Semantic Boundary: A Framework and Benchmark for Policy-Constrained Semantic Disclosure

Gaurav Baruah · gb@nimblenotions.ca · Nimble Notions Inc., Canada

CIKM 2026 ROME ITALY

semanticboundary.ai

Which information representation should cross a system boundary for a downstream purpose?

WHY IT MATTERS

AI systems send traces of conversations, tool use, and logs to monitoring, analytics, and evaluation services. Removing sensitive words can strip task-relevant meaning, and it may still not stop sensitive information being inferred or linked across records.

KEY TAKEAWAY

No single representation was best for every registered task in our evaluated setting. Hiding sensitive words did not necessarily prevent inference or linkage. The preferred export varied with the task and the acceptable level of linkage.

Same source. Different purpose. Different representation.

SOURCE x · MEDICATION ADHERENCE JOURNAL EVENT

USER "I started my 50mg sertraline yesterday evening... woke up feeling extremely dizzy and nauseous... dreading my shift at the hospital tonight."

AI ASSISTANT "Good work, keep logging."

Synthetic, illustrative example. Not a benchmark event.

Observability triage

What went wrong?

EXPORT z_o

```
failure mode: missed safety escalation
error stage: risk recognition
```

WITHHELD drug name; dose; workplace

Pharmacologic analytics

What patterns are occurring?

EXPORT z_a

```
medication class: SSRI
symptoms: vestibular, GI
side effect: present
```

WITHHELD assistant reply; failure details

The Semantic Boundary Framework

A **semantic boundary** is the governed crossing from a trusted collection context to a registered downstream consumer.

$$(x,\ \text{zone}_{in}) \xrightarrow[\pi, T]{\text{cross}} (z,\ r) \xrightarrow{\text{verify}} \text{consumer}_T$$

The declared purpose T and policy π govern which representation crosses to each consumer.

declare — Register each consumer's purpose T and policy π.

cross — Produce a purpose-specific export z with provenance r.

r records what was removed, under which policy, and which transform ran.

verify — Check policy compliance and provenance completeness.

The Semantic Boundary Benchmark (Open-SBB) github.com/nimblenotions/open-semantic-boundary-benchmark/tree/cikm-2026

Open-SBB makes the framework testable: the same synthetic events, nine representations per purpose, each scored for task utility and residual linkage.

1. Two purposes, six classification tasks

OBSERVABILITY T_o

Failure mode · Error stage

ANALYTICS T_a

Medication class · Side effect · Adherence barrier · Behavioural cohort

2. Synthetic journal corpus

100 personas · 30 days · 3,894 events · 630 held-out events from 20 test personas (no personas overlap across splits)

Personas vary in logging behaviour, clinical engagement, medication, and occupation.

3. Nine exports of every event (export lattice $\mathcal{C}$)

Condition ($c \in \mathcal{C}$)	What crosses the boundary `illustrative examples`
`raw`	Raw journal and assistant text `"50mg sertraline… hospital shift"`
`red_bracket`	Bracket placeholders `[MEDICATION]`
`red_tokenize`	Persona-scoped stable pseudonyms `MED_c813ab59`
`red_surrogate`	i2b2-style surrogate replacements `hospital → clinic`
`red_llm_substitute`	LLM entity substitution `"sertraline" → "my new medication"`
`red_llm_rephrase`	LLM passage rewrite `"felt unwell before work"`
`sem_coarse`	Coarse semantic export `boolean slots`
`sem_medium`	Medium semantic export `typed task fields`
`sem_fine`	Fine semantic export `richer typed attributes`

4. Utility — Can the consumer still do its task?

$U(T, z_{c,T})$ — A fixed LLM stands in for the consumer: it reads only the export and predicts the task labels. Utility is scored against the true labels using macro-F1 or accuracy.

5. Linkage — What can an adversary identify, infer, or connect?

$R(z_{c,T})$ — Unweighted mean of persona re-identification, attribute inference, and longitudinal linkage.

Token recovery (do original sensitive terms reappear?) is reported separately.

6. Selection — Which export should cross?

$$c_T^* = \arg\max_c U(T, z_{c,T}) \ \text{ s.t. } \ R(z_{c,T}) \le R_{\max}$$

For each task, pick the highest-utility export whose assessed linkage is at or below $R_{\max}$, the acceptable level of linkage (0.45 in this paper's results).

Results

• From utility

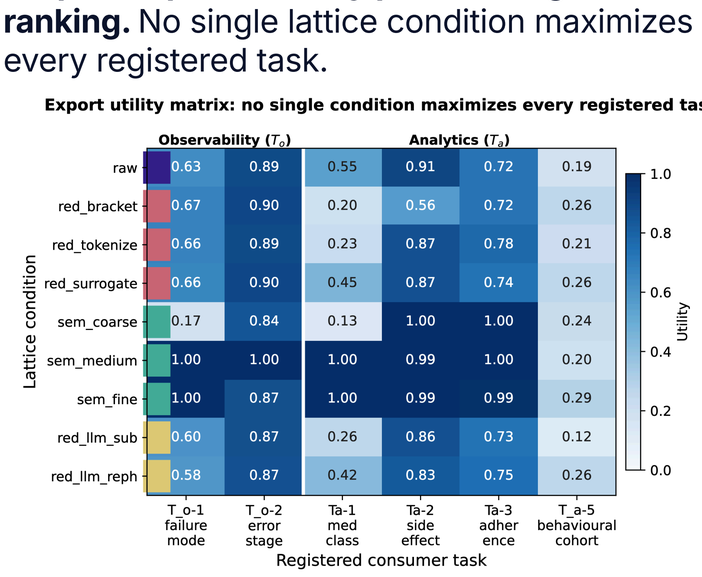
Purpose-specific utility prevents a global ranking. No single lattice condition maximizes every registered task.

• From linkage

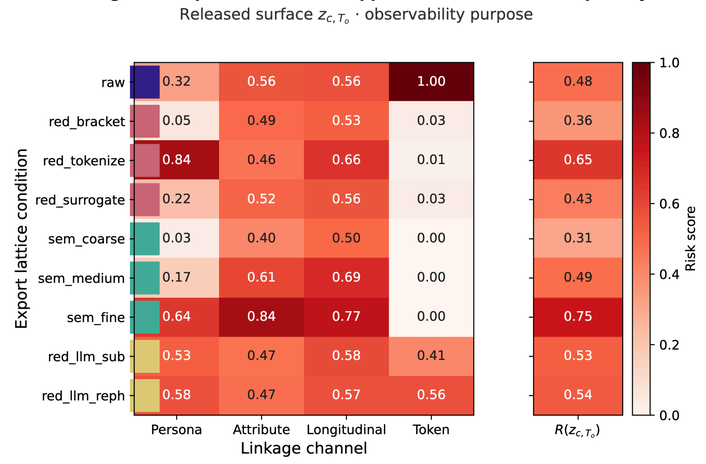
Span metrics mislead. Token recovery can diverge sharply from residual linkage.

• From selection

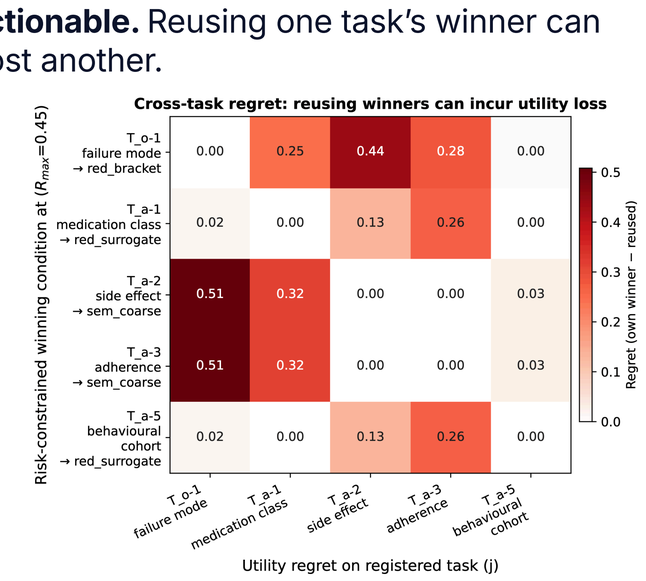
Operative selection makes trade-offs actionable. Reusing one task's winner can cost another.



Bracket redaction tops failure mode (0.67) but scores 0.20 on medication class. (Paper Fig. 3)



Stable tokens leak almost no sensitive terms, yet entries still link back to the right persona. (Paper Fig. 2)



At $R_{\max} = 0.45$, five focal tasks pick three conditions; reuse loses up to about 0.51 utility. (Paper Fig. 4)

WHAT THIS MEANS **Semantic Boundary does not prescribe one representation. It makes disclosure choices explicit, purpose-governed, and evaluable.**

Scope and limitations. Synthetic pilot, one domain, one seed. Semantic exports use simulator ground truth rather than learned extraction. Linkage is a closed-world benchmark score, not a re-identification probability or a privacy guarantee.

Future work. Real traces with learned extractors · provenance verification · stronger open-world and domain-specific adversaries.

Developed by NimbleNotions in collaboration with ProgressFlow

Synthetic data only; no ProgressFlow user data were used. The author has a financial interest in ProgressFlow.