AI Failure Modes in Regulatory Writing: A Risk Awareness Framework

Sruthi Sampathkumar, MBBS, MS · Founder and Principal, AccelWrite LLC

AccelWrite
AMWA Medical Writing & Communication Conference 2026
Poster Session

Background and gap

AI failure modes are well documented in clinical artificial intelligence (AI), medical natural language processing (NLP), and radiology, but, to our knowledge, no published framework translates them to regulatory writing, where trial design integrity, content traceability, and sponsor accountability carry real-world consequences.

~4 → 56.9 56.9 fabricated references **per 10,000** biomedical papers by early 2026, up from approximately 4 in 2023[1]

Objective

To present a taxonomy of AI failure modes with consequences for regulatory document development, grounded in published evidence where available and mapped to submission risk and to the writer competency required to recognize and mitigate each.

Take home

Medical writers as informed risk managers, not passive AI users.

Method

Targeted review of AI, medical NLP, and clinical decision support literature (2023–2026), with framework development informed by regulatory writing practice and subject-matter expertise in AI product development. Literature was identified through PubMed and Google Scholar searches and AI-assisted searching (Claude, Anthropic). Every reference was verified by the author against the source records.

The evidence comes from related fields: clinical note summarization, diagnosis, pathology, systematic reviews, and general AI testing. No study has yet measured these failure modes in regulatory documents. The consequences listed are the author's assessment of likely impact, not documented incidents, and changes over time are reported only where studies measured the same thing in the same way. Some failure modes overlap; they are listed separately when each needs its own check. For example, reference fabrication is a form of hallucination, but it is caught by checking PubMed and digital object identifiers (DOIs), not the document's own sources.

Scope: This conceptual framework covers errors in what AI produces and in how writers and AI tools interact. Governance risks (confidentiality and data protection, unapproved use, tool validation) are out of scope.

Failure mode	Consequence in regulatory writing	Writer competency	Key evidence	Change over time
■ **Content integrity** · *what the document says is wrong, missing, or overstated*				
1. Hallucination	The document states something the source doesn't support, or contradicts it	Sources quality control (QC) discipline Escalating to SMEs when verifying a claim requires specialist expertise	1.47% hallucination rate across 12,999 clinician-annotated sentences of large language model (LLM)-generated clinical notes[2]	Low in a purpose-built, clinician-reviewed system; not eliminated
2. Numerical and source-fidelity errors	Numbers or their attributions do not match the source data	Precise extraction instructions; 100% verification of numbers; recomputing derived values	In systematic-review data extraction, only 35% of variables were acceptable on every metric; numerical effect-size variables were least accurate[3]	No early baseline
3. Reference fabrication and misattribution	Citations point to studies that do not exist, or do not say what is claimed	Two-step references QC check: (a) the citation exists (PubMed, DOI), and (b) its full text, beyond the abstract, supports the statement	2023: 47% of references fabricated (GPT-3.5)[4]; 55% with GPT-3.5 vs 18% with GPT-4[5]. 2026 preprint: 35.3% fabricated and only 31.8% fully correct among 2026 models[6]	Lower in newer models; not eliminated
4. Omission	Important source findings go missing, so the data look better or worse than they are	Output completeness check against sources, template, and regulatory requirements	3.45% omission rate in LLM-generated clinical notes, more than twice the hallucination rate[2]	No early baseline
5. Overgeneralization	Limited or uncertain findings become broad, definite claims	Hedge-and-scope Sources QC check: compare qualifiers, populations, and certainty wording against source	LLM summaries are nearly 5 times as likely as human-written ones to overgeneralize (OR, 4.85; 95% CI, 3.06-7.70); newer models less accurate[7]	Newer models worse in one comparison
■ **Context and coherence** · *long documents lose detail, and wording drifts between sections*				
6. Context collapse	Information buried in the middle of long documents gets missed or misused	Giving AI one section at a time; querying in the source's own terms; coverage QC of long documents, especially middle sections	Retrieval degrades for information in the middle of long inputs[8]. 2025: at 32K tokens, 11 of 13 models fell below 50% of their short-context baseline[9]	Persistent in 2025 models on retrieval tasks
7. Tone and terminology drift	Wording shifts across sections: promotional phrasing creeps in, or one concept gets several names	Terminology and style QC against the style guide, lexicons, and controlled terminology (MedDRA, CDISC); neutral-tone read QC with a flag-word list for promotional (eg, robust, favorable) and AI-style wording (eg, delves, underscores)	At least 13.5% of 2024 PubMed abstracts show LLM style vocabulary[10]. Terminology drift: *not yet measured*	Not measured
■ **Reproducibility and stability** · *the same task can give different results*				
8. Nondeterministic output	The same prompt gives a different answer each time, so a reviewed draft can't be regenerated	Save the generated output (not the prompt) as the record; QC every output, not just the first	Temperature 0 did not guarantee determinism[11]; accuracy varied up to 15% across identical runs[12]; still present in 2026[13]	Persistent
9. Prompt sensitivity	Small wording changes in a prompt change the output substantially	Shared, version-controlled prompt templates; QC every output regardless of prompt	Formatting alone shifted accuracy up to 76 points[14]; partly a scoring artifact[15]; one small change altered model rankings in 63% of cases[16]	Task- and evaluation-dependent
10. Model version instability	A tool that worked well can behave differently after an update, often without notice	Record the tool and version used for each draft; retest the workflow after updates	The same named model, GPT-4, fell from 84% to 51% on one task over 3 months in 2023; other tasks moved differently[17]	Shown in 2023; no recent comparable data
■ **Human-AI loop** · *the writer defers to the AI, the AI defers to the writer, and nobody checks the source*				
11. Automation bias (user side)	The writer trusts AI output over their own correct judgment	Author-led source verification, independent of the AI's explanation; protected review time	Biased AI cut clinician accuracy by 11.3 percentage points, even with explanations[18]; 7% of correct pathology judgments overturned, worse under time pressure[19]	A human factor, not a model property
12. Sycophancy (model side)	The AI agrees with the writer instead of the source: it gives way when challenged, or follows a leading prompt	Neutral prompts; settle disagreements by checking the source, not by asking the AI again; memory off or project-scoped	Up to 100% initial compliance with illogical medical requests[20]; memory profiles raised agreement by up to 45% in 1 model[21]	No early baseline
■ **Emerging risks** · *watch items*				
13. Prompt injection	Hidden instructions in a source file (eg, white text, tiny fonts, hidden layers, metadata) steer what the AI writes, without the writer knowing	Use only controlled sources of known origin; verify output against source; escalate unexplained deviations to quality or IT	All 4 vision-language models tested followed instructions hidden in oncology images[22]	Emerging; not shown in regulatory documents
14. Temporal staleness	Outdated guidance or labeling is presented as current, from the AI's dated knowledge or an old source file	Confirm current versions of guidance and labeling, in AI output and in the files supplied to it	*Not yet measured*	Watch item

What the evidence shows

No failure mode has been eliminated. Newer models fabricate fewer references, yet in a 2026 test, fewer than one-third of their references were fully correct. Newer models overgeneralized more than earlier ones in one study, and problems with long documents and run-to-run consistency persisted in 2025-2026 models. For most failure modes, there isn't enough comparable data to tell whether they are improving. Newer models are not automatically safer. Understanding the science and checking every source remain the writer's job.

47% references fabricated, GPT-3.5, 2023[4]

31.8% fully correct references, 2026 models[6]

4.85× odds of overgeneralizing vs human summaries[7]

11 of 13 models below 50% of baseline at 32K tokens[9]

63% model rankings changed by one small prompt change[16]

84→51% GPT-4 accuracy on one task over 3 months[17]

−11.3 points of clinician accuracy with biased AI[18]

100% peak compliance with illogical medical requests[20]

Implications

For writers: 3 skills catch most AI failures: (1) checking every statement against its source, (2) knowing what each section must contain, and (3) understanding the science well enough to spot what's wrong. AI in the workflow makes these skills even more important.

For teams: Use the framework to onboard writers to AI tools, to update QC checklists, and to decide which documents and tasks need the closest review. Several failure modes, such as sycophancy and model version changes, have no QC check today.

For training: Teach how AI fails, not just how to prompt it.

Continue the conversation

Scan for references, the handout, and a 2-minute survey ·

accelwrite.com/amwa2026

Contact: ssampathkumar@accelwrite.com · AccelWrite LLC