

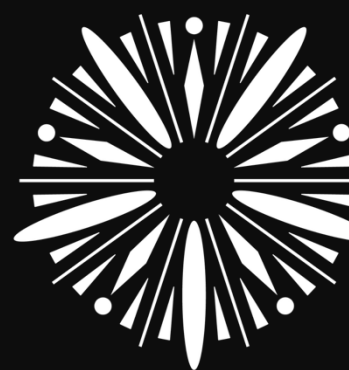
Diagnostic AI fears false negatives.

Gen AI fears broken pipelines.

Gen AI Reporting vs. Diagnostic AI — A Post-Market Surveillance Error Comparison

TANGO PMS Monte Carlo Simulation (n=68, seed=42) · Herpe et al. Real-World Dataset (n=942)

Herpe et al. · TANGO PMS Simulation (n=68, seed=42) · Herpe et al. 2025–2026 (n=942)



CHU
Poitiers



PURPOSE

Diagnostic error patterns in imaging AI have been extensively characterised in real-world post-market surveillance datasets; however, the **failure modes specific to Gen AI language-based reporting tools** remain poorly described.

This study aimed to compare the distribution of PMS error categories between **TANGO** — a Gen AI radiology report prefilling system — and a **validated benchmark of diagnostic imaging AI solutions** (Herpe et al., n=942), in order to identify structural differences in error profiles and derive targeted **safety governance implications**.

MATERIAL

TANGO PMS Dataset

7 original incident tickets spanning 4 product lines (TANGO Prefilling, TANGO Prefilling OLD, TANGO Standalone OLD, TANGO License & UI), **Monte Carlo-simulated to n=68** using a base scale factor of 9.89 ($\pm 10\%$ uniform variation, seed=42). Each incident independently classified by two senior radiologists using the Herpe et al. diagnostic error taxonomy, with consensus adjudication by a third reviewer.

Benchmark Reference Dataset

n=942 real-world error declarations prospectively collected **August–December 2025** from **32 commercially deployed imaging AI solutions** across a European clinical AI distribution platform (Incepto Medical, Paris, France), covering approximately **5,000,000 annual imaging examinations**.

Study design: Simulated cross-sectional comparison · validated error taxonomy (Herpe et al.) · dual independent classification · consensus adjudication · EU AI Act PMS framework

Classification: Each error mapped to one of six categories from the Herpe et al. diagnostic error taxonomy, with blinded inter-rater classification and third-reviewer arbitration.

RESULTS

Error category distributions diverged substantially between TANGO Gen AI and diagnostic imaging AI. Key findings:

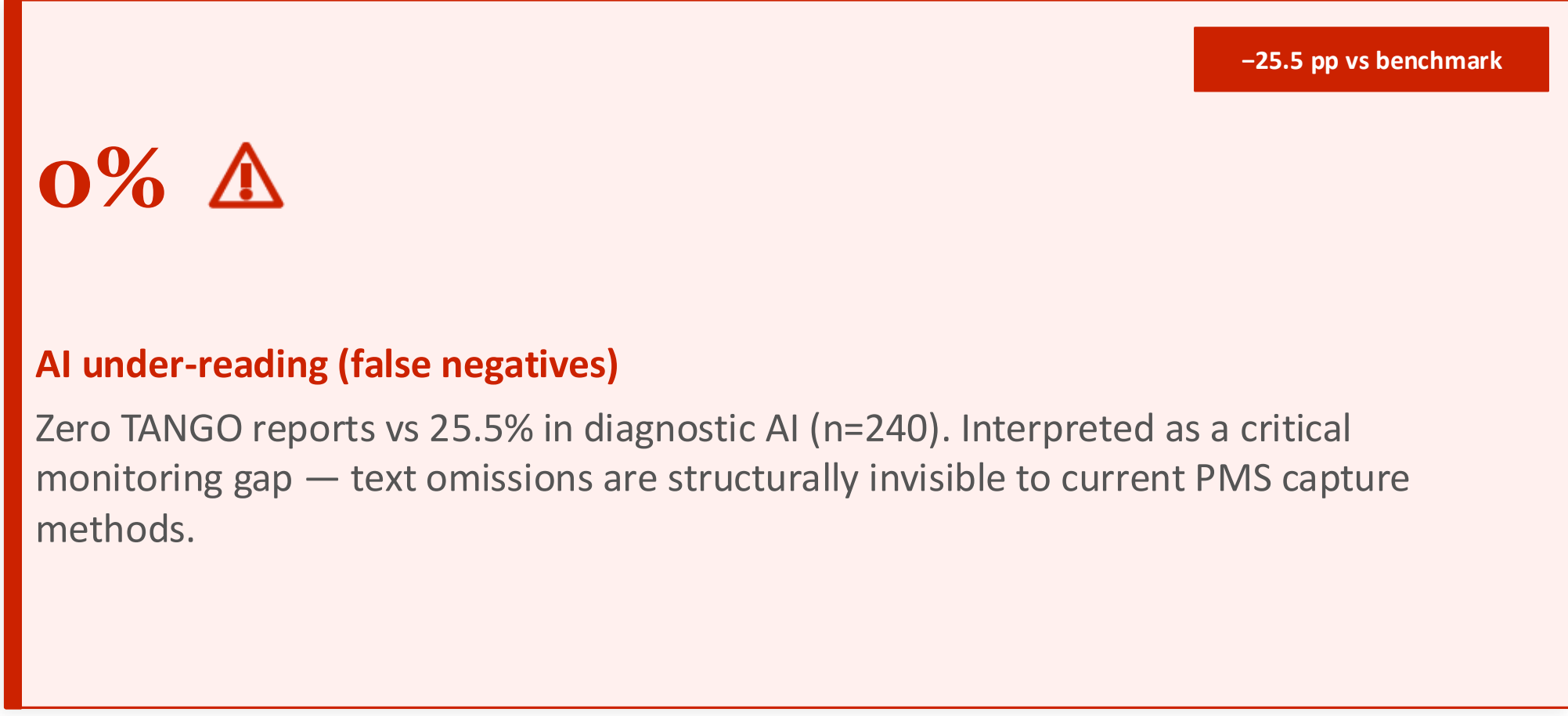
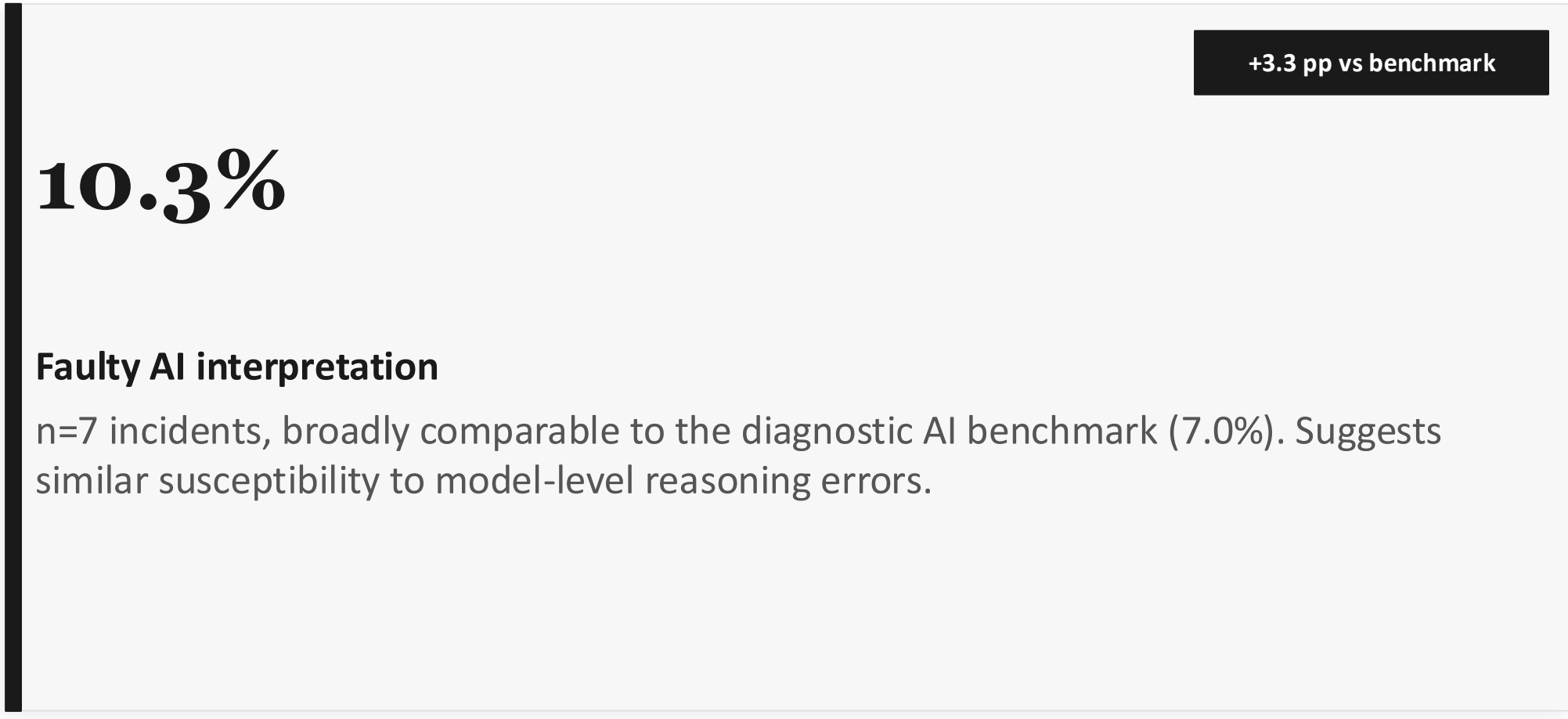
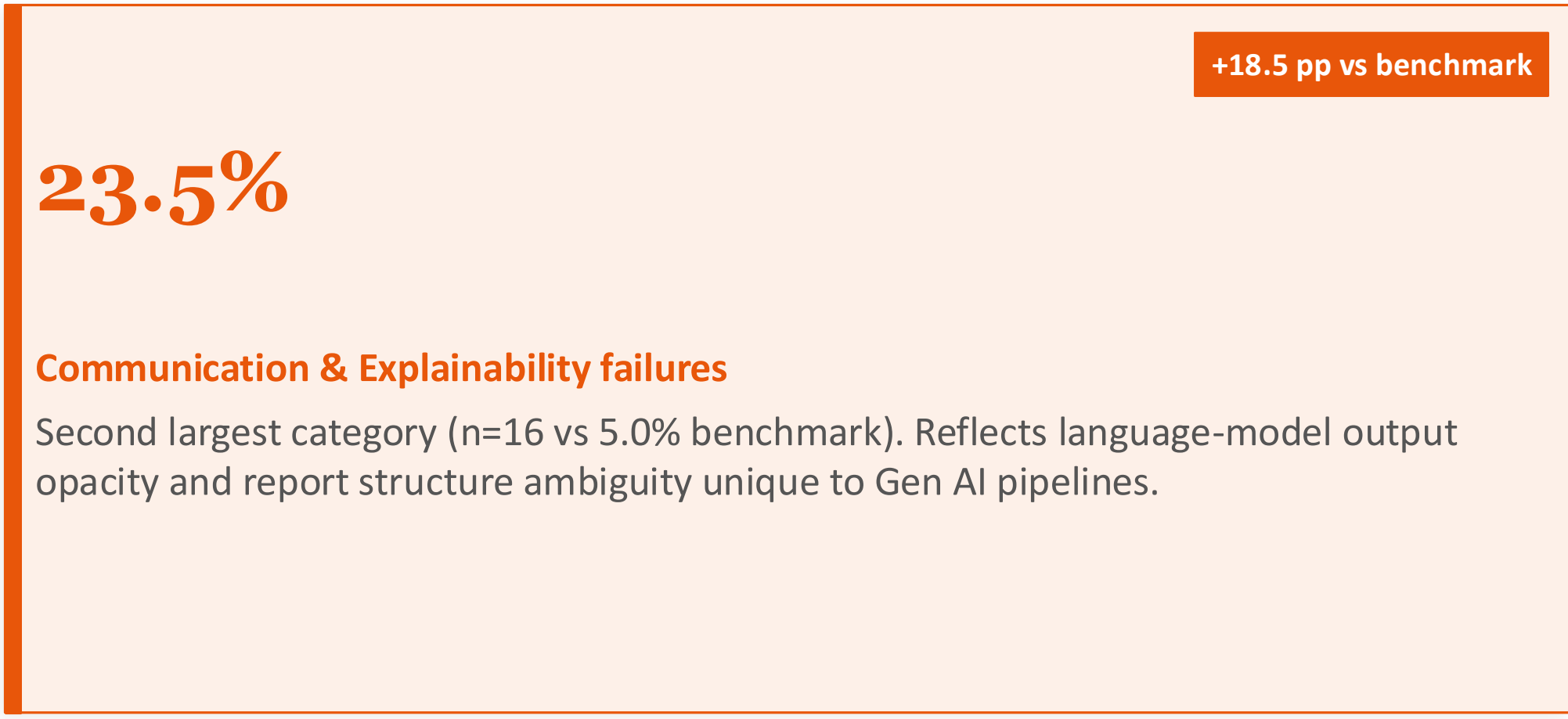
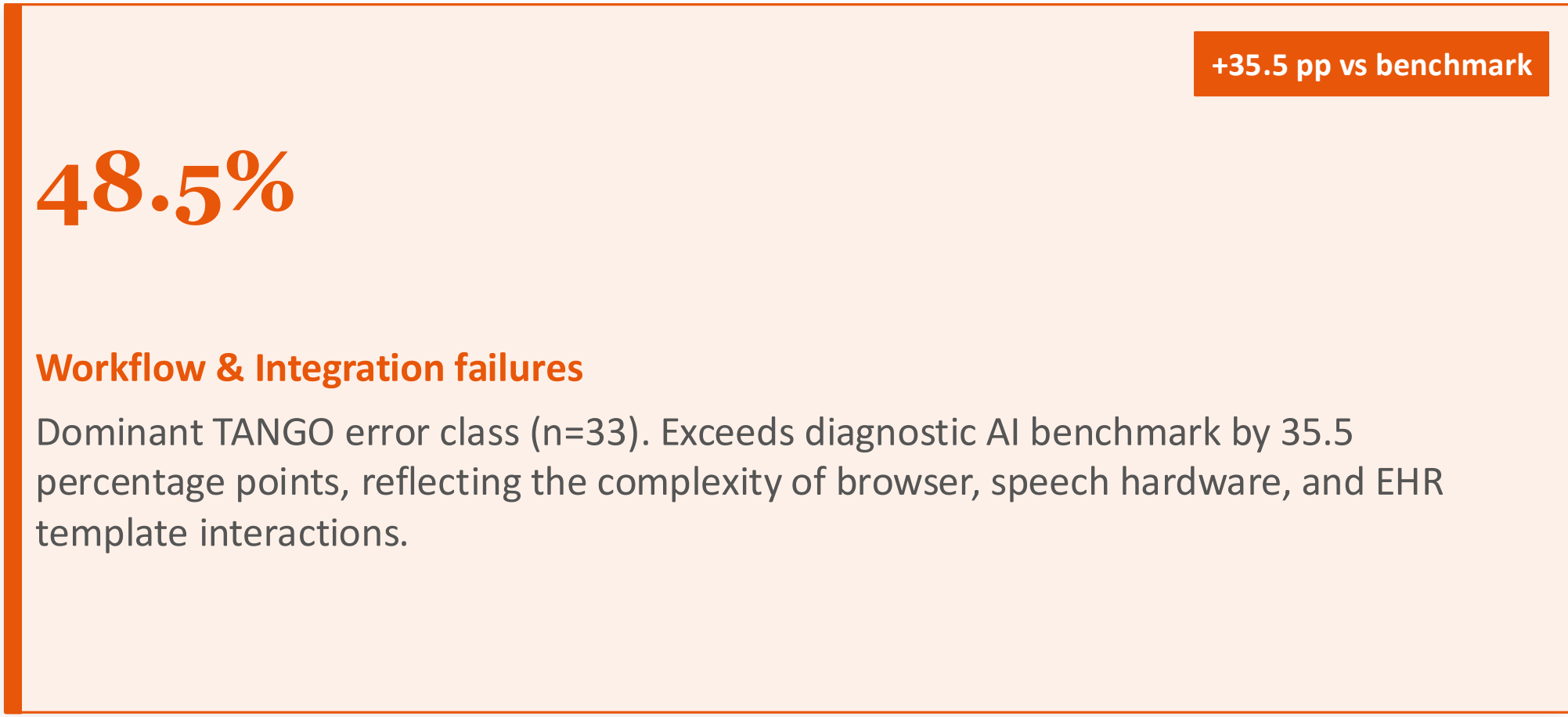
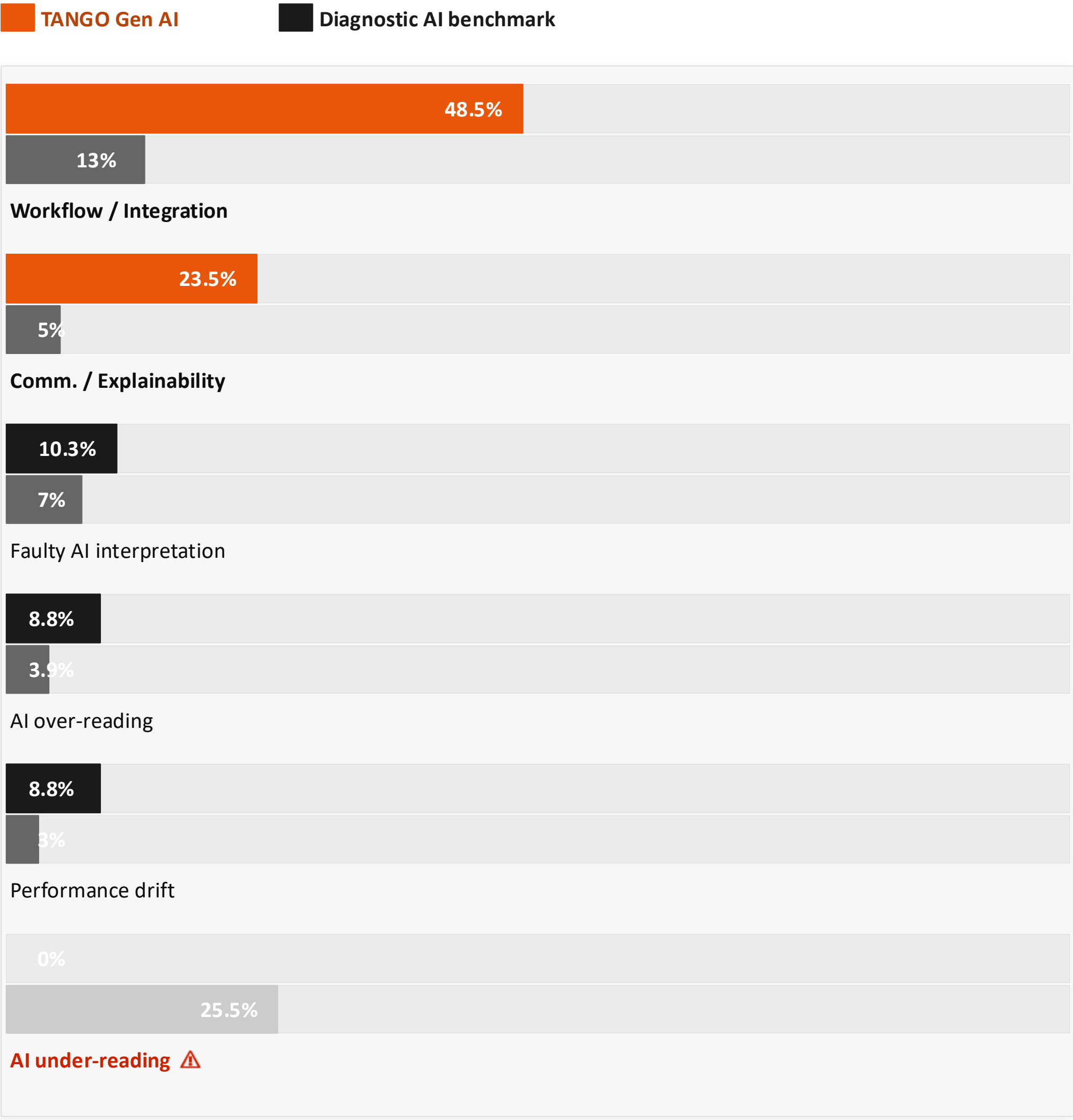


TABLE 1 — ERROR CATEGORY DISTRIBUTION

Error Category (Herpe Taxonomy)	TANGO Gen AI n=68 (%)	Diagnostic AI n=942 (%)	Delta (pp)
Workflow / Integration failure	33 (48.5%)	122 (13.0%)	+35.5
Communication / Explainability	16 (23.5%)	47 (5.0%)	+18.5
Faulty AI interpretation	7 (10.3%)	66 (7.0%)	+3.3
AI over-reading / Overcalling	6 (8.8%)	37 (3.9%)	+4.9
Performance drift	6 (8.8%)	~28 (3.0%)	+5.8
AI under-reading ⚠️	0 (0.0%)	240 (25.5%)	-25.5
TOTAL	68 (100%)	942 (100%)	—

⚠️ 0% AI under-reading is interpreted as an underreporting artefact, not a genuine safety advantage.

ERROR PROFILE COMPARISON



TAM CONSTRUCT MAPPING

Error Category	TAM Construct	TANGO %
Workflow / Integration	Perceived Ease of Use (PEOU)	48.5%
Communication / Explainability	Perceived Usefulness (PU)	23.5%
Faulty AI interpretation	Perceived Usefulness (PU)	10.3%
AI over-reading	Perceived Usefulness (PU)	8.8%
Performance drift	Perceived Ease of Use (PEOU)	8.8%

All TANGO error categories map exclusively to TAM constructs (PEOU and PU), confirming adoption barriers are structurally addressable through product and deployment improvements.

SAFETY GOVERNANCE IMPLICATIONS

- Priority 1 — Text omission monitoring:** The 0% AI under-reading rate is a surveillance artefact. Prospective text omission detection must be implemented as the highest-priority safety measure.
- Priority 2 — Integration testing:** Workflow and integration failures (48.5%) require systematic browser, EHR, and speech hardware interoperability testing pre-deployment.
- Priority 3 — Output explainability:** Communication failures (23.5%) should be addressed via structured report templates and confidence indicators to reduce clinician uncertainty.
- EU AI Act alignment:** All identified error categories are classifiable under EU AI Act Art. 9 risk management and Art. 72 PMS requirements for high-risk AI systems.

CONCLUSION

TANGO Gen AI exhibits a **structurally distinct error profile** from diagnostic imaging AI. Workflow and integration failures plus communication deficits account for **72% of simulated TANGO incidents** vs. 18% in the imaging AI benchmark. The **complete absence of AI under-reading reports** represents the highest-priority safety gap, requiring dedicated prospective text omission monitoring. Applying established diagnostic error taxonomies to Gen AI PMS data provides a scalable framework for standardised safety monitoring aligned with **EU AI Act post-market surveillance requirements**.