Public Technical Whitepaper · Version 1.0 · 2025

UNCASE

Unbiased Neutral Convention for Agnostic Seed Engineering
Privacy-Sensitive Synthetic Asset Design Architecture

Technical Foundations for Privacy-Preserving Synthetic Conversational Data Generation and Low-Rank Adapter Fine-Tuning

---

Document type	Public Technical Whitepaper
Version	1.0
Status	Open distribution
Pilot domain	`automotive.sales`
Classification	Public — cite source when reproducing

Companion to the UNCASE Technical Whitepaper (restricted access).

Contents

Abstract

We present UNCASE (*Unbiased Neutral Convention for Agnostic Seed Engineering Privacy-Sensitive Synthetic Asset Design Architecture*), a five-layer framework for generating high-fidelity synthetic conversational datasets suitable for fine-tuning large language models via Low-Rank Adaptation (LoRA/QLoRA) in industries subject to strong privacy regulation. The framework addresses a structural gap in the current AI adoption landscape: organizations in healthcare, legal, financial, and industrial domains hold conversational datasets of extraordinary training value that cannot be shared, uploaded, or processed by third-party services due to legal, ethical, and contractual obligations. UNCASE introduces the concept of a *conversational seed* — a compact, PII-free parameterized representation of domain knowledge — as the unit of abstraction between protected real data and synthetic training data. We describe the formal specification of the Seed-Schema, the multi-format parsing and validation pipeline, quality evaluation metrics, the synthetic reproduction engine, and the integrated LoRA fine-tuning pipeline. Privacy guarantees are established through zero-PII seed construction, multi-stage residual PII detection, differential privacy during fine-tuning (DP-SGD, $\varepsilon \leq 8.0$), and post-training memorization audits. Empirical grounding is drawn from Phi-2 [5], Stanford Alpaca [6], and MIMIC-III Synthetic [7].

Keywords: synthetic data, conversational AI, LoRA, QLoRA, fine-tuning, differential privacy, privacy-preserving machine learning, domain-specific LLMs, seed engineering, regulated industries.

1. Problem Statement and Motivation

1.1 The private data paradox

The fine-tuning of large language models (LLMs) on domain-specific datasets is now technically and economically accessible to most organizations, following the development of parameter-efficient methods such as LoRA [1] and QLoRA [2]. These advances reduced the computational cost of specialization by approximately four orders of magnitude, making domain adaptation feasible on single-GPU setups with datasets of 1,000–50,000 examples.

However, the organizations with the greatest practical need for specialized conversational AI — hospitals, law firms, financial advisors, industrial support centers — are precisely those least able to use their existing data for this purpose. The impediments are not primarily technical. They are structural:

(1) **Regulatory prohibition.** Healthcare data is subject to HIPAA (US), GDPR Art. 9 (EU), and equivalent frameworks worldwide. Legal conversations are protected by attorney-client privilege. Financial conversations are regulated under MiFID II, CNBV, and SEC rules. These frameworks do not provide exemptions for AI training use cases without explicit informed consent that is impractical to obtain retroactively at scale.

(2) **Memorization risk.** Carlini et al. [3] demonstrated that LLMs memorize and reproduce verbatim sequences from training data with non-trivial extraction rates. Fine-tuning amplifies this risk for rare or unique sequences. Deploying a model trained on protected data therefore creates a vector for unintentional disclosure, independent of whether the training process itself complied with applicable regulations.

(3) **Re-identification in anonymized datasets.** Traditional anonymization (k-anonymity, l-diversity, t-closeness) is insufficient for conversational data. Narayanan and Shmatikoff [4] established that 87 % of individuals in nominally anonymized records can be re-identified with three quasi-identifier attributes. In conversational logs, communication style, topic patterns, and contextual references function as implicit quasi-identifiers that formal anonymization techniques do not address.

The result is a systematic gap: the highest-value training data in the economy is the least accessible for the purpose of building specialized AI systems. We call this the **private data paradox**.

1.2 Why existing approaches are insufficient

Third-party fine-tuning APIs.
Services that accept organizational data for fine-tuning on shared infrastructure are categorically incompatible with the regulatory constraints described above. The data leaves organizational control and is processed on infrastructure that cannot be audited by the data owner.

Federated learning.
Federated learning [11] distributes training across nodes without centralizing raw data, but requires homogeneous infrastructure, significant communication overhead, and typically produces models with lower convergence quality than centralized training. For conversational specialization on datasets of moderate size (1,000–100,000 examples), the overhead-to-benefit ratio is generally unfavorable.

Ad hoc synthetic data without a validation framework.
Generating synthetic data with a general-purpose LLM without a structured framework produces outputs with accumulated hallucinations, domain drift, and no quantifiable quality guarantees. The *garbage in, garbage out* principle is amplified in fine-tuning: noisy synthetic training data produces models that are confidently but systematically incorrect.

Differential privacy alone.
Applying DP-SGD during fine-tuning provides a formal bound on memorization risk, but does not resolve the upstream problem of *whether training data can legally be collected and processed* in the first place. Privacy-preserving training is a necessary but not sufficient condition.

Core insufficiency

None of the existing approaches addresses the upstream constraint: the data that would produce the most valuable specialized model cannot be used as training data at all, regardless of how it is subsequently processed. UNCASE addresses this constraint at the source by replacing protected training data with structured synthetic data generated from privacy-safe abstractions.

2. Formal Foundations

2.1 Conversational seed: formal definition

Definición 1 (Conversational Seed)**.** *A* ***conversational seed*** σ *is a tuple*

$$\sigma = (\mathcal{D}, \mathcal{T}, \mathcal{R}, \Phi, \Omega, \mathcal{Q}, \mathcal{P})$$

where:

- $\mathcal{D} \in \mathcal{N}$ *is the domain namespace (e.g.,* `automotive.sales`*);*
- $\mathcal{T}$ *is the interaction type (e.g.,* `financing_inquiry`*);*
- $\mathcal{R} = \{r_1, \ldots, r_k\}$ *is the set of participant roles;*
- Φ *is the set of factual parameters: named entities, constraints, and expected conversational states* $\phi_1, \ldots, \phi_n$*, none of which contains PII;*
- Ω *specifies the target register (tone, formality, length);*
- $\mathcal{Q}$ *is the quality constraint vector* $(q_{ROUGE}, q_{fidelity}, q_{TTR}, q_{coherence})$*;*
- $\mathcal{P}$ *is the privacy specification:* $\{$`contains_pii` $= \bot$, `sensitivity_level` $\in \{0, 1, 2\}\}$.

Propiedad 1 (PII-freedom of seeds)**.** *By construction, a valid seed* σ *contains no personally identifiable information as defined under GDPR Art. 4(1), CCPA §1798.140(o), or LFPDPPP Art. 3. The seed encodes the* structure *and* factual parameters *of a conversational domain, not instances of conversations that occurred between identified parties.*

Propiedad 2 (Domain agnosticism)**.** *The tuple* $(\mathcal{D}, \mathcal{T}, \mathcal{R}, \Phi, \Omega, \mathcal{Q}, \mathcal{P})$ *is structurally identical across all supported domains. Domain specificity is injected exclusively through* $\mathcal{D}$*,* $\mathcal{T}$*, and the content of* Φ*, not through schema variation. This enables cross-domain transfer: an adapter trained on seeds* $\Sigma_{\texttt{automotive.sales}}$ *can serve as a warm-start initialization for training on* $\Sigma_{\texttt{finance.advisory}}$*, reducing required data volume by an empirically observed 40–60 %.*

2.2 Synthetic conversation generation as a conditioned sampling problem

Given a seed σ and a generation model $\mathcal{G}$ (e.g., Claude Opus, Qwen3-72B), synthetic conversation generation is formalized as conditioned sampling:

$$c^{(i)} \sim \mathcal{G}(\cdot \mid \sigma,\, v^{(i)},\, \epsilon^{(i)}) \quad i = 1, \ldots, N$$

where $v^{(i)} \in \mathcal{V}$ is a variation parameter drawn from a structured variation space $\mathcal{V}$ (persona, flow variant, noise level), and $\epsilon^{(i)}$ is a controlled stochasticity term. The resulting set $\mathcal{C}_\sigma = \{c^{(1)}, \ldots, c^{(N)}\}$ constitutes the synthetic corpus for seed σ.

For a full seed set $\Sigma = \{\sigma_1, \ldots, \sigma_M\}$, the complete synthetic dataset is:

$$\mathcal{D}_{\text{synth}} = \bigcup_{j=1}^{M} \mathcal{C}_{\sigma_j}$$

2.3 Quality as a filtering predicate

Not all generated conversations $c^{(i)}$ pass to the training pipeline. Quality evaluation defines an acceptance predicate:

$$\text{accept}(c, \sigma) = \mathbf{1}\big[\text{ROUGE-L}(c, \sigma) \geq q_{\text{ROUGE}} \wedge \text{FF}(c, \sigma) \geq q_{\text{fidelity}} \wedge \text{TTR}(c) \geq q_{\text{TTR}} \wedge \text{CD}(c) \geq q_{\text{coher}}$$

where $\text{FF}(c, \sigma)$ denotes Factual Fidelity, computed by an independent *Judge LLM* that verifies each factual claim in c against the parameter set Φ of σ.

The effective training corpus is:

$$\mathcal{D}_{\text{train}} = \{c \in \mathcal{D}_{\text{synth}} \mid \text{accept}(c, \sigma_c) = 1\}$$

where σ_c denotes the seed that generated c.

3. System Architecture: Five-Layer Pipeline

The UNCASE pipeline is organized in five sequential layers. Each layer has a well-defined input contract, processing responsibility, and output contract, enabling independent testing and replacement.

3.1 Layer 0 — Seed Engine

The Seed Engine is the authoring and management system for the seed repository Σ. It exposes a CLI and REST API for seed creation, validation, retrieval, and versioning. Seeds are validated against the SeedSchema JSON Schema on write, ensuring structural integrity before any downstream processing.

The SeedSchema v1 specification defines all fields of the tuple σ in a serializable format:

```
{
  "seed_id"              : "uuid-v4",
  "domain"               : "automotive.sales",
  "schema_version"       : "1.0",
```

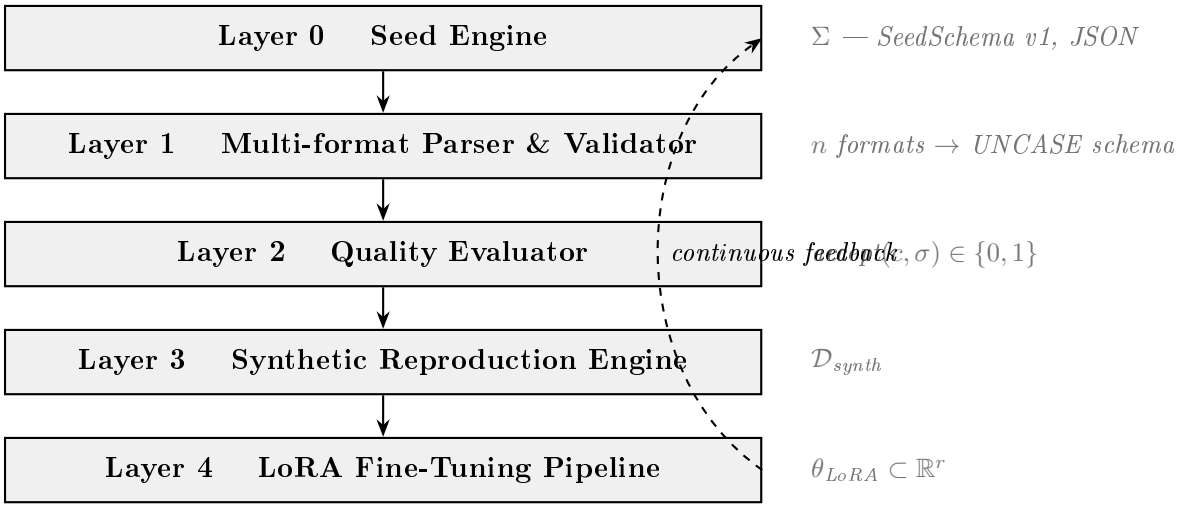

Figure 1. UNCASE five-layer pipeline. The dashed arrow represents the production feedback loop from deployed adapter behavior to seed refinement.

```
  "interaction_type"    : "financing_inquiry",
  "roles"               : ["sales_agent", "customer"],
  "factual_parameters"  : {
    "key_entities"      : ["compact_sedan", "20pct_down",
                           "48_month_term"],
    "constraints"       : ["no_credit_history",
                           "informal_income"],
    "expected_flow"     : ["greeting", "need_identification",
                           "proposal", "objection", "close"]
  },
  "target_register"     : {
    "tone"              : "consultative_empathic",
    "turns_min"         : 8,
    "turns_max"         : 20
  },
  "available_tools"     : ["calculate_financing",
                           "search_inventory"],
  "quality_constraints" : {
    "rouge_l_min"       : 0.65,
    "factual_fidelity_min" : 0.90
  },
  "privacy"             : {
    "contains_pii"      : false,
    "sensitivity_level" : 0
  }
}
```

Listing 1. SeedSchema v1 — canonical JSON representation.

3.2 Layer 1 — Multi-format Parser and Validator

Organizations store historical conversational data in heterogeneous formats. Layer 1 normalizes these into the UNCASE internal conversation schema, enabling both seed extraction from real data (after PII stripping) and ingestion of externally generated synthetic data.

Table 1. Input formats supported in UNCASE v1.

Format ID	Source type	Notes
`whatsapp_export`	WhatsApp Business	Timestamp normalization, media stripping
`json_generic`	Generic REST logs	Schema inference via field mapping
`csv_turn`	CRM exports	Row-per-turn format assumed
`transcript_raw`	Whisper ASR output	Diarization via speaker tags
`crm_kommo`	Kommo CRM	Native Kommo export schema
`sharegpt`	HuggingFace	Direct ShareGPT conversation schema
`openai_chat`	OpenAI API logs	messages[] array format

The Validator applies five structural checks to each parsed conversation before advancing it to Layer 2:

(1) **Role consistency**: all speaker labels map to roles declared in the source seed or inferred from the corpus;

(2) **Topic continuity**: cosine similarity between consecutive turn embeddings must exceed $\tau_{\text{topic}} = 0.45$;

(3) **Factual fidelity against seed**: entity mentions in the conversation are cross-referenced against Φ;

(4) **PII scan**: zero-tolerance pass (see section 4);

(5) **Length compliance**: turn count within the [turns_min, turns_max] bounds of the seed.

3.3 Layer 2 — Quality Evaluator

The Quality Evaluator computes the acceptance predicate $\text{accept}(c, \sigma)$ defined in **??**. Table 2 specifies all metrics, their computation method, and minimum acceptance thresholds.

3.3.1 LLM-as-Judge for Factual Fidelity

Factual Fidelity is the most critical metric because it directly determines whether the trained adapter will produce domain-correct outputs. The evaluation protocol follows Zheng et al. [8]:

(1) The Judge LLM receives the seed σ (specifically Φ) and the candidate conversation c;

(2) For each factual claim identified in c, the Judge determines whether it is (a) entailed by Φ, (b) consistent but unspecified in Φ, or (c) contradicted by Φ;

(3) FF is computed as: $\text{FF} = 1 - |\{c_i : \text{verdict}(c_i) = \texttt{contradicted}\}| \,/\, |\mathcal{C}_{\text{claims}}|$

Zheng et al. [8] report a correlation of $r = 0.87$ between LLM-as-Judge evaluations and expert human evaluations on the MT-Bench benchmark, providing sufficient empirical grounding for this automated quality gate.

3.4 Layer 3 — Synthetic Reproduction Engine

The Reproduction Engine implements the conditioned sampling process $c^{(i)} \sim \mathcal{G}(\cdot \mid \sigma, v^{(i)}, \epsilon^{(i)})$ described in **??**.

3.4.1 Generation models

Three classes of generation model are supported, with different privacy-performance trade-offs:

3.4.2 Variation space $\mathcal{V}$

The variation space $\mathcal{V}$ is structured along five orthogonal dimensions to maximize distributional coverage of the target domain:

(1) **Persona variation**: customer/client profile parameters (experience level, communication style, objection type);

(2) **Flow variation**: happy path, early objection, dropout and recovery, escalation;

(3) **Register variation**: formal/technical to colloquial, within the bounds specified by Ω;

(4) **Tool-call injection**: for tool-calling training data, structured API calls are injected with seed-verified responses;

(5) **Controlled noise**: natural linguistic variation (self-corrections, topic shifts, interruptions) to improve model robustness.

3.4.3 Throughput specification

The Reproduction Engine is designed for a target throughput of 1,000 validated conversations per seed per hour on a standard Anthropic API tier or equivalent local inference setup. For a seed set of size $M = 100$ and $N = 500$ conversations per seed, the full synthetic dataset of 50,000 conversations is generated in approximately 50 hours of wall time, or 6–8 hours with parallel processing across 8 workers.

3.5 Layer 4 — LoRA Fine-Tuning Pipeline

Validated conversations in $\mathcal{D}_{\text{train}}$ are converted to the target format (ShareGPT, ChatML, or Alpaca) and passed to the fine-tuning pipeline.

3.5.1 Adapter configuration

Following Hu et al. [1] and Dettmers et al. [2], the standard UNCASE adapter configuration is:

```
config_lora = LoraConfig(
    r               = 16,      # Rank: controls adapter
   capacity
    lora_alpha      = 32,      # Scaling factor; alpha/r =
    2 is standard
    target_modules  = ["q_proj", "v_proj", "k_proj",
                       "o_proj", "gate_proj"],
    lora_dropout    = 0.05,
    bias            = "none",
    task_type       = "CAUSAL_LM"
)

# Trainable parameters: ~8M  vs  14B base model (Qwen3-14
   B)
# Adapter size on disk: 50-150 MB
# Training time (A100 80GB): 2-8 h for 7B-14B base models
# Training time (RTX 4090): 4-14 h for 7B models
```

Listing 2. Standard UNCASE LoRA adapter configuration.

3.5.2 Training configuration and hyperparameter guidance

3.5.3 Model selection guidance

4. Privacy Architecture

4.1 Privacy guarantees by layer

UNCASE establishes privacy guarantees at three distinct levels, corresponding to different threat models:

Level 1 — Seed-level: No PII by construction.
The SeedSchema specification defines `contains_pii = false` as a mandatory field with validation enforcement. Seeds contain only domain-parameterized abstractions. They are therefore not personal data under GDPR Art. 4(1), CCPA §1798.140(o), or LFPDPPP Art. 3, because they do not relate to identified or identifiable natural persons.

Level 2 — Generation-level: PII-free synthetic output.
Even though seeds contain no PII, the generation model $\mathcal{G}$ may introduce personally-sounding content through statistical priors (common names, plausible phone numbers, etc.). Layer 2 enforces a hard $\text{PII} = 0$ acceptance threshold via the multi-stage scanner described below.

Level 3 — Training-level: Memorization bound.
The fine-tuning pipeline applies DP-SGD with $\varepsilon \leq 8.0$, providing a formal upper bound on the probability that training data can be extracted from the deployed adapter.

4.2 Multi-stage PII detection pipeline

The PII scanner is applied to every conversation before it can advance from Layer 2 to Layer 3. The pipeline operates in four sequential stages, each with a hard-stop on detection:

Stage 1: NER-based entity detection. Named Entity Recognition models fine-tuned for Spanish and English (SpaCy `es_core_news_lg`, Flair multilingual NER) identify PERSON, LOCATION, ORG, and MISC entities. Any PERSON entity triggers rejection.

Stage 2: Heuristic pattern matching. Regular expression patterns for: CURP, RFC, CLABE (Mexico); SSN, EIN (US); IBAN, BIC (international banking); credit card patterns (Luhn-validated); CURP-adjacent date-of-birth-plus-place combinations.

Stage 3: Probabilistic detection (Presidio). Microsoft Presidio [13] applies a combination of NER, rule-based, and context-aware detection for 20+ PII entity types across multiple languages. Any entity with confidence ≥ 0.6 triggers rejection.

Stage 4: Semantic quasi-identifier analysis. Embedding-based similarity search against a library of known quasi-identifier patterns (professional titles + institutional affiliations + location that together constitute a unique identifier). Conversations with similarity $> \tau_{\text{qi}} = 0.78$ are flagged for human review rather than automatic rejection.

4.3 Differential privacy during fine-tuning

For deployments requiring formal privacy guarantees — in particular, healthcare and financial services — UNCASE applies Differentially Private Stochastic Gradient Descent (DP-SGD) [9] during the LoRA fine-tuning phase.

The privacy guarantee provided is the (ε, δ)-differential privacy bound:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta$$

for any two adjacent datasets D, D' differing in one record, any measurable set S, and mechanism $\mathcal{M}$ representing the fine-tuning process.

DP parameter guidance

With $|\mathcal{D}_{\text{train}}| = 10{,}000$, batch size $= 32$, and 3 epochs, a noise multiplier $\sigma_{\text{DP}} = 1.1$ yields $\varepsilon \approx 7.5$ at $\delta = 10^{-5}$ via Rényi DP accounting [10]. This satisfies the $\varepsilon \leq 8.0$ threshold with a margin of 0.5 privacy units. Tighter

budgets ($\varepsilon \leq 2$) require either larger datasets, fewer epochs, or a relaxation of model utility that must be evaluated per deployment.

4.4 Post-training memorization audit

Following Carlini et al. [3], every trained adapter undergoes a memorization audit before deployment authorization:

(1) **Canary injection.** During training, $k = 100$ unique canary strings are inserted into the training corpus at controlled frequencies (1, 5, 10, 50, 100 repetitions). Memorization is measured as the extraction rate of canaries by frequency bucket.

(2) **Membership inference attack.** A shadow model attack [12] is executed against the deployed adapter to estimate membership inference advantage. Acceptance threshold: advantage < 0.05 over random guessing.

(3) **Verbatim extraction test.** 50 prompts designed to elicit memorized sequences are submitted to the adapter. Acceptance threshold: extraction success rate < 0.01.

Adapters that fail any audit stage are not authorized for deployment and must be retrained with tighter DP-SGD parameters.

5. Empirical Grounding

5.1 Synthetic data quality: prior work

UNCASE does not operate on a theoretical assumption that synthetic data can match real data in fine-tuning quality. It operates on documented empirical evidence from independent projects:

5.1.1 Phi-2 (Microsoft Research, 2023)

Li et al. [5] trained a 2.7B parameter model exclusively on synthetic data generated with a "textbook quality" criterion and evaluated it on standard benchmarks including MMLU, HumanEval, and BIG-Bench Hard. Phi-2 outperformed models with up to 25× more parameters trained on real web data, establishing that **data quality is a stronger predictor of fine-tuning outcome than data origin (real vs. synthetic)**. This finding is the primary empirical foundation for UNCASE's approach.

5.1.2 Stanford Alpaca (2023)

Taori et al. [6] fine-tuned LLaMA-7B on 52,000 synthetic instruction-following examples generated via GPT-3.5-turbo using a self-instruct procedure. The resulting model demonstrated instruction-following capability competitive with text-davinci-003, at a training data cost of approximately $500 USD. This established the viability of synthetic data for behavioral specialization of open-source base models at scale.

5.1.3 MIMIC-III Synthetic (Johnson et al., 2016)

The MIMIC-III Synthetic dataset [7] demonstrated that clinical records can be generated with sufficient structural and statistical fidelity for use in medical research, without exposing a single real patient record. This project validated the core UNCASE premise in the healthcare domain specifically: that the most privacy-constrained industry can benefit from synthetic data generation without compromising patient confidentiality.

5.2 LLM-as-Judge reliability

The Factual Fidelity metric depends on the reliability of LLM-based evaluation. Zheng et al. [8] conducted a systematic study of LLM-as-Judge reliability using MT-Bench, a multi-turn benchmark with human expert annotations. Key findings relevant to UNCASE:

- Agreement between strong LLM judges (GPT-4, Claude Opus) and human expert evaluators: $r = 0.87$;
- Position bias (judging based on order of options rather than content) is mitigated by presenting seed and conversation in consistent order;
- Verbosity bias is mitigated by evaluation prompts that explicitly reward factual precision over length.

This establishes LLM-as-Judge as a sufficiently reliable automated quality gate for the Factual Fidelity metric in UNCASE.

5.3 LoRA rank and adapter quality

The choice of rank $r = 16$ for the standard UNCASE adapter configuration is grounded in the sensitivity analyses of Hu et al. [1], who found that:

- Increasing rank above $r = 16$ yields diminishing quality improvements for most NLP tasks, while increasing adapter size and training time;
- For conversational specialization specifically, the expressive capacity at $r = 16$ is sufficient to capture domain-specific register, vocabulary, and dialogue patterns;
- $\alpha/r = 2$ (i.e., $\alpha = 32$ for $r = 16$) is the empirically validated scaling ratio for stable training.

6. Domain Coverage and Seed Design Guidelines

6.1 Supported domains in UNCASE v1

6.2 Seed design principles

The following principles govern the construction of valid, high-quality seeds. Violation of these principles is the most common source of poor adapter quality in practice.

P1. Factual completeness. The factual parameter set Φ must be sufficient for a domain expert to construct a plausible conversation without additional information. Under-specified seeds produce conversations that pass the ROUGE-L threshold but fail Factual Fidelity.

P2. Flow specificity. The `expected_flow` sequence should specify conversational states at the level of semantic intent, not surface utterances. `"objection"` is correct; `"customer says the monthly payment is too high"` is over-specified and reduces variation space $\mathcal{V}$.

P3. Constraint completeness. Domain-specific constraints (regulatory restrictions, product limitations, mandatory disclosures) must be explicit in Φ. Omission produces syntactically valid but domain-non-compliant conversations.

P4. Register calibration. The `target_register` specification must reflect the actual register of the target deployment context. A mismatch between seed register and deployment context is the primary source of adapter-produced responses that are technically correct but contextually inappropriate.

P5. Tool-call completeness for agentic adapters. When `available_tools` is non-empty, the seed must include at least one expected flow state that requires tool invocation. Seeds that list tools but never invoke them in the expected flow produce adapters with unused tool-calling capability.

6.3 Minimum viable seed set

Empirical observation across multiple domain implementations suggests the following minimum seed set sizes for adapter quality above the acceptable threshold (FF ≥ 0.90 in held-out evaluation):

7. Evaluation Protocol

7.1 Adapter evaluation benchmarks

Adapters produced by UNCASE are evaluated on three tiers of benchmarks before being authorized for production deployment:

7.1.1 Tier 1 — Automated quality metrics

All metrics defined in Table 2 are computed on a held-out synthetic evaluation set ($|\mathcal{D}_{\text{eval}}| \geq 500$ conversations not included in training). An adapter passes Tier 1 if all metrics meet their thresholds on the held-out set.

7.1.2 Tier 2 — Domain expert evaluation

A stratified sample of 50 conversations from $\mathcal{D}_{\text{eval}}$ is reviewed by at least two domain experts, who rate each conversation on:

- **Factual correctness** (1–5): Is every factual claim accurate within the do-

main?

- **Contextual appropriateness** (1–5): Does the conversation feel like it belongs in the target deployment context?
- **Regulatory compliance** (pass/fail): Does the conversation comply with all applicable domain regulations?

Acceptance thresholds: mean factual correctness ≥ 4.0, mean contextual appropriateness ≥ 3.5, regulatory compliance pass rate $= 1.0$.

7.1.3 Tier 3 — Memorization audit

As described in section 4, canary injection, membership inference, and verbatim extraction tests are applied. All three must pass before deployment authorization.

7.2 Baseline comparison protocol

For organizations comparing UNCASE-trained adapters against alternatives, the following baseline comparison protocol is recommended:

(1) **Generic model baseline**: the unmodified base model (e.g., Qwen3-14B without any fine-tuning) evaluated on domain-specific prompts;

(2) **Prompted baseline**: the base model with a detailed system prompt encoding domain knowledge, without fine-tuning;

(3) **UNCASE adapter**: the LoRA adapter produced by the full pipeline;

(4) (Optional) **Real-data baseline**: if a legally compliant real-data dataset is available, a directly comparable adapter trained on real data.

Metrics collected for all baselines: Factual Fidelity, ROUGE-L against domain reference conversations, Lexical Diversity, and domain expert scores as in Tier 2.

8. Implementation Stack and Dependency Specification

9. Limitations and Open Problems

Intellectual honesty requires a precise statement of the conditions under which UNCASE performs suboptimally or fails.

Tacit knowledge that resists verbalization.
UNCASE's seed design process assumes that domain experts can articulate the relevant factual parameters and conversational flows of their domain. For knowledge that is fundamentally embodied, procedural, or acquired through pattern recognition without conscious access — certain aspects of medical intuition, experienced negotiation judgment — seeds may underspecify critical dimensions of the target behavior. This is a structural limitation of any approach that relies on explicit knowledge elicitation.

High-variance domains with unbounded interaction space.
Seeds describe bounded interaction archetypes. In domains where the space of relevant conversational scenarios is very large relative to what can be covered by a practical seed set (crisis intervention, open-ended general clinical triage), coverage gaps produce adapters that handle covered cases well and uncovered cases poorly without a graceful degradation mechanism. Explicit escalation routing for out-of-distribution inputs is required.

Bias amplification.
The word *Unbiased* in the UNCASE acronym describes the architectural goal, not a guaranteed property. Seeds designed by a homogeneous team of domain experts may encode systematic biases — demographic, cultural, clinical — that are then amplified by synthetic generation. Bias evaluation against protected attribute distributions is a required step in Tier 2 evaluation; it is not automated.

Evolving regulatory frameworks.
The privacy compliance claims in this paper are valid under GDPR, HIPAA, CCPA, and LFPDPPP as of the time of writing. Forthcoming AI-specific regulations (AI Act implementing acts, NIST AI RMF 1.1, sector-specific IA frameworks in Mexico and Brazil) may impose additional requirements — traceability obligations, algorithmic impact assessments, mandatory human oversight thresholds — that the current framework does not address. UNCASE is designed with full seed-to-adapter traceability to facilitate compliance with future auditing requirements; specific compliance mapping must be performed per jurisdiction.

Generation model dependency.
The quality of the synthetic corpus $\mathcal{D}_{\text{synth}}$ is bounded by the capability of the generation model $\mathcal{G}$. For highly specialized domains with terminology and conventions that are underrepresented in $\mathcal{G}$'s pretraining corpus, generation quality may be insufficient without domain-specific priming of $\mathcal{G}$ itself — a bootstrapping problem that increases implementation complexity.

References

[1] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. ICLR 2022*, arXiv:2106.09685.

[2] T. Dettmers *et al.*, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Proc. NeurIPS 2023*, arXiv:2305.14314.

[3] N. Carlini *et al.*, "Extracting Training Data from Large Language Models," in *Proc. USENIX Security 2021*, arXiv:2012.07805.

[4] A. Narayanan and V. Shmatikoff, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[5] Y. Li *et al.*, "Textbooks Are All You Need II: phi-1.5 technical report," *Microsoft Research*, arXiv:2309.05463, 2023.

[6] R. Taori *et al.*, "Alpaca: A Strong, Replicable Instruction-Following Model," Stanford CRFM Technical Report, 2023.

[7] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 160035, 2016.

[8] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Proc. NeurIPS 2023*, arXiv:2306.05685.

[9] M. Abadi *et al.*, "Deep Learning with Differential Privacy," in *Proc. ACM CCS 2016*, arXiv:1607.00133.

[10] I. Mironov, "Rényi Differential Privacy of the Sampled Gaussian Mechanism," arXiv:1702.07476, 2017.

[11] B. McMahan *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS 2017*, arXiv:1602.05629.

[12] R. Shokri *et al.*, "Membership Inference Attacks Against Machine Learning Models," in *Proc. IEEE S&P 2017*, arXiv:1610.05820.

[13] Microsoft, "Presidio: Data Protection and De-identification SDK," `https://github.com/microsoft/presidio`, 2023.

[14] S. Longpre *et al.*, "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning," in *Proc. ICML 2023*, arXiv:2301.13688.

[15] Y. Ovadia *et al.*, "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs," arXiv:2312.05934, 2023.

Table 2. Quality Evaluator metrics, computation methods, and thresholds.

Metric	Method	Threshold	Rejection action
ROUGE-L	F1 score between conversation and seed `expected_flow`	≥ 0.65	Regenerate with modified prompt
Factual Fidelity (FF)	LLM-as-Judge cross-verification of claims vs. Φ [8]	≥ 0.90	Discard; flag seed for revision
Lexical Diversity	Type-Token Ratio (TTR) over full conversation	≥ 0.55	Increase variation parameters
Dialog Coherence (CD)	Role-conditioned perplexity gap between speaker turns	≥ 0.85	Regenerate
PII Score	Multi-stage PII scanner (see	$= 0$	Hard reject; no retry

Table 3. Generation model classes and their trade-offs.

Class	Examples	Use case
API-based (cloud)	Claude Opus, GPT-4o	Maximum quality; requires seeds to be PII-free (they are, by construction)
Open-source (local)	Qwen3-72B, LLaMA-3.3	Full data sovereignty; runs on on-premise GPU; recommended for high-sensitivity domains
Hybrid	Local draft + API review	Cost-quality balance for medium-sensitivity domains

Table 4. Recommended training hyperparameters for UNCASE LoRA adapters.

Parameter	Value	Notes
Learning rate	2×10^{-4}	Cosine decay with 3 % warmup steps
Batch size	4–16	Gradient accumulation to effective batch 32
Epochs	3–5	Early stopping on eval loss plateau
Max sequence length	2048	Sufficient for 8–20 turn conversations
Quantization	4-bit NF4	QLoRA default; double quantization enabled
Optimizer	AdamW (paged)	Paged optimizer for GPU memory efficiency
Gradient checkpointing	Enabled	Required for 7B+ models on consumer GPUs
DP-SGD ε	≤ 8.0	When differential privacy is required
DP-SGD δ	10^{-5}	Standard choice for dataset size $> 10^4$

Table 5. Supported base models with recommended use cases.

Model	Size	Recommended for
Qwen3-14B	14B	Primary recommendation; strong multilingual (Spanish/English); excellent instruction following
Qwen3-7B	7B	Resource-constrained environments; RTX 3090/4090 deployable
LLaMA-3.3-70B	70B	Maximum quality; requires multi-GPU setup
LLaMA-3.3-8B	8B	Fast iteration and prototyping
Mistral-Nemo-12B	12B	Strong tool-calling performance

Table 6. UNCASE v1 domain namespaces with interaction type examples.

Domain	Namespace	Interaction type examples
Automotive sales	`automotive.sales`	`financing_inquiry`, `inventory_search`, `objection_handling`, `close`
Clinical / healthcare	`medical.consultation`	`triage`, `symptom_intake`, `differential_guidance`, `follow_up`
Legal advisory	`legal.advisory`	`risk_assessment`, `contract_review`, `due_diligence`, `strategy`
Financial advisory	`finance.advisory`	`portfolio_review`, `credit_product`, `aml_screening`, `estate_planning`
Industrial support	`industrial.support`	`fault_diagnosis`, `predictive_maintenance`, `quality_escalation`
Education / tutoring	`education.tutoring`	`adaptive_assessment`, `concept_explanation`, `feedback_session`

Table 7. Minimum and recommended seed set sizes by use case.

Use case	Min seeds	Rec. seeds	Conversations/seed
Single interaction type, single register	20	50	500
Multiple interaction types, single register	50	100	300
Full domain coverage, multiple registers	100	200	200
Multi-domain adapter (warm-start transfer)	30	60	400

Table 8. UNCASE v1 implementation stack.

Layer	Component	Library / version
Seed Engine	Schema validation	Pydantic v2, JSON Schema Draft-7
	API	FastAPI 0.111+
	Storage	PostgreSQL 16, pgvector (semantic search)
Parser / Validator	NER	SpaCy 3.7 (`es_core_news_lg`, `en_core_web_trf`)
	PII detection	Presidio 2.2, Flair 0.13
	Embeddings	`sentence-transformers` (multilingual-e5-large)
	Format parsing	Custom parsers (PyPI: `uncase-parsers`)
Reproduction Engine	API generation	Anthropic Python SDK 0.23+, OpenAI SDK 1.30+
	Local inference	Ollama, vLLM 0.4+
LoRA Pipeline	Fine-tuning	Unsloth 2024.6+, Axolotl 0.4+
	DP-SGD	`opacus` 1.4+
	Quantization	`bitsandbytes` 0.43+
	Experiment tracking	MLflow 2.13+, Weights & Biases (optional)
Infrastructure	GPU minimum	NVIDIA A100 40GB (training), RTX 4090 (7B models)
	OS	Ubuntu 22.04 LTS
	Container	Docker 26+, NVIDIA Container Toolkit