Quantified Hallucination Rates in Contemporary AI Systems (2023–2025)

Below is a table restricted to 2023–2025, using only peer-reviewed papers, official benchmarks, or widely cited technical reports and leaderboards that (a) define hallucination, (b) provide an explicit rate or percentage, (c) specify the task or benchmark evaluated, and (d) identify model versions.

Hallucination rates vary substantially depending on task design, definition, and evaluation methodology. Across citation tasks, open-domain QA, clinical reasoning, RAG systems, and knowledge benchmarks, contemporary research consistently reports non-zero hallucination rates — ranging from low single digits in tightly constrained, retrieval-grounded systems to 50–80% in adversarial or high-uncertainty contexts.

Study (Author, Year)	Model(s)	Hallucination Definition (max 12 words)	Reported Hallucination Rate (%)
Chelli et al., 2024 [jmir]	GPT-3.5, GPT-4 (ChatGPT), Bard/Gemini	Generated reference not matching any real paper metadata [jmir]	GPT-3.5: 39.6%; GPT-4: 28.6%; Bard: 91.4% [jmir]
Multi-model clinical vignettes, 2025 [pmc.ncbi.nlm.nih]	GPT-4o, Distilled-DeepSeek-Llama, Phi-4, Gemma-2-27B-it, Qwen-2.5-72B	Elaborates on deliberately fabricated clinical detail as if true [pmc.ncbi.nlm.nih]	Overall 50–82.7%; GPT-4o ~50–53.3%; DeepSeek-Llama 80–82.7% [pmc.ncbi.nlm.nih]
Yang et al., 2025 (MetaQA) [arxiv]	GPT-4, GPT-3.5, Llama-3, Mistral	Answer factually incorrect when model chooses to respond [arxiv]	GPT-4: 17–28%; others up to 55% (by dataset) [arxiv]
Magesh et al., 2025 [dho.stanford]	GPT-4, Lexis+ AI, Ask Practical Law	Output containing incorrect or unsupported legal statements vs. gold [dho.stanford]	General-purpose GPT-4 ≈40%; legal RAG tools substantially lower (varied) [dho.stanford]
Nishisako et al., 2025 [cancer.jmir]	GPT-4, GPT-3.5 (RAG vs. conventional chatbots)	Response contradicts or not supported by cancer information sources [cancer.jmir]	CIS-RAG GPT-4: 0%; CIS-RAG GPT-3.5: 6%; Google-RAG GPT-4: 6%; Google-RAG GPT-3.5: 10%; conventional ≈40% [cancer.jmir]

Study (Author, Year)	Model(s)	Hallucination Definition (max 12 words)	Reported Hallucination Rate (%)
Kelkar et al., 2024 [app.got-it]	Mixtral-8x7B, GPT-3.5-Turbo-1106, GPT-4-Turbo-1106	Answer not grounded in retrieved docs or factually false [app.got-it]	Mixtral: 8.3%; Mixtral+filter: 3.4%; GPT-3.5: 8.7%; GPT-4-Turbo: 2.6% [app.got-it]
Vectara, 2025 [vectara]	DeepSeek-R1, DeepSeek-V3, GPT-o1, GPT-4o	Unsupported or contradicted answer under FACTS hallucination metric [vectara]	DeepSeek-R1: 14.3%; DeepSeek-V3: 3.5%; GPT-o1/4o: ~0.9–1.9% [vectara]
AllAboutAI, 2025 [allaboutai]	GPT-4, Claude, Gemini	Factually incorrect answer vs. trusted references on mixed tasks [allaboutai]	Real-world tasks: GPT-4 ~21%; Claude ~13%; Gemini ~19%; structured: GPT-4 ~1.5–3.7% [allaboutai]