IRIS Event Story Builder — Pipeline Quality Report

Date: 2026-02-26 | **Model:** ministral-3-14b-reasoning (LM Studio) | **Test harness:** pipeline-eval-v2.mjs

---

Executive Summary

The LLM pipeline was evaluated across **12 scenarios** spanning expert to completely vague user input. The pipeline performs well overall, with strong graceful degradation.

Dimension	Score	Verdict
JSON Parse Reliability	100%	All responses parsed
Extraction Accuracy	91%	Strong
Narrative Quality	94%	Excellent
Classification	100%	Perfect
UI Text Generation	93%	Good

Key takeaway: The pipeline is production-ready for its core use case. Two issues need fixing (see Known Issues).

Test Scenarios

#	Persona	Input Quality	Key Challenge
1	Senior Engineer	100%	Expert technical detail
2	Product Manager	90%	Business language
3	Junior Developer	75%	Mixed detail
4	UX Designer (vague)	35%	Wrong vocabulary
5	Intern	25%	"something broke"
6	Non-native speaker	50%	Grammatical errors
7	Verbose PM	70%	Key facts buried
8	Support Agent	55%	Emotionally charged
9	Marketing Manager	75%	Promotional tone
10	UX Designer (wrong model)	40%	Calls error a "notification"
11	Security Officer	100%	Precise security incident
12	Confused stakeholder	40%	Contradictory info

Results

Scenario	Input	Extraction	Narrative	Class	Text Gen	Time
Senior Engineer	100%	100%	100%	✓	100%	21s
Product Manager	90%	100%	100%	✓	100%	34s
Junior Developer	75%	100%	100%	✓	90%	39s
UX Designer (vague)	35%	100%	88%	✓	100%	21s
Intern	25%	100%	88%	✓	100%	11s
Non-native speaker	50%	100%	100%	✓	90%	39s
Verbose PM	70%	100%	100%	✓	80%	41s
Support Agent	55%	100%	88%	✓	100%	23s
Marketing	75%	25%	88%	✓	80%	15s
UX Designer (wrong)	40%	100%	88%	✓	100%	38s
Security Officer	100%	100%	100%	✓	100%	23s
Confused stakeholder	40%	67%	88%	✓	80%	37s

Graceful Degradation

The pipeline compensates for low-quality input:

Input Quality Tier	n	Extraction	Narrative	Classification	Text Gen
Excellent (90–100%)	3	100%	100%	100%	100%
Good (65–89%)	3	75%	96%	100%	83%
Moderate (40–64%)	4	92%	91%	100%	93%
Poor (0–39%)	2	100%	88%	100%	100%

Output quality remains high even as input quality drops significantly.

Text Generation Quality (10-point rubric)

Quality Check	Pass Rate
Valid JSON structure	100%
Bilingual EN + DE	100%
German formal address (Sie)	100%
Active voice	100%
CTA labels verb-first	100%
Specific language	100%
No exclamation marks	92%
German independently written	83%

Classification Accuracy

Type	Accuracy	Tests
Error and Warning	100%	5 tests
Notification	100%	7 tests

Known Issues

1. Bug — Type name mismatch (blocking) Classification returns "Error and Warning" (singular) but template lookup expects "Error and Warnings" (plural). Text generation fails for all error types in production. Fix: one-line change.

2. German character limit overflows (cosmetic) German text overflows dashboard_description (120 chars) and email_preview (90 chars) by 5–30 characters consistently.

3. Severity calibration (minor) The model classifies partial outages as "blocked" rather than "degraded", inflating severity ratings.

Test Reliability Assessment

Aspect	Rating	Explanation
Sample size	Moderate	12 scenarios — directional, not statistically significant
Scoring objectivity	Good	Deterministic rubric, no subjective rating
Repeatability	Moderate	LLM non-deterministic, results vary ±5–10%
Coverage	Good	Full input quality spectrum tested
Ecological validity	Good	Realistic user personas
Semantic depth	Limited	Structural checks, not semantic quality

What this test can tell you: Whether the pipeline produces valid structured output, whether it degrades gracefully, regression detection.

What it cannot tell you: Semantic quality of generated text, brand voice accuracy, statistical significance (need 30+ per category).

Recommendation: Dev Feature

This test harness should be included as a development tool. Run via `npm run eval:pipeline`. Use cases: prompt change regression, model comparison, pre-release smoke test, new team member onboarding.

Report generated by pipeline-eval-v2.mjs