings where no such evidence appeared. In the cross-model stage, Gemini emphasized alignment across model interpretations to ensure that conclusions reflected collective understanding rather than a single-model reading.

Final Summaries of Each AI Model's Research

GPT - Summary of Assessment Across the Four Firms

Overall stance
GPT consistently concludes that all four firms operate under a **generation-first, post-hoc mitigation paradigm**. Hallucinations are treated as inherent, probabilistic properties of LLMs and managed through governance, RAG, and human oversight after inference.

By firm

- **Deloitte**: Hallucinations framed as reliability and model-risk failures; inevitable and governed after generation via QA, governance, and review.
- **PwC**: Hallucinations described as plausible but baseless outputs; mitigated through RAG and prompt discipline during/after generation, not structurally prevented.
- **KPMG**: Hallucinations treated as endemic reliability risks; addressed primarily through post-generation governance, auditability, and monitoring.
- **Microsoft**: Hallucinations framed as “ungrounded content”; mitigated through grounding, correction loops, and governance after inference.

Key takeaway
GPT finds **no evidence** of pre-inference execution denial or structural prevention across any firm.

Gemini - Summary of Assessment Across the Four Firms

Overall stance
Gemini aligns closely with GPT, emphasizing hallucinations as **unavoidable artifacts of probabilistic models**, managed through layered controls after generation.

By firm

- **Deloitte**: Hallucinations = output reliability risk; inevitable; controlled post-generation through governance and human review.
- **PwC**: Hallucinations framed as probability–reality mismatch; addressed via "membranes," RAG, and review, but generation always occurs.
- **KPMG**: Hallucinations treated as persistent risk; governed using traditional risk frameworks after generation.
- **Microsoft**: Hallucinations formalized as ungroundedness; reduced via grounding and correction, not eliminated.

Key takeaway
Gemini explicitly states that **none of the four firms challenge the generation-first assumption**.

Copilot - Summary of Assessment Across the Four Firms

Overall stance
Copilot characterizes all four organizations as assuming hallucinations are **expected and unavoidable**, with mitigation focused during and especially after generation.

By firm

- **Deloitte**: Hallucinations are business and governance risks; managed after generation via validation and oversight.
- **PwC**: Hallucinations are inherent GenAI risks; mitigated through prompts, RAG, and Responsible AI governance.
- **KPMG**: Hallucinations embedded in reliability and trust concerns; handled through post-generation controls and audits.
- **Microsoft**: Hallucinations treated as groundedness failures; mitigated via platform tooling (RAG, correction, safety layers).

Key takeaway
Copilot finds **no advocacy of pre-generation blocking**; refusal is treated as a configured behavior, not a default architecture.

Perplexity - Summary of Assessment Across the Four Firms

Overall stance
Perplexity provides the most granular lifecycle detail but reaches the same conclusion: hallucinations are **structurally inevitable** and mitigated through compensatory controls layered onto generation-first systems.

By firm

- **Deloitte**: Hallucinations are statistically expected reliability failures; mitigated via RAG, QA, and governance after generation.
- **PwC**: Hallucinations are inherent GenAI risks; reduced through prompting, RAG "double checks," and review.
- **KPMG**: Hallucinations framed as endemic reliability risks; managed via governance, validation, and monitoring post-generation.
- **Microsoft**: Hallucinations defined as ungrounded outputs; addressed through grounding, detection, and correction after initial inference.

Key takeaway
Perplexity explicitly notes that RAG is treated as a compensatory mechanism, not a hard gate preventing inference.

Cross-Model Meta-Observation (Derived from All Four Summaries)

All four AI models independently agree that:

- **Hallucinations are treated as inevitable by all four firms.**
- **Inference proceeds by default.**
- **Mitigation occurs during and especially after generation.**
- **No firm is described as enforcing structural, pre-inference execution denial.**