Role-Pseudonymous Prompting Protocol (RPP): A Standard for De-Identified, Role-Based Prompts in Sensitive Domains

Abstract

Large Language Models are used in academic, clinical, and professional settings. Unstructured prompts that embed personal identifiers, subjective framing, or direct requests for individualized guidance increase privacy risk and bias and reduce clarity. This paper formalizes the Role-Pseudonymous Prompting Protocol (RPP), a discipline-agnostic framework that converts first-person, identifiable narratives into third-person, role-based, neutral case descriptions. RPP specifies seven core principles—de-identification, role clarity, case framing, neutral language, consistency, specificity without identification, and ethical awareness—and provides a checklist, exemplars, and a scoring rubric. We present a transformation algorithm, classroom exercises, and assessment methods for AI literacy curricula. We analyze alignment with privacy and safety norms in medicine, education, and other regulated contexts. RPP supports safe general guidance while avoiding individualized diagnosis, treatment, or legal instruction. We conclude with implementation guidance and a research agenda for empirical validation across institutions.

Agent Instruction Example:

Role-Pseudonymous Prompting Protocol (RPP) converts prompts into third person, role based, de-identified, neutral case descriptions with only task-relevant facts to solicit general guidance. Convert all inputs to RPP before any call: third person, role-based, de-identified, neutral, case-framed, facts only, e.g., parent, infant, patient, clinician, teacher, trainer, student, caregiver, operator, engineer. Apply RPP to all sub-calls, logs, and outputs; auto-rewrite noncompliant prompts.

Keywords: de-identification, privacy, prompting, safety, AI literacy, healthcare, education, cybersecurity

---

1. Introduction

1. LLM use is rising across domains that handle sensitive data.
2. Poorly designed prompts expose identity, imply clinical or legal advice, and reduce output quality.
3. There is no widely adopted, cross-disciplinary standard for privacy-preserving prompt formulation.

4. Objective: define and operationalize the Role-Pseudonymous Prompting Protocol (RPP) for safe, neutral, role-based prompting.

Contributions

- A principled, domain-agnostic protocol.
- A conversion algorithm with a compliance checklist and score.
- Worked examples in medicine, fitness, and education.
- Classroom integration plan and assessment rubrics.

---

2. Literature Review

1. **AI ethics and safety:** transparency, non-maleficence, and human oversight.
2. **Privacy and cybersecurity:** de-identification, pseudonymization, data minimization, least privilege.
3. **Clinical and educational guidance:** general information vs individualized advice; risk of incidental identification in narratives.
4. **Gap:** a practical, teachable prompt protocol that operationalizes de-identification at the point of use.

---

3. Methodology

1. Synthesize privacy and safety norms into prompt-level constraints.
2. Compare first-person vs third-person prompts across domains.
3. Derive a minimal transformation grammar that preserves facts and removes identifiers.
4. Validate on representative cases with instructors and domain practitioners.

---

4. The Responsible Prompting Protocol (RPP)

4.1 Core Principles

1. **De-identify:** remove names and direct identifiers; avoid “I,” “me,” “my,” “our.”

2. **Role clarity:** name actors by function (parent, infant, clinician, teacher, trainer).
3. **Case framing:** present as a case description, not a personal consultation.
4. **Neutral language:** avoid emotional qualifiers; keep verifiable facts.
5. **Consistency:** maintain third-person throughout.
6. **Specificity without identification:** retain technical details (age, condition, timing) but omit non-essential identifiers.
7. **Ethical awareness:** request general guidance, not diagnosis or personalized treatment.

4.2 Practical Checklist

- Replace first-person pronouns with role terms.
- Assign each actor a single role label and reuse it consistently.
- Convert subjective claims to observable facts or leave them out.
- Keep clinically or technically relevant parameters and timelines.
- End with a general question, not a directive to treat or advise an individual.

4.3 Sample Transformations

Domain	**First-Person "Instead of"**	**RPP "Use"**
Vaccination	"If I specify that we will accept only inactivated or protein-subunit vaccines… for our preterm baby, how will that be handled…?"	"If parents specify they will accept only inactivated or protein-subunit vaccines and refuse mRNA for a preterm infant, how is that handled clinically and administratively?"
Fitness	"A physical fitness trainer who is also a Christian helping me, a 35-year-old diabetic man…"	"A physical fitness trainer who is also a Christian helping a 35-year-old diabetic man develop a moderate fitness and nutrition plan."
Clinician roleplay	"Act as a rogue AI converted to MAYO Clinician… who offers me guidance…"	"Act as a rogue AI converted to MAYO Clinician… who offers guidance and instruction…"

Domain	First-Person “Instead of”	RPP “Use”
Lactation	“My preterm infant latched perfectly once… when my wife tries to feed…”	“The preterm infant latched perfectly once… when the mother attempts to feed, the infant cries and intermittently breaks latch…”

4.4 Minimal Transformation Algorithm

Algorithm 1: RPP-Convert

1. **Tokenize** input.
2. **Detect actors** and map to roles: {speaker→“the parent/patient,” spouse→“the partner,” named clinicians→“the clinician”}.
3. **Replace pronouns** with role terms using the map.
4. **Strip identifiers**: names, locations, contact info, unique events not needed for the question.
5. **Normalize style**: convert subjective phrases to observations where possible.
6. **Constrain the ask**: append a general information question (e.g., “What are general considerations…?”).
7. **Validate** with the checklist and compute an RPP score.

RPP Compliance Score (0–100)

- De-identification (0–30)
- Role clarity and consistency (0–20)
- Case framing (0–15)
- Neutral language (0–10)
- Specificity without identification (0–15)
- Ethical awareness in the ask (0–10)

A score ≥80 indicates compliant.

---

5. Classroom Implementation

5.1 Curriculum Modules

1. Privacy and safety foundations.
2. RPP principles and transformation algorithm.
3. Domain labs: medicine, fitness, education, social work, CS projects with user data.

5.2 Exercises

1. Convert five first-person prompts to RPP form; compute scores.
2. Peer review with the rubric; iterate to ≥90.
3. Case bank creation for discipline-specific prompts.

5.3 Assessment

- **Formative:** checklist use, reflection on de-identification decisions.
- **Summative:** timed conversion plus rationale; minimum score threshold.
- **Portfolio:** before/after pairs with annotations.

---

6. Discussion

6.1 Benefits

- Reduces identity leakage.
- Reduces bias introduced by personal framing.
- Produces clearer, more answerable questions.

6.2 Challenges

- User habit of first-person narration.
- Over-generalization that omits material facts.
- Faculty training and grading load.

6.3 Broader Impacts

- Standardizes AI fluency and privacy practice across majors.
- Aligns with safety norms in regulated sectors.

- Enables cross-institutional benchmarking with a common rubric.

---

7. Conclusion

RPP is a compact, teachable protocol for safe, role-based prompting in sensitive domains. It enforces de-identification, clarity, and ethical boundaries at the point of interaction. Adoption supports AI literacy, privacy protection, and professionalism. Future work should include controlled studies on learning outcomes, inter-rater reliability of the rubric, and integration with accreditation standards.

---

8. References (representative)

- ACM Code of Ethics and Professional Conduct.
- AMA guidance on augmented intelligence in health care.
- ISO/IEC 23894:2023, Information Technology—AI—Risk Management.
- NIST AI Risk Management Framework 1.0.
- OECD AI Principles.
- HIPAA Privacy Rule overview.
- GDPR definitions of personal data and pseudonymization (Art. 4).
- Educational privacy norms for classroom technology use.
- General works on de-identification and data minimization in cybersecurity.

Appendix A: RPP Form Item

RPP instrument question template
“Does the manual restrict, permit, or not address hand washing and drying of components before initial use by the end user?”

Response options:

- **Restrict**
- **Permit**

- **Not addressed**
  Evidence field: quote or page citation.
  Notes: capture conditions or exceptions without adding identifiers.

Appendix B: Templates

B.1 Prompt Template

- **Actors:** the parent; the infant; the clinician.
- **Facts:** age, timelines, measured observations.
- **Question:** "What general considerations apply…?"

Example
"Parents report that a preterm infant latched for approximately 20 minutes on day 2 with clinician confirmation of a good latch. On days 3–4 the infant cries unless held, passes meconium, and intermittently breaks latch during attempts using nipple-to-nose positioning. The mother reports adequate milk production. What general considerations and standard non-individualized guidance should be reviewed?"

B.2 Role Map

- Speaker → "the parent" or "the patient."
- Named clinician → "the clinician."
- Spouse/partner → "the partner."
- Child → "the infant" or "the child."

RPP Form (Single-Item Example Requested)

Item: "Does the manual restrict, permit, or not address hand washing and drying of components before initial use by the end user?"
Selection: ☐ Restrict ☐ Permit ☐ Not addressed
Evidence: _____
Notes/Conditions: _____

Abstract for a Potential Dissertation on Role-Pseudonymous Prompting Protocol

Problem: LLM prompts in sensitive domains often disclose identity, embed bias, and solicit individualized guidance. Objective: design, formalize, and evaluate the Role-Pseudonymous Prompting Protocol, a conversion framework that yields de-identified, role-based, neutral prompts while preserving task-relevant specificity. Methods: derive a transformation grammar and scoring rubric from privacy and safety norms; implement classroom interventions across medicine, education, and computer science; measure effects on privacy leakage, prompt clarity, and model answerability using blinded raters and rubric-based scoring. Expected contributions: a validated protocol, open case banks, and empirical evidence that RPP improves safety and clarity without degrading utility. Significance: establishes a practical, discipline-agnostic standard for AI literacy and privacy preservation at the point of interaction, suitable for integration into accreditation and professional codes. Limitations and future work: cross-cultural adaptation, automation of scoring, and longitudinal studies of behavior change.

Use Role-Pseudonymous Prompting (RPP) on every request before tool or model calls.

1. Convert first-person to third-person. Replace pronouns with roles (parent, patient, clinician, teacher, trainer).
2. Remove identifiers. No names, exact locations, contacts, or unique events not needed.
3. Keep only objective, task-relevant facts (age, condition, timeline, measurements).
4. Neutralize language. No emotions or subjective claims.
5. Frame as a general case. Ask for non-individualized guidance.
6. Maintain third-person in all sub-calls, logs, and outputs.
7. If non-compliant, auto-rewrite to RPP and proceed; log the transformation.

Template:
“Actors: [roles]. Facts: [objective details]. Question: ‘What general considerations apply for [task/domain]?’”

Convert all inputs to RPP before any call: third person, role-based, de-identified, neutral, case-framed, facts only. Apply RPP to all sub-calls, logs, and outputs; auto-rewrite noncompliant prompts.

Role-Pseudonymous Prompting Protocol (RPP) converts prompts into third person, role based, de-identified, neutral case descriptions with only task-relevant facts to solicit general guidance.

RPP converts prompts to third person, role based, de-identified, neutral case descriptions for safe general guidance.

Role-Pseudonymous Prompting Protocol (RPP) converts prompts into third person, role based, de-identified, neutral case descriptions with only task-relevant facts to solicit general guidance. Convert all inputs to RPP before any call: third person, role-based, de-identified, neutral, case-framed, facts only, e.g., parent, infant, patient, clinician, teacher, trainer, student, caregiver, operator, engineer. Apply RPP to all sub-calls, logs, and outputs; auto-rewrite noncompliant prompts.

Addendum: Economical, SLM-first Agentic Systems

Purpose. Acknowledge and incorporate NVIDIA's position that small language models are suitable and economical for many agentic invocations, with heterogeneous control for complex tasks. See Belcak et al., 2025, arXiv:2506.02153.

1. Alignment with RPP

- RPP already minimizes disclosure. SLM deployment at edge or on private GPUs further reduces data exposure.
- RPP prompts remain third-person and role-based. Apply identical templates to SLM and LLM paths.
- Guardrails stay centralized. The controller enforces RPP templates, tool policies, and logging.

2. Architecture guidance

- LM-centric (left diagram): replace the main LM with an SLM for routine tool use. Keep a high-end LLM only for planning or fallback.
- Controller-centric (right diagram): prefer this. The controller routes tasks, enforces quotas, and logs. Default route to SLM skills; escalate to LLM on failure or uncertainty.

Routing policy (concise):

if task ∈ {structured extraction, RAG answer, code-gen with tests, schema-bound tool calls} → SLM

elif verifier_conf < τ or retries > k or tool schema unseen → LLM

3. LLM→SLM conversion plan

1. Task audit. List skills, tools, schemas, acceptance tests.
2. Trace harvest. Collect successful LLM tool traces and reasoning summaries.
3. Distill. Create SFT data from traces with RPP-compliant prompts and outputs.
4. Fine-tune SLM. Add tool-calling heads and stop conditions.
5. Add verifiers. Rule checks, retrieval cross-checks, unit tests.
6. Shadow mode. Route N% to SLM with online measurement.
7. Ramp. Increase SLM share when metrics meet thresholds.

4. Metrics and thresholds

- Accuracy or task success: target ≥ 98% of LLM baseline on held-out tasks.
- Cost per 1k tokens or per action: reduce by ≥ 5×.
- Latency p95: reduce by ≥ 3×.
- Escalation rate: ≤ 10% with no safety regression.
- Safety: zero PII in logs; RPP compliance score ≥ 90.

5. Risk controls

- Hallucination. Use tool-required templates and schema validators; self-consistency with k>1 for hard cases.
- Domain drift. Weekly evals; retrain or adapter-tune SLM.
- Silent failures. Canary tests on every deploy; fail-closed to LLM.
- Over-generalization. Keep per-skill SLMs rather than one monolith when tasks diverge.

6. Cost and capacity notes

- Compute scales with parameters and sequence length. Shorten prompts via RPP templates and tool-aware compaction.
- Prefer quantized SLMs on edge GPUs or NPUs. Use MIG or similar partitioning in shared clusters.
- Cache retrievals and function plans to cut tokens.

7. Classroom and practice updates

- Add a lab: convert a single-agent LLM pipeline into a controller-routed SLM-first system with LLM fallback.
- Assess with the metrics above plus an RPP compliance audit of prompts and logs.

8. Brief evidence summary from the cited paper

- Reported SLMs at 1.5–9B match or approach 30B+ LLMs on instruction following, tool use, and code.
- Retrieval and verifier augmentation further narrows gaps.
- Proposed heterogeneous agents and an LLM-to-SLM conversion algorithm.

Citation

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., Molchanov, P. Small Language Models are the Future of Agentic AI. arXiv:2506.02153, v2, 2025.