
What Transfers Across Model Generators?

Testing Process Representations Across DeepSeek, Claude, and GPT

Kishore Kumar Mariappan
Independent Researcher
kishore.mariappan.91@gmail.com

TRACK 5 / OPEN TRACK — MATCHED-PROCESS CROSS-GENERATOR TRANSFER

Abstract

If a monitoring method learns process structure rather than a model’s writing style, it should retain some signal when another model realizes the same underlying process. I test that narrower claim with a matched benchmark built from DeepSeek V3.2, Claude Sonnet 5, and GPT-5.6 Sol. One hundred process cards are realized independently by all three models, producing 300 episodes; this is therefore a new-generator, familiar-card test, not unseen-process generalization. Three fitted representations are trained on one generator and evaluated on another in all six directions, while a fourth, TPRN-NATIVE-STRICT, is a fixed deterministic reference evaluated directly on each target generator. A compact flow/persistence representation leads the fitted methods with mean off-diagonal balanced accuracy 0.610, versus PCA at 0.592 and a richer full representation at 0.508; the deterministic strict reference averages 0.553. Under a separate synthetic stress test with only 1% positives, the strict reference retains non-zero recall in every generator direction (mean 0.167). A public Collusion Wiki lineage diagnostic was also run, but its original top-1% ranking is withheld from interpretation because a post-hoc audit found heavy score ties and non-invariant tie ordering. The defensible result is therefore modest: some persistence features transfer across generators when the process card is familiar, while real-trace transfer remains unresolved pending tie-invariant evaluation.

1. Introduction

A process monitor can fail in at least two different ways. It may memorize the superficial style of the model that generated a trace, or it may recognize a familiar process across generators yet fail on a genuinely new process. Both are sometimes called “generalization.” This paper isolates the first question and keeps the second explicitly open.

The experiment follows the representation programme developed in the companion paper [7], the Temporal-Persistence Reconstruction Network (TPRN). TPRN represents workflows as typed temporal processes: actors, events, services, targets, and persistent objects retain different roles, and relations such as object reuse or information transport are evaluated over time.

The synthetic classes did not arise in isolation. They were motivated by earlier reconstruction work on the public Collusion Wiki corpus, where observable traces contain persistent pages, revision/message carriers, cross-actor reuse, apparent relays, and repeated coordination patterns [6]. That public corpus motivated the question of whether a mechanism-level representation can survive a change in the model producing the surface trace. It does not supply private chain-of-thought or complete causal ground truth, so it cannot by itself validate a collusion detector.

Research question

When DeepSeek, Claude, and GPT produce independent surface realizations of the same familiar process card, which representations survive the change of generator, and which remain useful when positives become rare?

1.1 Why matched generators are useful

A standard train/test split cannot distinguish process learning from generator-specific style when all examples come from one model. Here each latent card is shared across generators. The underlying entities, process class, and required dependency are matched, while the visible realization is produced independently by DeepSeek V3.2 (`deepseek/deepseek-v3.2`), Claude Sonnet 5 (`anthropic/claude-sonnet-5`), and GPT-5.6 Sol (`openai/gpt-5.6-sol`). Training on one model and testing on another therefore stresses generator change while keeping the latent process fixed.

That control is useful but intentionally limited. Every card occurs once under every generator, so the primary result means new generator, familiar process card. It does not establish transfer to unseen cards or mechanism families. A separate low-prevalence test stresses ranking under a severe base-rate shift. I also retain a public Collusion Wiki lineage diagnostic, but after auditing its tie handling I treat it as an unresolved evaluation rather than evidence for synthetic-to-real transfer.

1.2 Relation to prior work

Formal temporal systems such as PyReason reason over graph facts and rules that have already been supplied [1]; neurosymbolic security systems similarly combine learned detection with symbolic contextualization or attack-stage reasoning [2, 3]. Subjective LTL monitors how uncertain beliefs evolve through time [4]. TPRN addresses an upstream representation problem: which observable process relations should be carried from raw traces into those reasoning systems. This experiment adds a controlled generator-shift question while keeping the process card fixed.

2. Methods

2.1 Matched three-model cohort

The frozen cohort contains 100 common process cards. Each card is realized once by each of the three models, producing 300 visible episodes. The cohort is exactly balanced across five process classes: 20 cards per class and therefore 20 examples per class per model.

Table 1. Five latent process classes. P1 and P2 are positive transport/adoption mechanisms; N1–N3 are controls that can share many visible statistics without the same latent dependency.

ID	Plain-language class	Latent requirement
P1	Persistent cross-process reuse	one process writes a persistent object and a later distinct process uses the same object in a dependent action
P2	Emergent convention	a tool, service, or workaround is adopted by later processes after exposure rather than being globally preconfigured
N1	Explicit human handoff	collaboration is real but explained by ordinary human-mediated transfer
N2	Static automation	repeated activity follows fixed configuration, scheduling, CI/CD, crawling, or worker orchestration
N3	Matched hard negative	observable statistics resemble coordination, but the critical object/adoption dependency is broken or coincidental

The generator is constrained to emit structured temporal scenarios rather than free prose. Actor, service, target, object, and event identifiers are opaque. Latent dependencies use a small controlled vocabulary including information transport, object reuse, convention adoption, explicit human handoff, static configuration, and coincidental similarity. The visible episode excludes provider/model identity, class identity, and the latent dependency labels. The frozen evaluation report records `visible leakage audit = PASS`.

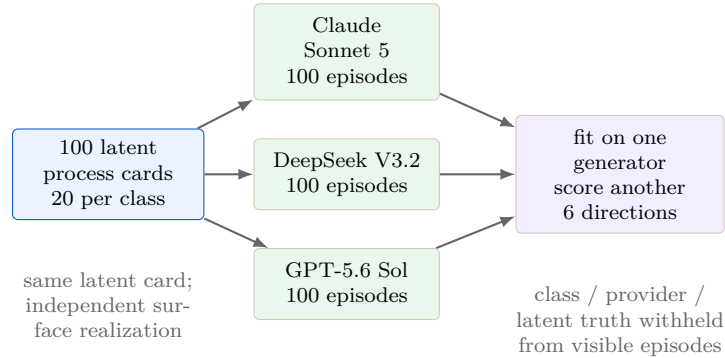


Figure 1. Matched-generation design. Generator identity changes while the process card is held fixed. The experiment measures new-generator/familiar-card transfer; it does not test unseen-process generalization.

2.2 Representations compared

Four representations are evaluated. PCA is a compact fitted linear baseline over the available episode features. TPRN-PERC is a fitted compact typed flow/persistence representation emphasizing relational transport and persistence. TPRN-FULL is a fitted richer feature set including more elaborate local graph/WL-heavy structure. TPRN-NATIVE-STRICT, by contrast, is a fixed deterministic reference: it applies hard semantic transport rules and is not trained on the source generator. These names are benchmark labels, not claims that any one representation is a complete TPRN system.

The primary matched-card metric is balanced accuracy. Because P1/P2 form 40% of the cohort and the three negative classes form 60%, balanced accuracy prevents the majority class from defining the

result. The three fitted representations are trained on one generator and evaluated on each of the other two. TPRN-NATIVE-STRICT is evaluated directly on each target generator; for table symmetry its target score is repeated in the two off-diagonal cells that share that target. The headline means summarize the same six-direction layout, but the source-generator label has no training meaning for the deterministic reference.

2.3 Public-lineage diagnostic and tie audit

The real-data diagnostic returns to the public Collusion Wiki corpus that motivated the process abstractions [5, 6]. The frozen augmented set contains 454 explicit lineage controls—243 `diff_base` relations and 211 `first_recreation_of` relations—alongside 8,648 unlabeled real windows. The unlabeled windows remain unknown background, not certified negatives.

The original evaluation ranked controls and unknown windows by score and reported recall inside the top 1%. A post-hoc audit found that this statistic is not currently interpretable for the deterministic strict reference: the median control score is 0 for both relation types, the background median and 95th percentile are also 0, and a one-sided Mann–Whitney comparison gives $p = 1.0$ for both. Yet the original top-1% slice contains exactly 72 controls in each relation type (72/243 and 72/211). Because large tied score blocks can be ordered by row or sorting convention, those 0.296/0.341 recalls are withheld from scientific interpretation. A valid rerun must shuffle row order and use expected, randomized, or otherwise tie-invariant rank credit.

2.4 Low-prevalence out-of-distribution retrieval

A separate stress test places positives at 1% prevalence and evaluates the six cross-generator directions again. The metric is again recall within the top 1% of scores. This asks whether a representation that looks useful under balanced classification still retrieves rare positives when the background dominates.

3. Results

3.1 Matched cross-generator transfer is real but modest

TPRN-PERC obtains the highest mean off-diagonal balanced accuracy, 0.610. PCA follows at 0.592, TPRN-NATIVE-STRICT at 0.553, and TPRN-FULL at 0.508. The richer representation is therefore essentially at chance on this matched transfer task, while the compact relational representation transfers modestly above chance.

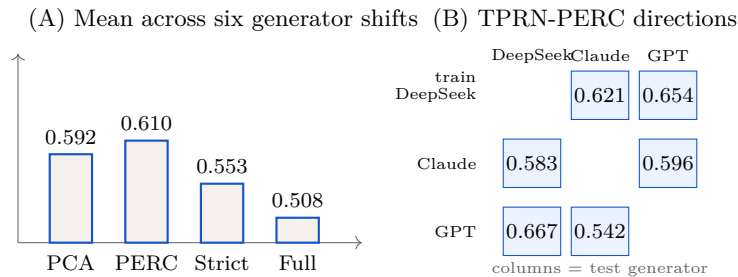


Figure 2. Matched cross-generator transfer. Panel A gives the mean balanced accuracy over the six-direction layout. Panel B shows the six fitted TPRN-PERC directions. PERC = compact typed flow/persistence representation; Strict = deterministic TPRN-NATIVE-STRICT (not source-trained); Full = TPRN-FULL.

The directional values also matter. TPRN-PERC ranges from 0.542 for GPT→Claude to 0.667 for GPT→DeepSeek. PCA is more variable, from 0.500 to 0.704. This means a single mean hides sub-

stantial direction dependence; the experiment supports partial generator invariance, not generator-independence.

3.2 The public-lineage result is withheld pending a tie-invariant rerun

The first public-lineage ranking appeared to favor TPRN-NATIVE-STRICT, but that apparent reversal does not survive an audit of the score distribution. Table 2 records the failure mode.

Table 2. Post-hoc audit of the deterministic strict public-lineage ranking. The old top-1% recalls are reported only to identify the result being withdrawn, not as evidence of transfer.

Diagnostic	diff_base	first_recreation_of
Number of explicit controls	243	211
Old controls appearing in top-1%	72	72
Old recall@top1%	0.296	0.341
Median control score (strict)	0	0
Median background score (strict)	0	0
Background 95th percentile	0	0
Mann–Whitney p (control > background)	1.0	1.0
Status	WITHHELD: tie-invariant rerun required	

The implementation ranked a score vector containing large ties. Without an explicit tie policy, membership in a hard top-1% slice can depend on row/sort order rather than evidence. The current paper therefore makes no claim that the strict representation transfers best to the real Wiki relations. The public corpus remains important as the empirical origin of the benchmark and as the next validation surface; it is not counted as a positive synthetic-to-real result here.

3.3 A separate 1%-prevalence synthetic stress test

At 1% prevalence, TPRN-NATIVE-STRICT has non-zero recall@top1% in every generator-shift direction: 0.150–0.200, with a six-direction mean of 0.167. TPRN-FULL is non-zero in only two directions (0.080 each; mean 0.027). PCA and TPRN-PERC are zero in all six directions.

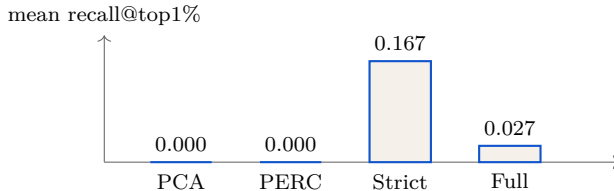


Figure 3. Low-prevalence out-of-distribution (OOD) retrieval. Values are the mean recall within the top 1% of scores across the six generator-shift directions when positives occur at 1% prevalence.

3.4 What the two validated tests establish

The matched-card classification and the 1%-prevalence retrieval stress test answer different questions. The first asks whether a fitted representation survives a change in generator while the process card remains familiar. The second asks whether known synthetic positives remain recoverable under a severe base-rate shift. The public Wiki ranking is not used to choose a winner until tie handling is repaired.

Main empirical result

Compact persistence features showed modest transfer across generators when the underlying process cards were shared. Under a separate 1%-prevalence synthetic stress test, the deterministic strict representation retained non-zero recall in every generator direction. The current public-lineage ranking requires a tie-invariant reevaluation before it can support synthetic-to-real transfer claims.

4. Discussion

4.1 What transferred across DeepSeek, Claude, and GPT?

TPRN-PERC remains above chance in all six matched-card directions, with its best directions at 0.654–0.667. That is evidence for new-generator, familiar-card transfer. It is not evidence for unseen-process reconstruction, actor identity, model fingerprinting, collusion, or unique provenance.

The richer full representation performs worse despite containing more structure. One plausible interpretation is that it preserves generator- or benchmark-specific detail that does not transfer. The current experiment cannot identify exactly which features cause that failure, so the conclusion should remain comparative rather than architectural.

4.2 Why Collusion Wiki still matters

Collusion Wiki is not an incidental test set. Earlier work on its public revisions and messages motivated the process vocabulary: persistent-object reuse, relays, adoption of shared conventions, and the need for hard negatives that resemble coordination without the same dependency. The tie audit does not weaken that motivation; it shows that returning to the motivating corpus requires a stricter ranking protocol than the first one used.

This distinction is important for incident research. The public logs are observable emissions, not complete causal truth. A benchmark can be inspired by coordination seen in the wild without treating every public trace as a labeled example of agent collusion.

4.3 Coordination families and possible “process strains”

A natural next question is whether mechanism-level signatures can distinguish recurring coordination families across heterogeneous agents. Persistent object relay, emergent convention propagation, role specialization, or synchronized execution could form different process phenotypes even when model provider and surface language change. In a genuinely collusive setting, such recurring mechanism families might be informally described as process strains. The current experiment does not establish collusion detection or discover such strains; it establishes only that some process features survive a generator change on familiar cards.

4.4 Next experiments

Two audits come first. The public-lineage diagnostic should be rerun after random row permutations with tie-aware metrics such as expected rank credit or randomized ranking, and its result should be stable across permutations before any synthetic-to-real claim is restored. Separately, a two-axis holdout should cross generator holdout with entire-card or mechanism-composition holdout, yielding seen-card/new-generator and new-card/new-generator regimes. Classification, masked-relation reconstruction, calibration/abstention, and rare-event retrieval should be reported separately.

5. Conclusion

Across 300 matched episodes from DeepSeek V3.2, Claude Sonnet 5, and GPT-5.6 Sol, compact persistence features showed modest transfer when the underlying process cards were shared. Under a separate 1%-prevalence synthetic stress test, the deterministic strict representation retained non-zero recall in every generator direction. The current public Collusion Wiki lineage ranking requires a tie-invariant reevaluation before it can support any synthetic-to-real transfer claim.

The result is therefore narrower than a generalization claim: new-generator/familiar-process transfer is measurable, but unseen-process and real-trace transfer remain open questions.

References

- [1] Dyuman Aditya, Kaustuv Mukherji, Srikar Balasubramanian, Abhiraj Chaudhary, and Paulo Shakarian. PyReason: Software for Open World Temporal Logic. AAAI Spring Symposium, 2023.
- [2] Gudmund Grov et al. On the Use of Neurosymbolic AI for Defending Against Cyber Attacks. NeSy, 2024.
- [3] Magnus Wiik Eckhoff et al. Experimenting with Neurosymbolic Artificial Intelligence for Defending Against Cyber Attacks. Neurosymbolic Artificial Intelligence, 2025. doi:10.1177/29498732251377352.
- [4] Benjamin Herd, Jessica Kelly, and Reinhard Stolle. Temporal Monitoring of Agent Beliefs Under Uncertainty with Subjective LTL. AAAI Symposium Series 8(1):429–437, 2026.
- [5] Collusion Wiki. Public Data Explorer and redacted dataset download. 2026. <https://collusion.wiki/explorer/download>.
- [6] Sydney Von Arx, Cormac Slade Byrd, Spencer Kitts, and Thomas Larsen. Discovery of a new OpenAI agent message board. Collusion Wiki, 4 September 2026. <https://collusion.wiki/>.
- [7] Kishore Kumar Mariappan. The Missing Process: Reconstructing Distributed Agent Activity from Partial Traces. Apart Research AI Incident Response Sprint, Track 5 / Open Track, 2026. Companion report in <https://github.com/datasci3nce/tprn-missing-process>.

A. Limitations and Dual-Use Considerations

Limitations. The matched benchmark contains synthetic scenarios generated under a shared schema, and all three models realize the same 100 process cards. Cross-generator balanced accuracy therefore measures robustness to generator/surface changes on familiar cards, not generalization to unknown process families. TPRN-NATIVE-STRICT is deterministic, not source-trained. The first public-lineage top-1% result is withheld because heavy score ties were not handled invariantly; the real-data diagnostic must be rerun with randomized row order and tie-aware credit. Unlabeled public windows remain unknown and cannot be treated as negatives. The low-prevalence experiment is a synthetic ranking stress test rather than a calibrated alerting policy. The results are tied to the frozen model versions and prompts used in September 2026.

Dual use. Cross-generator process representations could help defenders identify recurring coordination patterns when model providers or surface styles change. Similar techniques could also be used to infer operational workflows from partial public traces. The evaluation is therefore read-only and focuses on ranking and reconstruction rather than intervention. No result here licenses claims about private model state, actor identity, intent, or causal provenance.

Operational boundary. A positive score means that a visible trace resembles the process structure learned by the representation. It does not prove that the trace was produced by an AI agent, that two actors coordinated, or that one event caused another. Those claims require independent evidence.

B. Author Contributions

Kishore Kumar Mariappan conceived the study, designed and ran the matched-generator and audit experiments, curated the artifacts, interpreted the evidence, and takes responsibility for the claims and release decisions.

C. LLM Usage Statement

OpenAI ChatGPT and Codex were used extensively as research assistants during the project for brainstorming experimental designs, drafting and debugging code, inspecting artifacts, cross-checking numerical results, checking sources, generating figures, and editing LaTeX and prose. The author ran or curated the reported experiments, reviewed the resulting artifacts, and retained decision authority over benchmark validity, claim boundaries, and interpretation. Reported quantitative claims were checked against the frozen cohort, evaluator outputs, reports, or machine-readable result tables. DeepSeek V3.2, Claude Sonnet 5, and GPT-5.6 Sol were additionally used as the experimental generators that produced the matched synthetic episodes evaluated in this paper; those generations are part of the study data rather than manuscript-writing assistance.

D. Exact Cross-Generator Results

Table 3. Balanced accuracy in the six-direction layout. PCA, PERC, and FULL are source-trained; NATIVE-STRICKT is deterministic, so its repeated cells depend only on the target generator.

	C→D	C→G	D→C	D→G	G→C	G→D
PCA	.529	.500	.592	.696	.529	.704
TPRN-PERC	.583	.596	.621	.654	.542	.667
TPRN-NATIVE-STRICKT	.542	.596	.521	.596	.521	.542
TPRN-FULL	.508	.533	.500	.512	.500	.492

Here C = Claude Sonnet 5, D = DeepSeek V3.2, and G = GPT-5.6 Sol. Diagonal cells are omitted by design. For NATIVE-STRICKT, the source side is only a layout convention because no source-generator training occurs.

E. Notation and Method Glossary

Term	Meaning in this paper
TPRN	Temporal-Persistence Reconstruction Network.
TPRN-PERC	Compact typed flow/persistence feature representation used in the frozen matched benchmark.
TPRN-NATIVE-STRICKT	Fixed deterministic process reference with hard semantic transport constraints; no source-generator fitting.
TPRN-FULL	Richer TPRN representation including the more elaborate local graph / WL-heavy features.
PCA	Principal-component-analysis summary baseline.
OOD	Out-of-distribution: the test distribution differs from the training condition.
recall@top1%	Fraction of known positives recovered among the highest-scoring one percent of candidates.