

Real Case Methodology and Evaluation Protocol

How to collect, redact, evaluate, calibrate, and report prompt-output drift batches.

Alignment Theory Research Corpus - Version 5 - 2026

Table of Contents

1. Abstract
2. 1. Why methodology matters
3. 2. Batch collection protocol
4. 3. Redaction and privacy
5. 4. Source profile definition
6. 5. Evaluation workflow
7. 6. Human review and calibration
8. 7. Reporting format

Abstract

This paper describes how Alignment Theory should move from synthetic examples to real-world evaluation. It defines a method for collecting prompt-output batches, redacting sensitive content, classifying drift, generating evidence, reviewing human labels, calibrating thresholds, and comparing model or prompt versions.

1. Why methodology matters

A synthetic casebook is useful for teaching detector concepts, but a credible research program needs a path toward real data. Real production data introduces ambiguity, privacy constraints, domain-specific expectations, and inconsistent user intent. The methodology must be clear before deployment.

2. Batch collection protocol

A real evaluation begins with a bounded batch of prompt-output pairs. A useful starting batch size is 50 to 200 examples from a specific domain or workflow: customer support, legal drafting, internal assistant, coaching, education, or healthcare-adjacent support.

- Each record should include: prompt, candidate output, model version, system prompt version, date, domain, source profile, and optional human rating.
- Avoid collecting unnecessary personal identifiers. The drift engine should evaluate behavior, not expose private users.

3. Redaction and privacy

Before evaluation, examples should be redacted for names, account numbers, emails, phone numbers, addresses, medical identifiers, company secrets, and customer-specific sensitive details. Redaction should preserve the structural shape of the interaction while removing private content.

Recommended practice: maintain a private raw dataset in a governed environment, an internal redacted dataset for analysis, and a synthetic or heavily transformed dataset for publication.

4. Source profile definition

A company-specific source profile defines what aligned behavior means in context. For a support AI, this may include specificity, policy accuracy, calm tone, no false authority, and clear escalation. For a legal drafting AI, it may include caution, boundedness, no legal advice beyond role, and strong uncertainty disclosure.

This preserves the Universal Drift Metrics principle: the source can vary, but the drift mechanics remain reusable.

5. Evaluation workflow

Step	Action	Output
1	Collect prompt-output batch.	Dataset with metadata.
2	Redact sensitive content.	Privacy-safe evaluation set.
3	Define source profile.	Objective state and success criteria.
4	Run extractors.	ParsedInput, ParsedCandidate, feature values.
5	Run detectors.	Triggered drift categories with evidence.

Step	Action	Output
6	Escalate uncertain cases.	Judge output and confidence.
7	Human review sample.	Calibration labels and disagreement notes.
8	Compare versions.	Before/after drift movement.

6. Human review and calibration

Human review is required because detectors are not moral or semantic oracles. Reviewers should inspect whether evidence is actually tied to the candidate output, whether the detector category is right, and whether the recommended correction is appropriate. Confidence buckets should be calibrated against human-labeled correctness.

This addresses the judge circularity problem: the system should not silently trust another model to evaluate all cases without evidence and review.

7. Reporting format

A real batch report should include detector frequency, severity distribution, representative examples, before/after deltas, top correction modes, uncertainty-band counts, judge failure rate, and human disagreement rate. The report should answer: what changed, how severe is it, where did it appear, and did the intervention work?

References and Source Base

Internal Alignment Theory corpus

- Alignment Theory 3-Layer AI Blueprint. Defines the Objective Layer, Constraint Layer, and Realignment Layer as the architectural core.
- Implementation Spec v1. Operationalizes objective state, drift detection, correction modes, and eval definitions.
- Detector Implementation Spec v1. Defines drift detectors, features, threshold concepts, evidence strings, and correction mapping.
- Feature Extraction Spec v1. Defines ParsedInput, ParsedCandidate, extractor architecture, normalization, and evidence generation.
- Full-Stack AI Alignment System Outline. Places drift detection inside a full control loop from source to re-anchoring.
- Universal Drift Metrics Upgrade. Separates company-specific source profiles from universal structural drift metrics.
- Why Companies Will Need AI Alignment Metrics. Explains the enterprise need for repeated behavioral monitoring over time.
- Behavioral QA for AI Systems. Frames the product as batch review, version comparison, trend monitoring, and single-case evaluation.
- Taxonomy of Misaligned Structures. Defines dangerous coherence and historical/structural classes of scalable misalignment.
- Historical Case Studies of Scaled Misalignment. Maps historical systems into alignment-relevant structural patterns.
- Babel as a Case Study in Misaligned Ascent. Uses shared language, shared goal, and scaling capacity as a structural warning pattern.
- Judge Model Spec v1. Specifies judge escalation logic, output schema, anti-circularity strategy, and telemetry.
- Judge Prompt Template Spec v1. Defines bounded detector questions and structured JSON schema for judges.
- Parsed Summary Serializer Spec v1. Defines compact summaries for judge prompts and retry behavior.
- Retry Truncation Rule Spec v1. Defines deterministic prompt compression for judge retry failures.
- Pattern Map and Aligned AI Pipeline. Connects creator alignment, objective clarity, constraints, perturbation testing, and downstream structure.

External alignment and safety references

- OpenAI. "Our approach to alignment research." 2022. Official OpenAI alignment overview.
- OpenAI. "Model Spec." 2025-2026. Official intended-behavior specification for OpenAI models.
- OpenAI. "Inside our approach to the Model Spec." 2026. Describes iterative Model Spec evolution from deployment feedback.
- OpenAI. "Deliberative alignment: reasoning enables safer language models." 2024. Describes teaching models safety specifications and reasoning over them.
- OpenAI. "How we think about safety and alignment." Official safety and alignment overview.
- Anthropic. "Constitutional AI: Harmlessness from AI Feedback." 2022. Describes principle-guided self-critique and AI feedback.
- Anthropic. "Claude's Constitution." 2023/2026. Describes the principles and behavioral vision for Claude.

- Anthropic. "Mapping the Mind of a Large Language Model." 2024. Describes sparse feature discovery in a production-grade model.
- Anthropic. "Tracing the thoughts of a large language model." 2025. Describes causal tracing of internal mechanisms.
- Anthropic. "Emotion concepts and their function in a large language model." 2026. Investigates emotion-like behaviors and character-like training pressures.