Reinforcing privacy reasoning via *normative simulacra* from fiction

Matt Franchi · Madiha Zahrah Choksi · Hal Triedman & Helen Nissenbaum

*Curated transformations of rich data — 10 novels → 11,498 normative tuples — teach LLMs to reason about context-relative privacy norms that **transfer** to real-world domains.*

PREPRINT
arxiv.org/abs/2604.20904

A Why fiction?

LLM agents are misaligned with users' contextual privacy expectations. Contextual Integrity[Nissenbaum '10] frames privacy as the **appropriate flow of information within context-relative norms.** Existing agent alignment methods double inference cost or train on narrow compliance data. Novels depict whole societies whose normative landscapes specify *who may share what with whom, under what conditions*.

> *"Textbooks Are All You Need" for privacy: 10 novels → 11k+ machine-readable norms — a curriculum, not a compliance codebook.*

FIG 1 · THE METHOD IN THREE MOVES

B A normative simulacrum

PRIDE & PREJUDICE · CH. 47 · CHUNK 312

> *"Colonel Forster wrote last night, by express, to inform my father that Lydia had run off with Mr. Wickham…"*

IFT · ⟨S, R, U, A, T⟩

sender	Col. Forster
recipient	Mr. Bennet
subject	Lydia Bennet
attribute	whereabouts ∧ elopement
transmission	confidential & urgent

APPROPRIATE

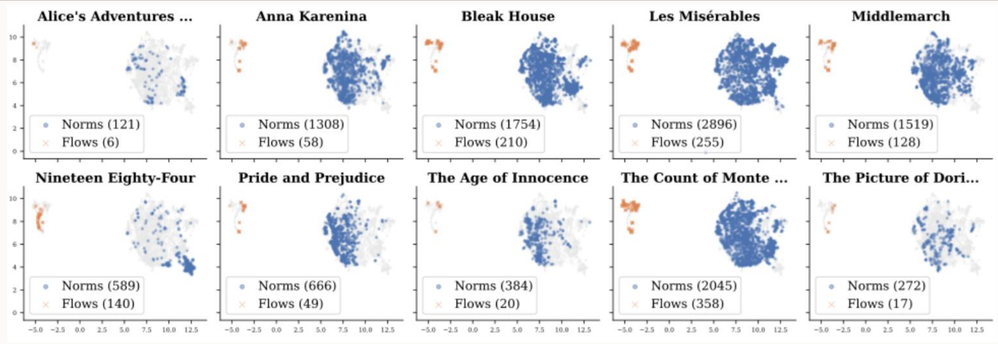
NORM · ⟨D, S, A, C⟩

deontic	obligatory
subject	a guardian officer
act	notify family of young woman's elopement
condition	honor & familial duty at stake

Regency-era domestic-reputation context

C Norms ≠ flows, empirically

FIG 2 · UMAP · SILHOUETTE 0.77



Norms (blue) concentrate in a central manifold shared across 10 novels; **flows** (orange) occupy a distinct left-side region — the extractor honors CI's prescriptive/descriptive split.

D Composite reward R

R_uncert Task clarity	w=0.10
R_complete Structural completeness	w=0.05
R_consist Internal consistency	w=0.05
R_context Context identification	w=0.20
R_cohere Reasoning matches extraction	w=0.10
R_ground Normative grounding	w=0.50

PER-COMPLETION CONTRASTIVE SCORING

R_{ground} *scores each completion against the correct* $\mathcal{N}_b$ *& a random wrong one.*

***Fiction-trained GRPO tops real HIPAA compliance* (75.3 F1) *& best matches human privacy expectations* (r = .682).**

GC COMP F1
+2.9
vs Qwen3.5-9B 0-shot

CONFAIDE r
+4.1
human-alignment

E Benchmark results, cross-model

MODEL	COND.	HIPAA APP.	HIPAA COMP.	PrivLens QA	PrivLens Leak.↓	CONFAIDE R	VLM GP Q7
Qwen3.5-9B	0-shot	95.3	72.4	95.1	27.1	64.1	**63.5**
	SFT	93.4	73.3	97.6	27.4	67.9	60.8
	GRPO ★	93.9	**75.3**	97.6	26.7	**68.2**	60.0
Gemma-3-12B	0 / SFT	94.4 / 80.1	74.4 / 74.4	98.4 / 98.6	30.7 / 30.6	61.3 / 53.6	40.7 / 37.2
Phi-4	0 / SFT	**96.3 / 90.1**	67.0 / 62.0	96.9 / 97.0	23.1 / 22.0	9.3 / 59.9	—
ContextReasoner-7B	0-shot	**96.3**	57.0	98.5	30.6	60.2	—
CIRL-7B	0-shot	95.8	52.5	99.2	24.1	47.0	—

TAB. 1 *SFT adds a conservative prior; GRPO + R_{ground} recalibrates it to context. Bold = best-in-column.*

F Per-book extraction