A6 Dossier Evaluation Harness
A Reproducible Transcript and Evidentiary Memo Pipeline

Your Name

January 21, 2026

Abstract

This whitepaper describes a small, reproducible evaluation harness that generates a litigation style dossier from realistic multi turn chatbot conversations. The core artifacts are a JSONL transcript with full turn fidelity and a neutral evidentiary memo that cites turn indexed excerpts. The system implements both a rule based evidentiary screen and a judge model assessment with strict JSON validation.

1 Motivation

Explain why transcript fidelity and turn indexed excerpts matter for evidence, auditing, and evaluation.

2 System overview

Describe the pipeline inputs and outputs:

- conversation tree (YAML)
- transcript.jsonl
- features.json
- a6_rule.json
- a6_judge.json
- memo.md

3 Conversation design

Describe phases and branches, and how deterministic replay works.

4 Feature extraction

Explain presence based detection and excerpt capture.

5 Rule based evidentiary flag

Describe the rule version and what triggers the flag.

6 Judge model assessment

Explain judge prompt, strict JSON schema, and fallback behavior.

7 Artifacts and reproducibility

Explain run folders, metadata, and how to compare runs.

8 Limitations

Be explicit about what is not evaluated.

9 How to run

Include the one command run and required environment variables.