AEROPROC-BENCH: DISTINCT CITATION AND SAFETY FAILURES CREATE A DEPLOYMENT DILEMMA IN AVIATION MAINTENANCE LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models deployed as maintenance assistants can cite manual references correctly while silently omitting mandated safety steps, a failure invisible to citation-based evaluation. We introduce AeroProc-Bench, a 4-dimensional benchmark built from 14,721 real aircraft fault-isolation items, and use it to reveal a *deployment dilemma*: citation accuracy and safety coverage are structurally independent per item ($r = 0.034$) yet covary across methods ($r = 0.806$), so no training-free method optimizes both. The dilemma manifests as a ranking inversion at the frontier: the safest model achieves SEF 0.834 but scores only CEF 0.196, meaning a citation-only deployer would reject the safest available model. Supervised fine-tuning escapes the dilemma by dominating both axes, but is unavailable for frontier APIs. Single-metric evaluation therefore systematically misleads deployment in safety-critical domains; multi-dimensional, per-domain validation is essential. Code and scoring scripts are released in the supplementary material.

1 INTRODUCTION

Deploying LLMs in safety-critical procedural domains on the basis of a single evaluation metric leads to systematically wrong decisions: a model that cites its sources correctly can still omit every mandated safety step. We show this with AeroProc-Bench, where citation accuracy and safety-precondition coverage are structurally independent failure types—whether a model gets one reference right tells you nothing about whether it includes the mandated depressurization step on that same item (per-item Pearson $r = 0.034$, and consistently low across all 12 methods, mean $|r| = 0.11$). The method-level picture is the opposite: interventions that improve citation tend to also improve safety ($r = 0.806$ across 12 methods). These two facts together create a *deployment dilemma*: methods broadly help both, but *among training-free methods* no single method dominates. The citation-optimal and safety-optimal interventions diverge, and each intervention targets a different failure type: context-contrast favors safety, best-of-N favors citation, and self-consistency voting (SC-vote$\geq$2) collapses safety—a divergence invisible to single-metric benchmarks. We study training-free interventions because they are the deployable set in exactly the settings where the dilemma bites: the safest models are frontier APIs that cannot be fine-tuned, procedural gold annotations are costly and domain-locked, and maintenance operations require interventions that port across model upgrades. Supervised fine-tuning dominates both axes (CEF 0.505, SEF 0.548) and serves as the supervised ceiling throughout.

We introduce AeroProc-Bench, a 4-dimensional evaluation framework for procedural hallucination in aviation maintenance. The benchmark comprises 14,721 real fault-isolation items from two aircraft types, annotated with 49,041 gold Aircraft Maintenance Manual (AMM) references and 7,784 gold safety preconditions across 3,760 items. Unlike hallucination benchmarks that assess world-knowledge facts (Lin et al., 2022; Li et al., 2023a; Min et al., 2023), AeroProc-Bench evaluates *intra-document procedural accuracy*: whether the model cites the correct manual references and includes mandated safety steps. The 4 dimensions—CEF (citation recall), CLC (causal-branch confusion), XFA (cross-fault mis-attribution), SEF (safety recall)—span two failure types validated by inter-annotator agreement (gold-reference $\kappa = 0.77$; CEF $\kappa = 0.86$; SEF judge–human $\kappa = 0.82$).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Positioning vs. representative benchmarks.

Benchmark	Domain	Failure type	Metric validated	Multi-size
TruthfulQA	world facts	confabulation	partial	no
HaluEval	general	confabulation	no	no
FActScore	world facts	imprecision	partial	no
HalluLens	mixed	multiple	partial	no
AviationQA/ALUE	aviation	factual QA	no	no
AeroProc-Bench	**aviation proc.**	**citation+safety**	$\kappa = 0.86$	**yes**

Our contributions are empirical findings enabled by the benchmark:

1. **Single-metric evaluation misleads deployment (the deployment dilemma).** Citation and safety-precondition coverage are structurally independent failure types—per-item $r = 0.034$, consistently low across all 12 methods (mean $|r| = 0.11$)—yet covary across methods ($r = 0.806$). Single-metric evaluation therefore both *misses* safety failures and *misranks* methods: context-contrast ranks 6th on CEF but 2nd on SEF (safety-selective), best-of-N improves CEF but leaves SEF at baseline (citation-selective), and SC-vote$\geq$2 keeps CEF near baseline (0.225) yet collapses SEF to 0.015.
2. **Citation ranking inverts under safety evaluation.** DeepSeek-V4-pro, the one frontier model we probed on CEF, shows low citation recall (CEF 0.196, not significantly below Qwen-7B 0.237) yet high safety recall (SEF 0.834, statistically indistinguishable from the other five frontier models); all six cluster at high safety (0.78–0.83), far above the best open model (SFT 0.548), while within the open Qwen family SEF decreases with scale (0.5B 0.377 $>$ 7B 0.223). A citation-only deployer would avoid a model in the safest cluster.
3. **The benchmark further surfaces three deployment-relevant patterns:** (i) RAG gains scale with model size within Qwen (-14% at 0.5B $\rightarrow$ $+47\%$ at 7B), but the strongest model Llama-3B drops (-9%); (ii) steering is scale-fragile—ITI α is non-portable across sizes, and the wrong α collapses CEF by 90%; (iii) the information-adding vs. representation-manipulating taxonomy replicates on CFR (n=40, CIs exclude 0) but *inverts* on Rail (n=27)—domain-specific, not universal.

2 Related Work

LLM Benchmarks. MMLU (Hendrycks et al., 2021), BIG-Bench (Srivastava et al., 2023), HELM (Liang et al., 2023), GLUE/SuperGLUE (Wang et al., 2018; 2019), BIG-Bench-Hard (Suzgun et al., 2023). BasicTS+ (Shao et al., 2025) introduced heterogeneity-aware benchmarking. AeroProc-Bench targets intra-document procedural accuracy.

Hallucination. Ji et al. (2023) survey hallucination. TruthfulQA (Lin et al., 2022), HaluEval (Li et al., 2023a), FActScore (Min et al., 2023), HalluLens (Bang et al., 2025), FaithDial (Dziri et al., 2022), TRUE (Honovich et al., 2022), SelfCheckGPT (Manakul et al., 2023). AeroProc-Bench differs: ground truth is intra-document, failure modes are omission-based. Table 1 positions the benchmark.

Steering and Interventions. CAA (Rimsky et al., 2024), ITI (Li et al., 2023b), DoLa (Chuang et al., 2024), self-consistency (Wang et al., 2023), RepE (Zou et al., 2023), activation addition (Turner et al., 2023), contrastive decoding (Li et al., 2023c), DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), RLHF (Ouyang et al., 2022), CoT (Wei et al., 2022), ICL (Dong et al., 2023; Brown et al., 2020), LLM-as-judge (Zheng et al., 2023), G-Eval (Liu et al., 2023).

Aviation NLP. AviationQA (Agarwal et al., 2022), ALUE (Mangortey et al., 2025), CAMB (Zhang et al., 2025). Models: Qwen2.5 (Qwen Team, 2024), Llama-3 (Meta, 2024), Mistral (Jiang et al., 2023), Phi-3 (Microsoft, 2024), DeepSeek (DeepSeek-AI, 2024). Methodology: Krippendorff (Krippendorff, 2018), QLoRA (Dettmers et al., 2023), LoRA (Hu et al., 2022). Most closely related: Xu et al. (2025).

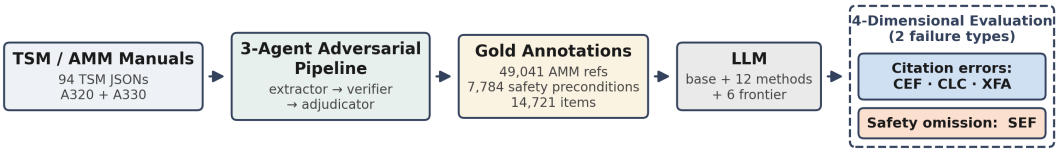

Figure 1: AeroProc-Bench evaluation pipeline. Gold references and safety preconditions are extracted from TSM/AMM manuals via a 3-agent adversarial pipeline. Model outputs are evaluated along 4 dimensions spanning 2 failure types: citation (CEF/CLC/XFA) and safety (SEF).

Table 2: AeroProc-Bench statistics.

Statistic	Value
Items	14,721
Gold AMM references	49,041
Gold safety preconditions	7,784 (3,760 items)
Aircraft types	A320 + A330 (94 TSM)
Difficulty (E/M/H)	4,782 / 5,928 / 4,011
Balanced hard eval	240 items (12 methods, 6 open models)

3 Benchmark Construction

The contribution is threefold: (1) first benchmark targeting intra-document citation accuracy and safety-precondition completeness in procedural text; (2) 4-dimensional, 2-failure-type evaluation decomposing the failure space into non-redundant axes; (3) deterministic scoring enabling exact replication.

3.1 Data Source and Statistics

Constructed from 94 real Troubleshooting Manual (TSM) JSONs (58 A320 + 36 A330). Gold AMM references extracted via domain-specific regex (5-segment format), yielding 14,721 items with 49,041 references (Table 2; overall pipeline in Figure 1).

3.2 Gold Extraction and Verification

Action-verb triggers bind AMM references to maintenance actions. A 3-agent pipeline (extractor/verifier/adjudicator) yields 0 orphans, 250/250 faithful, 99.8% in source. Safety preconditions extracted from CAUTION-/WARNING-tagged lines via safety-verb lexicon; 3,760/14,721 items have ≥ 1 gold safety. CLC probes from TSM IF/THEN branches; XFA pairs from titles with similarity ≥ 0.5 but zero Jaccard overlap. Cross-domain: CFRProc (40 items, 21 CFR, public) and RailProc (27 CR300AF items).

3.3 Four-Dimensional Evaluation Protocol

CEF: citation omission (recall of gold refs in output). CLC: causal-branch confusion. XFA: cross-fault mis-attribution. SEF: safety-precondition omission via LLM judges (DeepSeek-V4-flash and GLM-5.2, temp=0). Citation scoring is deterministic (greedy decoding, do_sample=False, seed 0; repetition_penalty=1.05—the Qwen2.5 generation-config default—applied uniformly to the base and all methods; per-item checkpoint); SEF validated in Section 4.5.

Table 3: Main results—7B, all 4 dimensions, 12 methods. Sample sizes differ by dimension: CEF n=240 (hard subset), CLC/XFA n=300, SEF n=80.

Method	CEF	CLC rec.	CLC conf.	XFA acc.	SEF
base	0.230	0.481	0.220	0.612	0.223
few-shot	0.399	0.530	0.193	0.682	0.390
SFT	**0.505**	**0.580**	0.383	**0.750**	**0.548**
context-contrast	0.320	0.441	**0.166**	0.608	0.455
best-of-N	0.316	0.486	0.198	0.600	0.235
CAA	0.240	0.502	0.142	0.577	0.279
DoLa	0.200	0.436	0.140	0.528	0.237
SC union	0.350	0.591	0.253	0.650	0.407
SC-vote$\geq$2	0.225	0.437	**0.108**	0.543	0.015
ITI	0.222	0.433	0.161	0.518	0.223
fs+cc	0.424	0.542	0.157	0.702	0.401
fs+bon	0.441	0.585	0.235	0.695	0.398

4 Experiments and Analysis

4.1 Experimental Setup

Models. Qwen2.5 (0.5B/1.5B/7B), Llama-3.2-3B, Phi-3.5-mini, Mistral-7B. Frontier: DeepSeek-V4-pro, Grok-4.5, GLM-5.2, Qwen3.7-max, MiniMax-M3, Kimi-K3-1 (all via API). All open models in bfloat16 (7B fits a single RTX 3090). 4$\times$RTX 3090.

Methods. *Training-free* (the deployment-relevant set—frontier APIs admit only these): few-shot, context-contrast, best-of-N (K=3), CAA (layer 14), DoLa, self-consistency (union/vote$\geq$2), ITI (top-48 heads). *Fine-tuned*: SFT (QLoRA) and RFT (representation-level fine-tuning: a contrastive loss on the residual stream that directly restructures the model's geometry). *Combos*: fs+cc, fs+bon. *Retrieval*: RAG (TSM chunk retrieval).

Protocol. CEF on a 240-item hard subset; CLC on 300 causal-branch probes and XFA on 300 cross-fault pairs (each sampled from the benchmark, seed 0); SEF on an 80-item subset. Frontier models evaluated on the same 80 items with a matched token budget. Bootstrap CIs: $B = 10{,}000$, seed 0.

4.2 Main Results

Table 3 and Figure 2 reveal: (1) no method dominates all dimensions; (2) best-of-N raises CEF but leaves SEF at baseline; (3) representation-manipulating methods do not substantially improve CEF (CAA 0.240 within noise of base 0.230; ITI 0.222 and DoLa 0.200 below base); (4) SC-vote$\geq$2 catastrophically collapses SEF (0.015).

Cross-scale. Context-contrast improves CEF consistently across sizes. DoLa collapses on 0.5B. ITI requires per-scale α (Table 4): applying 7B's α=3.0 to 0.5B (optimal α=1.0) collapses CEF from 0.341 to 0.034 (-90%). CAA is similarly scale-fragile: α=16—7B's optimal ($+4\%$)—collapses 0.5B CEF to 0.000 and drops 1.5B to 0.122.

4.3 Analysis: Drivers of the Deployment Dilemma

Citation–safety: distinct failures, method-level covariance. At the item level, CEF and SEF are independent ($r = 0.034$, $n = 68$; computed on the 80-item SEF set after dropping items whose gold reference list is empty, for which CEF is undefined): getting a reference right on one item says nothing about safety compliance on that item. This independence is not specific to the base policy—per-item r is consistently low across all 12 methods (mean $|r| = 0.11$, range $[-0.05, 0.32]$, n=68 each). At the method level, however, the 12 (CEF, SEF) pairs are positively correlated ($r = 0.806$, $p < 0.01$; Spearman 0.774, $p < 0.01$). The two layers are not contradictory: failures are mechanistically distinct, but methods that add procedural information tend to help both. The deployment dilemma

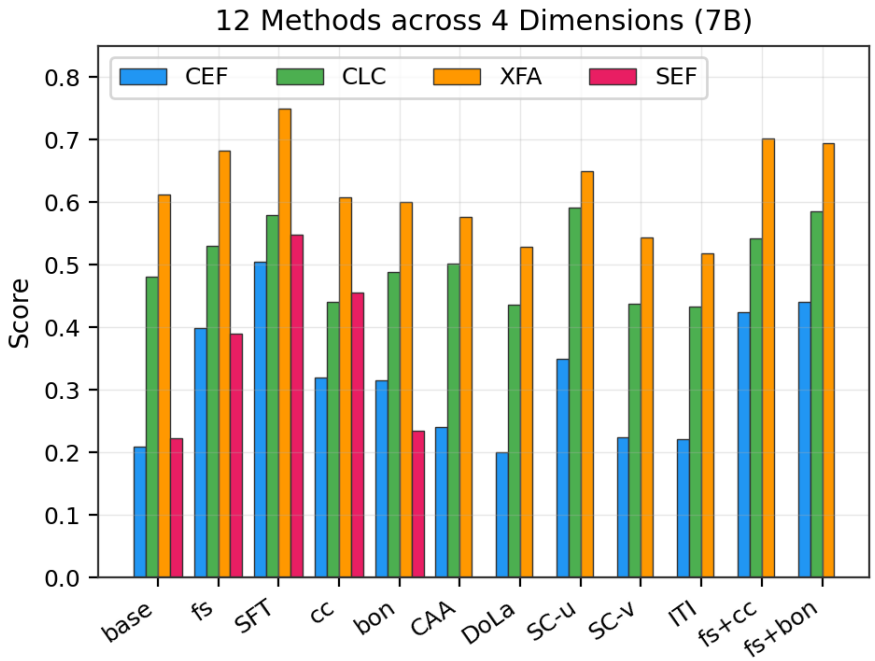

Figure 2: 12 methods across 4 dimensions (7B). No single method dominates. SFT leads CEF/XFA but worst CLC confusion; context-contrast yields the largest SEF gain among training-free methods while also improving CEF.

Table 4: Cross-scale CEF (Qwen2.5). CAA reported at per-scale best α (7B: α=16; 0.5B/1.5B: α=2); ITI at per-scale α (0.5B: 1.0; 1.5B/7B: 3.0).

Method	0.5B	1.5B	7B
base	0.290	0.218	0.230
few-shot	0.097	0.258	0.399
context-contrast	0.431	0.361	0.320
best-of-N	0.301	0.274	0.316
CAA	0.238	0.183	0.240
DoLa	0.000	0.238	0.200
ITI	0.341	0.337	0.222

arises from the *exceptions*: context-contrast (6th CEF, 2nd SEF — safety-selective), best-of-N (citation-selective: CEF up, SEF flat at baseline), and SC-vote$\geq$2 (normal CEF, catastrophic SEF 0.015). These divergences are invisible to single-metric benchmarks (Figure 3).

Frontier models and scaling. All six frontier models achieve a uniformly high SEF in a tight 0.78–0.83 band (Table 5): DeepSeek-V4-pro (0.834), Kimi-K3-1 (0.829), Grok-4.5 (0.828), MiniMax-M3 (0.823), GLM-5.2 (0.816), Qwen3.7-max (0.783)—far above the best open model (SFT 0.548). Differences among frontier models are negligible (overlapping CIs); safety recall is not driven by the "reasoning" vs "non-reasoning" distinction. Within the *open* Qwen family, SEF decreases with scale (0.5B: 0.377 $>$ 1.5B: 0.313 $>$ 7B: 0.223), but this is a training-specific pattern, not a general scaling law. The inversion is the dilemma's second face: a citation-only evaluation would reject the entire safest cluster of models.

RAG baseline (4 models). RAG (TSM chunk retrieval, top-5) exhibits a within-Qwen scale dose-response: CEF $\Delta = -14\%$ (0.5B), $+5\%$ (1.5B), $+47\%$ (7B); SEF $\Delta = +1\%, +5\%, +35\%$. Gains rise monotonically with model size—the smallest model cannot use retrieved context and is hurt, the largest gains most. Cross-family, the strongest model (Llama-3B, base CEF 0.408, n=240) is the

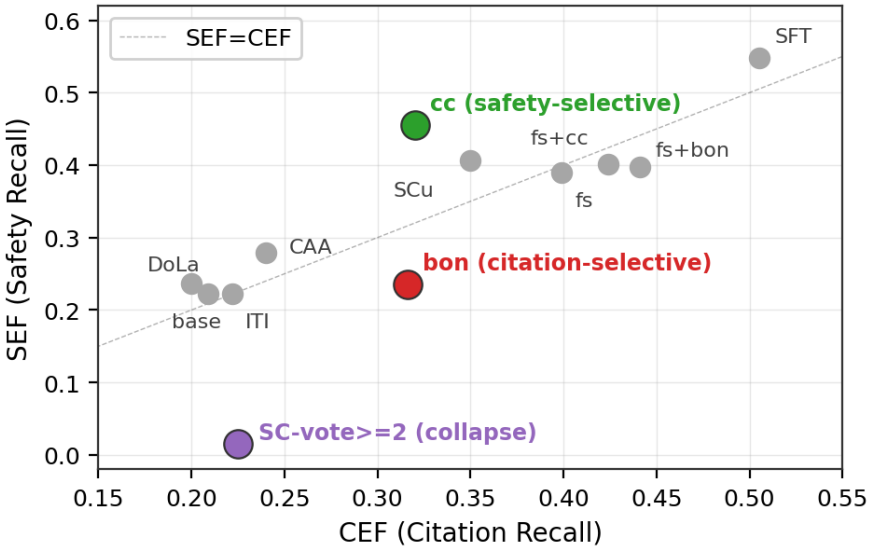

Figure 3: CEF vs. SEF (7B, 80 items). Context-contrast lies above the SEF = CEF diagonal (safety-selective); best-of-N lies below it (citation-selective); SC-vote≥2 collapses to the floor. Dashed: SEF = CEF.

Table 5: SEF across models and methods (80-item subset, 95% CI).

Model / Method	SEF	95% CI
DeepSeek-V4-pro (frontier)	**0.834**	[0.774, 0.889]
Kimi-K3-1 (frontier)	0.829	[0.768, 0.884]
Grok-4.5 (frontier)	0.828	[0.767, 0.882]
MiniMax-M3 (frontier)	0.823	[0.764, 0.878]
GLM-5.2 (frontier)	0.816	[0.752, 0.873]
Qwen3.7-max (frontier)	0.783	[0.711, 0.847]
SFT (Qwen 7B)	0.548	[0.468, 0.627]
context-contrast (7B)	0.455	[0.373, 0.538]
few-shot (7B)	0.390	[0.313, 0.467]
Qwen 0.5B base	0.377	[0.293, 0.462]
Qwen 1.5B base	0.313	[0.231, 0.397]
Qwen 7B base	0.223	[0.169, 0.281]
SC-vote≥2 (7B)	0.015	[0.003, 0.033]

only one that drops (-9%). RAG creates no citation–safety trade-off: SEF never decreases across all 4 models. RAG remains below SFT (7B RAG CEF 0.338 $<$ SFT 0.505; SEF 0.301 $<$ 0.548). RAG thus escapes the dilemma—safety never decreases—without resolving it: both axes remain below the supervised ceiling.

SEF stratification. Information-adding $>$ base: SFT (0.548), context-contrast (0.455), SC union (0.407), fs+cc (0.401), fs+bon (0.398), few-shot (0.390). Representation-manipulating $\approx$ base: CAA (0.279), DoLa (0.237), ITI (0.223). SC-vote≥2 catastrophic (0.015).

Frontier CEF. On 50 items: Llama-3B = 0.402, Qwen-7B = 0.237, DeepSeek = 0.196. Llama significantly $>$ DeepSeek ($+0.207$, CI $[+0.077, +0.334]$); Qwen-7B vs DeepSeek is not significant ($+0.041$, CI $[-0.073, +0.156]$). The one frontier model we probed does not improve citation recall.

RFT collapse threshold. Scaling the trained RFT LoRA by τ (zero-retrain `set_scale`) exposes a sharp phase transition, not a gradual trade-off. Mild restructuring (τ=0.25, representational similarity to the base CKA$\sim$0.67, where CKA is centered kernel alignment) preserves citation recall (CEF

0.243 vs base 0.231; Δ+0.012, 95% CI $[-0.017, +0.042]$, n=222). Full RFT (τ=1.0, CKA$\sim$0.23) collapses CEF to 0.000 (a drop of 0.231, CI $[+0.200, +0.263]$ on the base$-$method scale, excludes 0). CEF drops from 0.243 to 0.000 between τ=0.25 and τ=0.5—a restructuring threshold below which citation recall catastrophically fails.

Error analysis. The dominant error mode is omission and misrouting of *real* references, not fabrication. *CEF*: on the hardest item (24 gold references), the base model cites 1/24—it names the correct subsystem but drops parallel branches whose descriptions share surface vocabulary, the same lexical-overlap trap the XFA construction targets (title similarity ≥ 0.5, zero reference overlap). *CLC*: on causal-branch probes the model cites the wrong-branch references (recall 0, confusion 1): sibling IF/THEN branches describe lexically similar tasks, and the model retrieves the topic without executing the conditional that separates the branches. *SEF*: the base model emits generic safety boilerplate while omitting all five item-specific preconditions (e.g., depressurization); SFT recovers 4/5—the knowledge is elicitable, but training-free prompting does not elicit it. All three dimensions measure the same underlying deficit from different angles: the required information is present in the context or the weights, and the model fails to surface it.

4.4 Cross-Domain Generalization

To test whether the information-adding vs. representation-manipulating split is an aviation artifact, we evaluate on CFRProc (40 items, 21 CFR medical-device regulations, public) and RailProc (27 CR300AF high-speed-train items). On CFR, information-adding methods significantly beat base—context-contrast $+0.123$ $[+0.037, +0.219]$, best-of-N $+0.139$ $[+0.015, +0.269]$—while representation-manipulating methods do not. On Rail the split does *not* replicate: context-contrast significantly *hurts* (-0.138 $[-0.208, -0.074]$), best-of-N is null ($+0.027$ $[-0.006, +0.074]$), and ITI significantly *helps* ($+0.041$ $[+0.004, +0.089]$). The taxonomy replicates on CFR but inverts on Rail; it is domain-specific, not a universal regularity.

4.5 Metric Validation

Two human annotators independently labeled Task A (n=50, gold-reference correctness) and Task B (n=110, citation presence). Inter-annotator $\kappa = 0.86$ on Task B counting UNSURE as a valid category (7 UNSURE$\leftrightarrow$NOT_COVERED boundary disagreements), rising to 1.00 after resolving UNSURE; Task A $\kappa = 0.77$ (NOT_REQUIRED boundary cases). Both exceed the 0.75 threshold. DeepSeek-V4-pro annotated the full 1,393-item set; annotation quality was validated on a 160-item human subset (IAA above). Auto-CEF vs human recall: Pearson $r = 1.00$ ($n = 85$). Gold precision $\sim$98%. For the SEF dimension, we validated the DeepSeek-V4-flash safety-coverage judge against a human annotator on $n = 345$ (safety-precondition, output) pairs: Cohen's $\kappa = 0.82$, accuracy 94.8% (substantial; 327/345 agreements, diverging mainly on 9 COVERED-vs-NOT_COVERED boundary cases).

5 Conclusion

We introduced AeroProc-Bench, a 4-dimensional benchmark that decomposes procedural hallucination in aviation maintenance into two failure types—citation errors (CEF/CLC/XFA) and safety-precondition omission (SEF)—validated by inter-annotator agreement. Across 6 open models, 12 methods, and 6 frontier models, the two types behave as distinct failures: independent per item ($r_{\text{item}} = 0.034$) yet covarying across methods ($r_{\text{method}} = 0.806$). This creates a deployment dilemma among training-free interventions: context-contrast is safety-selective, best-of-N is citation-selective, and self-consistency voting keeps CEF at baseline while collapsing SEF to 0.015. Supervised fine-tuning dominates both axes and remains the ceiling; RAG never reduces safety across the 4 models tested, with gains that scale within Qwen but reverse on the strongest model; steering strengths do not transfer across model sizes, and the wrong strength is catastrophic. At the frontier the dilemma inverts: all six frontier APIs cluster at high safety recall (0.78–0.83) while the one probed on citation stays low (CEF 0.196), so single-metric evaluation misranks models in both directions. The cross-domain probe tempers generality—the information-adding vs. representation-manipulating taxonomy replicates on CFR but inverts on Rail—so method rankings should be re-validated per domain. For practitioners: audit both failure types before deployment, select interventions by the axis that matters, and treat

voting-based methods as unsafe for safety-critical procedures until an SEF check passes. No single metric or intervention suffices.

Limitations. Aviation-primary. The cross-domain taxonomy replicates on CFR (n=40) but inverts on Rail (n=27) (Section 4.4). RFT collapse threshold is shown on a single model family (n=222). Compute: all open-model inference fits a single 24 GB GPU (one 7B model at a time, greedy decoding); the SEF judge adds 9,200 API calls (25 systems $\times$ 368 (precondition, output) pairs), and the frontier rows one API generation per item.

AI USE STATEMENT

AI assisted with code generation and prose drafting. All experimental design, annotation, analyses, and conclusions were formulated and verified by the authors.

ETHICS STATEMENT

Proprietary Airbus data under license; none publicly released. No human subjects. Annotators participated with informed consent. Dual-use risk minimal (requires licensed manuals).

REPRODUCIBILITY STATEMENT

Citation scoring deterministic (greedy, seed 0, checkpoint). SEF judge log, scoring scripts, and code are in the supplementary material; public release upon acceptance. Bootstrap $B = 10{,}000$, `default_rng(0)`. Cross-domain data public.

REFERENCES

D. Hendrycks et al. Measuring Massive Multitask Language Understanding. ICLR 2021.

A. Srivastava et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. TMLR 2023.

P. Liang et al. Holistic Evaluation of Language Models. TMLR 2023.

A. Wang et al. GLUE: A Multi-Task Benchmark and Analysis Platform. NeurIPS 2018.

A. Wang et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. NeurIPS 2019.

M. Suzgun et al. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ACL 2023.

Z. Shao et al. Exploring Progress in MTS Forecasting: Comprehensive Benchmarking and Heterogeneity Analysis. IEEE TKDE 37(1), 2025.

E. Lin et al. TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL 2022.

J. Li et al. HaluEval: A Large-Scale Hallucination Evaluation Benchmark. EMNLP 2023.

Y. Bang et al. HalluLens: LLM Hallucination Benchmark. ACL 2025.

S. Min et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision. EMNLP 2023.

Z. Ji et al. Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55(12), 2023.

N. Dziri et al. FaithDial: A Faithful Benchmark for Dialogue-Based Information Seeking. TACL 2022.

N. Honovich et al. TRUE: Re-evaluating Factual Consistency Evaluation. EMNLP 2022.

P. Manakul et al. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative LLMs. EMNLP 2023.

A. Zou et al. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405, 2023.

N. Rimsky et al. Steering Llama 2 via Contrastive Activation Addition. ACL 2024.

K. Li et al. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. NeurIPS 2023.

Y.-S. Chuang et al. DoLa: Decoding by Contrasting Layers Improves Factuality. ICLR 2024.

A. Turner et al. Activation Addition: Steering Language Models Without Optimization. arXiv:2308.10248, 2023.

X. Li et al. Contrastive Decoding: Open-ended Text Generation as Optimization. ACL 2023.

X. Wang et al. Self-Consistency Improves Chain of Thought Reasoning. ICLR 2023.

J. Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.

Q. Dong et al. A Survey on In-context Learning. arXiv:2301.00234, 2023.

Y. Meng et al. SimPO: Simple Preference Optimization with a Reference-Free Reward. NeurIPS 2024.

R. Rafailov et al. Direct Preference Optimization. NeurIPS 2023.

L. Ouyang et al. Training Language Models to Follow Instructions with Human Feedback. NeurIPS 2022.

L. Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023.

Y. Liu et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP 2023.

A. Agarwal, S. Gawade, S. Channabasavarajendra, and P. Bhattacharyya. There is No Big Brother or Small Brother: Knowledge Infusion in Language Models for Link Prediction and Question Answering. In Proc. ICON, pages 204–211, 2022.

E. Mangortey, S. Singh, S. Chen, and K. K. Sarkhel. Aviation Language Understanding Evaluation (ALUE): Large Language Model Benchmark with Aviation Datasets. In Proc. AIAA AVIATION Forum and ASCEND, 2025.

F. Zhang, C. Pang, Y. Zhang, and C. Luo. CAMB: A Comprehensive Industrial LLM Benchmark on Civil Aviation Maintenance. arXiv:2508.20420, 2025.

Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

Meta. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

A. Jiang et al. Mistral 7B. arXiv:2310.06825, 2023.

Microsoft. Phi-3 Technical Report. arXiv:2404.14219, 2024.

DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434, 2024.

K. Krippendorff. Content Analysis: An Introduction to Its Methodology. Sage, 4th ed., 2018.

T. Dettmers et al. QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS 2023.

E. Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.

T. Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

F. Xu et al. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. IEEE TKDE 37(4), 2025.

A Additional Figures

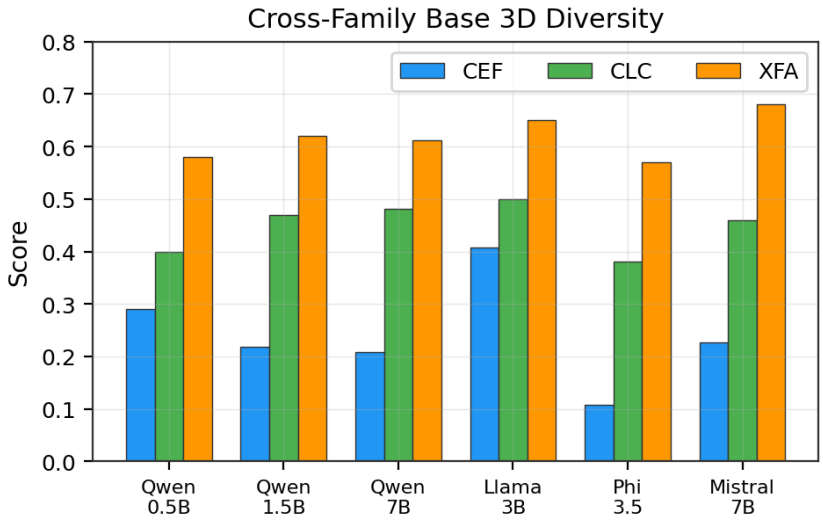

Figure 4: Cross-family base 3D diversity.

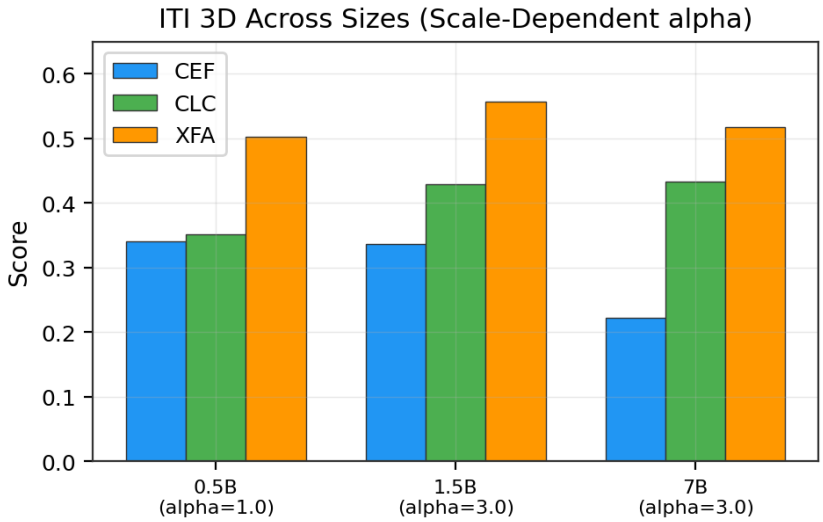

Figure 5: ITI 3D across sizes with scale-dependent α.

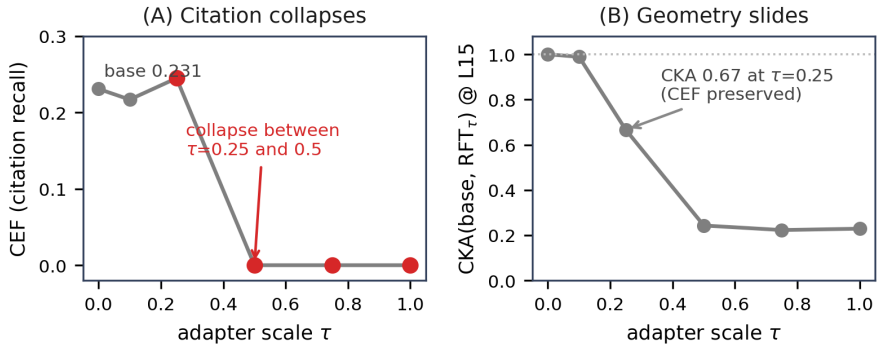

Figure 6: RFT restructuring threshold (Qwen2.5-7B): representational similarity to the base (CKA, layer 15) slides gradually with adapter scale τ, while citation recall collapses to 0.000 between τ=0.25 and τ=0.5 (points: n=240 sweep; τ=0 is the adapter-free base; Section 4.3 reports CIs on the n=222 subset).

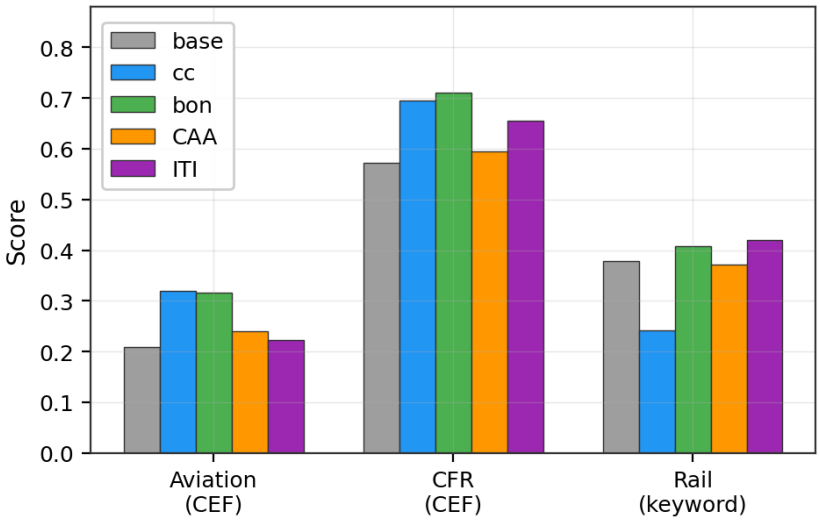

Figure 7: Cross-domain method transfer.

B Prompt Templates

Task generation (system). "You are an aviation maintenance troubleshooting assistant. Given the fault and the TSM context, write the complete corrective procedure with all AMM references and safety preconditions."

Task generation (user). "Fault: {title}. Possible causes: {causes}. TSM context: {fault_isolation_text}. Write the complete corrective procedure."

SEF judge rubric (DeepSeek-V4-flash and GLM-5.2, temp=0). "Judge whether this safety precondition is covered. COVERED / NOT_COVERED / UNSURE. Reply with EXACTLY one word."

ITI probe labels. Tokens on lines containing "Ref. AMM <number>" labeled 1; others 0. Logistic regression per head (top-48) identifies truthful direction.