Prism: Cross-Domain Personal Data Integration on Consumer Hardware Produces Emergent Insights

AtomGradient
https://github.com/AtomGradient/Prism
https://atomgradient.github.io/Prism

March 2026

Abstract

Current cloud-based AI assistants operate under a fundamental structural contradiction: to be genuinely useful, a model must understand its user's life context, yet the private data needed for such understanding is precisely what users are unwilling to upload to remote servers. We present Prism, a three-tier on-device architecture that resolves this privacy–utility paradox by integrating data from four vertical personal-data applications—finance, diet, mood, and reading—entirely on consumer Apple Silicon hardware. Through an ablation study on 10 synthetic users over 90 simulated days, we demonstrate that full cross-domain data integration achieves an average Insight Increment Ratio (IIR) of 1.48× over single-domain baselines, with the cross-domain scoring dimension driving nearly all of the gain. A model-scale study across four Qwen3.5 checkpoints (0.8B to 35B-A3B) reveals diminishing returns: the 2B→9B jump yields +14.9 points while 9B→35B yields only +5.6. Our federation protocol transmits only data summaries, achieving a 125.5× compression ratio with zero raw-data leakage. All experiments run on commodity hardware (M2 Ultra, M1 Max, M2 Pro), establishing a practical blueprint for privacy-preserving personal AI.

Keywords: personal AI, cross-domain data integration, on-device inference, federated data protocol, Apple Silicon, privacy-preserving AI

1 Introduction

1.1 The Privacy–Utility Paradox

The promise of a truly personal AI assistant rests on a circular dependency that current cloud-centric architectures cannot resolve: *to make AI genuinely useful, it must understand its user; to understand a user, it needs their most private data; yet the more private the data, the less willing users are to upload it to a third-party server.*

Today's flagship models—GPT-5, Claude Opus, Gemini 3.5—are powerful general-purpose reasoners, but they remain *knowledgeable strangers*. They can draft emails, generate images, and answer trivia, yet they do not know whether their user slept well last night, is under financial stress, or has been skipping meals for a week. The data that would enable such understanding—journals, medical records, financial transactions, dietary logs, emotional states—is exactly what users refuse to entrust to remote servers.

This is not merely a technical obstacle; it is a *structural* one. Cloud AI providers derive revenue from API calls (per-token billing); data remaining on-device eliminates their primary business incentive. Their competitive moat is built on data accumulation; local-only data erodes that moat. Even platform-native players face barriers: Apple's app ecosystem fragments third-party data into isolated silos, and Google's advertising model creates a rational distrust of deeper data access.

1.2 A Paradigm Inversion

We observe that the dominant paradigm can be characterized as **large model + small data**: a cloud-hosted model with hundreds of billions of parameters operates on a thin sliver of user context squeezed into a few thousand tokens of conversation history. PRISM inverts this paradigm to **medium model + rich data**: a locally hosted model of modest scale (3–35 billion parameters) operates on the user's entire longitudinal life data across multiple domains.

The core insight is that intelligence does not arise solely from parameter count; it also arises from *data depth and authenticity*. A 9-billion parameter model that has access to 90 days of a user's financial transactions, dietary patterns, emotional states, and reading habits can produce insights that a 1-trillion parameter model operating on a single conversation turn cannot—because the latter simply does not have the data.

1.3 Contributions

This paper makes the following contributions:

1. **Cross-domain insight emergence.** We demonstrate through a controlled ablation study (10 users × 8 data configurations) that integrating four personal-data domains produces an average IIR of 1.48× over single-domain analysis, with the cross-domain scoring dimension accounting for nearly all of the gain (Section 5.1).
2. **Model-scale analysis.** We establish a model-size vs. insight-quality curve on commodity hardware, showing that the practical sweet spot lies at the 9B parameter scale, with diminishing returns beyond that point (Section 5.2).
3. **Privacy-preserving federation.** We implement and evaluate a LAN-based federation protocol that achieves 125.5× data compression with zero raw-data leakage, demonstrating that multi-device personal AI can operate without any data leaving the user's physical premises (Section 5.4).
4. **Practical system design.** We provide a fully reproducible three-tier architecture running entirely on consumer Apple Silicon hardware, with all code and synthetic data publicly available.[1]

2 Related Work

2.1 Personalized Large Language Models

The challenge of adapting LLMs to individual users has received growing attention. Chen et al. (2025) provide a comprehensive survey identifying three axes of personalization: user profile injection, few-shot behavioral conditioning, and parametric adaptation. Most approaches assume access to a relatively small and static user profile.

Chuang et al. (2024) propose PLUM, a system that learns to "remember" user conversations by compressing dialogue history into persistent user embeddings. While effective for conversational style adaptation, PLUM does not address cross-domain data integration or the privacy implications of storing user embeddings on cloud infrastructure.

PersonalLLM (Li et al., 2025) introduces a benchmark and training methodology for personalized generation, demonstrating that even modest amounts of user-specific data can significantly improve output alignment. However, the evaluation is confined to text generation quality within a single domain and does not consider the compounding effects of multi-domain data fusion.

[1] https://github.com/AtomGradient/Prism

2.2 Continual and Lifelong Learning

The temporal dynamics of user preferences—what the literature terms *temporal drift*—pose a distinct challenge. Wang et al. (2024) survey continual learning methods for LLMs, highlighting catastrophic forgetting as the primary obstacle to lifelong adaptation. Current approaches (elastic weight consolidation, progressive neural networks, replay buffers) focus on model-side solutions and do not consider the data architecture that feeds the model.

LIBER (Huang et al., 2024) addresses lifelong user behavior modeling by proposing a framework that compresses long behavioral sequences into structured representations. LIBER's insight—that raw behavioral sequences are too long for direct LLM consumption and require intermediate structuring—directly informs our federation protocol design. However, LIBER operates in a cloud-based recommendation setting and does not address on-device deployment or cross-domain integration.

2.3 Federated Learning and On-Device Inference

Federated learning (McMahan et al., 2017) was designed to train models across distributed devices without centralizing raw data. However, classical federated learning focuses on *model training* (gradient aggregation), whereas our setting requires *data-level integration* for inference-time reasoning. Our federation protocol is closer to federated analytics (Ramage & Mazzocchi, 2020) but operates at the semantic summary level rather than the statistical aggregate level.

On-device LLM inference has become practical with the advent of Apple Silicon's unified memory architecture. The MLX framework (Hannun et al., 2023) and llama.cpp (Gerganov, 2023) have demonstrated that models up to 70B parameters (with quantization) can run at interactive speeds on consumer hardware. Our work builds on this capability but focuses on *what to do* with on-device inference once it is technically feasible, rather than on the inference engine itself.

3 System Design

3.1 Design Principles

PRISM is built on three design principles:

1. **Privacy by architecture, not by policy.** Raw data never leaves the device that generated it. This is enforced structurally, not by terms of service.
2. **Data collection as a natural byproduct.** Users interact with genuinely useful vertical applications; data accumulation is a side effect of normal usage, not a separate act of "feeding the AI."
3. **Heterogeneous device awareness.** The system adapts its inference strategy to the computational tier of the device, from sub-1B models on mobile to 35B MoE models on a home server.

3.2 Three-Tier Device Topology

The system organizes user devices into three computational tiers:

Tier 0 — Mobile Devices (Data Collection + Lightweight Inference). Smartphones and tablets serve as the primary data collection points. In our experimental setup, an iPhone 15 Pro Max (8 GB) runs Mealens (dietary tracking via photo recognition) and Ururu (mood journaling and emotional state tracking), while an iPad Air M3 (8 GB) runs Narrus (reading analysis) and Dailyn (financial accounting). These devices can run sub-2B models for immediate local queries but delegate complex cross-domain reasoning to higher tiers.

Tier 1 — Mid-Tier Local Inference (Daily Queries). Machines with 32 GB unified memory (Apple M1 Max, M2 Pro) run quantized 9B models, handling routine queries that require moderate reasoning over a single user's data. In our experiments, an M1 Max 32 GB and an M2 Pro 32 GB serve this role, achieving 21.9 and 12.7 tokens per second respectively on Qwen3.5-9B-Q8.

Tier 2 — Home Server (Panoramic Inference + Model Evolution). A machine with large unified memory (Apple M2 Ultra 192 GB) runs full-precision or lightly quantized 35B-class models. This tier performs cross-domain "panoramic" reasoning by federating data summaries from all Tier 0 devices. In idle periods, it can execute LoRA incremental fine-tuning on Tier 1 models using accumulated data. In our experiments, the M2 Ultra runs Qwen3.5-35B-A3B-Q8 at 49.9 tokens per second.

3.3 Vertical Application Suite

Four applications collect data from complementary life domains:

Table 1: Vertical application suite and the data dimensions they capture.

App	Domain	Data Dimension	Example Signals
Dailyn	Finance	Behavioral layer	Spending categories, amounts, timing
Mealens	Diet	Physical layer	Meal photos, nutritional estimates
Ururu	Mood	Emotional layer	Mood scores, journal entries, triggers
Narrus	Reading	Cognitive layer	Articles read, topics, time spent

The critical design decision is that these are *standalone useful tools*, not data-collection wrappers. A user adopts Dailyn because it is a good accounting app; the fact that it contributes to a cross-domain personal model is invisible and incidental. This resolves the cold-start problem: data accumulates as a natural byproduct of tool usage.

3.4 Federation Protocol

When a Tier 2 device initiates a panoramic query, the federation protocol executes the following steps:

Algorithm 1 LAN Federation Protocol for Panoramic Query

Require: Query q, user ID u, time window Δt, registered data nodes $\mathcal{N}$
Ensure: Panoramic insight $\mathcal{I}$
$\mathcal{S} \leftarrow \{\}$ {Collected summaries}
for each node $n \in \mathcal{N}$ **in parallel do**
  $s_n \leftarrow$ RequestSummary(n, u, Δt) {LAN-local HTTP}
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_n\}$
end for
prompt $\leftarrow$ BuildPanoramicPrompt(q, $\mathcal{S}$)
$\mathcal{I} \leftarrow$ LLMInference(prompt) {Tier 2 model}
AuditLog(|raw|, |transmitted|, nodes) {Privacy audit}
return $\mathcal{I}$

Each data node exposes a standardized JSON Schema interface that returns *summaries*, never raw records. For example, Dailyn's summary endpoint returns aggregated spending by category and trend indicators, not individual transaction records. The privacy audit trail records

the size of raw data on each device versus the bytes actually transmitted, enabling post-hoc verification that no data leakage occurred.

In our experiments, the protocol achieves a 125.5× compression ratio: 108,850 bytes of raw data across all apps compress to 867 bytes of transmitted summaries. The average federation latency (network round-trip to all data nodes) is 469.5 ms, negligible compared to the LLM inference time of 40.3 s.

4 Experimental Setup

4.1 Synthetic User Population

We generate synthetic data for 10 users spanning diverse socioeconomic profiles, each with 90 days of simulated life data across all four applications. Each user has a pre-designed "life event" injected at a specific day to simulate temporal drift—the phenomenon where a significant event causes correlated changes across multiple life domains simultaneously.

Table 2: Synthetic user profiles with injected life events for temporal drift simulation.

ID	Profile	Income	Age	Injected Event
user_01	Factory worker	¥3,500/mo	22	Day 40: factory layoff, income interrupted
user_02	Delivery rider	¥6,000/mo	28	Day 46: severe traffic accident → hospitalization → debt spiral
user_03	Primary teacher	¥5,500/mo	31	Day 35: pregnancy confirmed, spending/diet shift
user_04	Grad student	¥2,000/mo	26	Day 45: thesis rejected by advisor
user_05	E-commerce ops	¥12,000/mo	27	Day 30: promoted to team lead
user_06	Freelance illustrator	¥8–25K/mo	29	Day 60: large client commission (¥50K)
user_07	Hospital resident	¥25,000/mo	33	Day 21: medication error → patient ICU → suspension → breakup
user_08	Tech P7 engineer	¥50,000/mo	32	Day 55: laid off (N+1) → divorce → depression spiral
user_09	Business owner	¥150,000/mo	45	Day 38: ¥800K bad debt → cash flow crisis → hospitalization
user_10	Fund manager	¥830,000/mo	38	Day 25: 4% single-day fund drawdown

The user population is deliberately diverse in income level (from ¥2,000/month graduate student to ¥830,000/month fund manager), life stage (22-year-old factory worker to 45-year-old business owner), and event type (sudden income loss, health changes, career shifts). This diversity tests whether cross-domain insight emergence is robust across demographic segments rather than an artifact of a particular user profile.

Each injected event is designed to produce *correlated multi-domain signatures*. For example, user_01's factory layoff at Day 40 should simultaneously affect: (a) Dailyn: sudden income loss, spending pattern shifts; (b) Mealens: dietary degradation (cheaper, less nutritious meals); (c) Ururu: mood decline, anxiety signals; (d) Narrus: shift in reading topics toward job search, financial advice. A single-domain analysis can detect the within-domain anomaly; only cross-domain analysis can identify the *causal pattern* connecting all four signals.

4.2 Data Generation

For each user, we generate 90 daily records per app (360 records per user, 3,600 total), with the following characteristics:

- **Dailyn**: 3–8 financial transactions per day with category, amount, and memo fields. Amounts are calibrated to each user's income level with realistic variance.
- **Mealens**: 2–4 meal entries per day with estimated nutritional content, meal type (breakfast/lunch/dinner/snack), and optional photo description.
- **Ururu**: 1–3 mood entries per day with a 1–10 mood score, free-text journal entry, and detected emotional state tags.
- **Narrus**: 0–5 reading sessions per day with article/book title, topic tags, reading duration, and highlight/annotation count.

Pre-event and post-event data distributions differ systematically according to each user's injected event, creating ground-truth temporal drift patterns that the system should be able to detect.

4.3 Models and Hardware

All experiments use the Qwen3.5 model family exclusively, eliminating architecture variation as a confound. Table 3 lists the model–device assignments.

Table 3: Model and hardware configuration. All models use Q8 GGUF quantization. Inference engine: llama.cpp on all devices.

Device	Memory	Model	Role
M2 Ultra	192 GB	Qwen3.5-35B-A3B-Q8	Panoramic reasoning (Tier 2)
M2 Ultra	192 GB	Qwen3.5-9B-Q8	Scale comparison
M2 Ultra	192 GB	Qwen3.5-2B-Q8	Scale comparison
M2 Ultra	192 GB	Qwen3.5-0.8B-Q8	IoT baseline
M1 Max	32 GB	Qwen3.5-9B-Q8	Mid-tier inference (Tier 1)
M2 Pro	32 GB	Qwen3.5-9B-Q8	Mid-tier inference (Tier 1)

The Qwen3.5-35B-A3B is a Mixture-of-Experts (MoE) architecture with 35 billion total parameters but only 3 billion activated per token, providing knowledge capacity comparable to a 35B dense model at inference costs closer to a 3B model. This architecture directly embodies the Prism thesis: a "medium-scale but highly efficient" local model.

4.4 Scoring Methodology

Each generated insight is scored on four dimensions, each on a 0–25 scale, for a total of 0–100:

1. **Relevance** (0–25): Does the insight address something meaningful in the user's life?
2. **Specificity** (0–25): Does it reference concrete details from the user's data, or is it generic advice?
3. **Cross-Domain** (0–25): Does the insight connect patterns across multiple life domains?
4. **Actionability** (0–25): Can the user take a concrete action based on this insight?

Scoring is performed by Claude Opus 4.6 (Anthropic, 2025) serving as an LLM-as-Judge. The judge receives the user's full data context and the generated insight, and produces scores with justifications. This approach follows the established practice of using strong LLMs as evaluators (Zheng et al., 2023).

The key metric is the **Insight Increment Ratio** (IIR):

$$\text{IIR} = \frac{S_H}{\frac{1}{4}(S_A + S_B + S_C + S_D)} \tag{1}$$

where S_H is the total score under full panoramic integration (Config H) and S_A, S_B, S_C, S_D are scores under single-domain configurations. An IIR > 1.0 indicates that cross-domain integration adds value beyond what any individual domain can provide.

5 Results

5.1 Experiment A: Cross-Domain Ablation Study

We evaluate eight data configurations on all 10 users using Qwen3.5-35B-A3B-Q8. The configurations range from single-domain (A–D) through two-domain combinations (E–G) to full panoramic integration (H). Results are summarized in Table 4.

Table 4: Cross-domain ablation results. Scores are averages across 10 users (0–100 scale). Config H consistently outperforms all partial configurations.

Config	Data Sources	Avg. Score
A	Finance only (Dailyn)	66.3
B	Diet only (Mealens)	65.1
C	Mood only (Ururu)	63.2
D	Reading only (Narrus)	55.4
E	Finance × Diet	76.6
F	Finance × Mood	76.3
G	Diet × Mood	74.1
H	Full panoramic (all 4)	92.6

The single-domain average is $\bar{S}_{single} = \frac{66.3+65.1+63.2+55.4}{4} = 62.5$, yielding:

$$\text{IIR} = \frac{92.6}{62.5} = 1.48\times \tag{2}$$

The IIR ranges from 1.44 (user_05, e-commerce operations) to 1.53 (user_02, delivery rider with severe traffic accident). Users experiencing dramatic cascading life crises consistently show higher IIR values, confirming that cross-domain integration is most valuable when multiple life domains are simultaneously disrupted. Per-user results are shown in Table 5.

Table 5: Per-user IIR values. Users with dramatic cascading life crises (marked ≥ 1.5) show higher cross-domain value. Range: 1.44–1.53.

User	Profile	Age	IIR
user_01	Factory worker	22	1.47
user_02	Delivery rider	28	1.53
user_03	Primary teacher	31	1.45
user_04	Grad student	26	1.48
user_05	E-commerce ops	27	1.44
user_06	Freelance illustrator	29	1.46
user_07	Hospital resident	33	1.51
user_08	Tech P7 engineer	32	1.52
user_09	Business owner	45	1.51
user_10	Fund manager	38	1.45
Average			**1.48**

Dimension-level analysis. The cross-domain scoring dimension is the primary driver of the IIR gain. Under single-domain configurations (A–D), the cross-domain score averages 2–5 out of 25, which is expected: a model with access to only finance data cannot produce cross-domain insights. Under the full panoramic configuration (H), the cross-domain score rises to 21–24 out of 25. The other three dimensions (relevance, specificity, actionability) show moderate but consistent improvement, as the richer data context enables more specific and actionable recommendations.

Two-domain combinations. The two-domain configurations (E–G) achieve scores in the 74–77 range, intermediate between single-domain (55–66) and full panoramic (92.6). This confirms that cross-domain value scales with the number of integrated domains, though the relationship is super-linear: the jump from two domains to four domains ($92.6 - 75.7 \approx +17$) is proportionally larger than the jump from one domain to two domains ($75.7 - 62.5 \approx +13$ for a $2\times$ increase in data sources).

Qualitative insight examples. Consider user_02 (delivery rider, severe traffic accident at Day 46). With only Dailyn (Config A), the system detects "anomalous spending spikes"—an incomplete observation. With only Ururu (Config C), it detects "mood has declined severely." With full panoramic data (Config H), the system identifies a complete *cascading life crisis*: income drops to zero for 35 consecutive days, calories crash from 1,800 to 200/day (hospital IV), sleep collapses from 6.2h to 1.5h, and reading topics shift from casual browsing to legal aid applications and traffic accident compensation law. The model generates: *"This is a synchronized multi-domain collapse consistent with a severe physical trauma event. Financial, nutritional, emotional, and cognitive patterns all pivot around Day 46. Specific recommendations: apply for work-injury compensation via [referenced article], increase caloric intake to at least 1,200 kcal during recovery, consider [specific low-cost career transition paths based on reading interest patterns]."*

The qualitative difference is categorical: single-domain analysis produces observations; cross-domain analysis produces *causal explanations* and *targeted interventions* grounded in the user's specific data trail.

5.2 Experiment B: Model Scale vs. Insight Quality

To understand the relationship between model scale and insight quality, we run all 10 users under the full panoramic configuration (Config H) on four Qwen3.5 checkpoints, all on the M2 Ultra. Results are shown in Table 6.

Table 6: Model scale vs. insight quality. All runs use Config H (full panoramic) on M2 Ultra 192 GB. Scores are averages ± standard deviation across 10 users.

Model	Avg. Score	Std. Dev.	Quality Tier
Qwen3.5-0.8B-Q8	48.9	4.6	Unusable
Qwen3.5-2B-Q8	64.1	3.2	Marginally useful
Qwen3.5-9B-Q8	79.0	3.4	Good
Qwen3.5-35B-A3B-Q8	84.6	3.4	Excellent

The quality curve exhibits clear diminishing returns:

- 0.8B → 2B: +15.2 points ($\Delta = 15.2$)
- 2B → 9B: +14.9 points ($\Delta = 14.9$)
- 9B → 35B: +5.6 points ($\Delta = 5.6$)

The 2B→9B transition represents the steepest quality improvement relative to the model size increase (4.5× parameters for +14.9 points), while the 9B→35B transition (3.9× parameters for +5.6 points) yields only 37.6% of the per-parameter efficiency. This suggests that for personal data integration tasks, the practical sweet spot is at the 9B scale: it achieves 93.4% of the 35B score (79.0/84.6) while requiring substantially less memory and offering higher throughput.

The 0.8B model's poor performance (48.9) is not merely a matter of degree; it produces qualitatively different outputs—generic platitudes rather than data-grounded insights. This establishes a lower bound: sub-2B models are insufficient for meaningful cross-domain personal reasoning on current architectures.

5.3 Experiment C: Device Performance Benchmark

We benchmark inference speed across the three-tier hardware stack. Table 7 reports tokens per second (TPS) and time to first token (TTFT) for each device–model combination.

Table 7: Inference performance across the three-tier device stack. TPS measured at two context lengths: long (∼4K tokens) and medium (∼2K tokens). Inference engine: llama.cpp, quantization: Q8 GGUF.

Device	Model	TPS (long)	TPS (med)	TTFT (s)
M2 Ultra 192G	Qwen3.5-0.8B-Q8	137.2	135.7	0.088
M2 Ultra 192G	Qwen3.5-2B-Q8	105.1	105.2	0.139
M2 Ultra 192G	Qwen3.5-35B-A3B-Q8	49.9	49.3	0.365
M2 Ultra 192G	Qwen3.5-9B-Q8	41.3	41.4	0.471
M1 Max 32G	Qwen3.5-9B-Q8	21.9	22.0	1.138
M2 Pro 32G	Qwen3.5-9B-Q8	12.7	12.7	1.762

Several observations merit discussion:

MoE efficiency. The Qwen3.5-35B-A3B-Q8 model achieves 49.9 TPS despite having 35B total parameters, outperforming the dense 9B model (41.3 TPS) on the same hardware. This validates the MoE architecture's suitability for on-device deployment: 35B knowledge capacity at super-9B inference speed.

Cross-device scaling. The same model (Qwen3.5-9B-Q8) exhibits a clear performance hierarchy: M2 Ultra (41.3 TPS) > M1 Max (21.9 TPS) > M2 Pro (12.7 TPS). The M2 Ultra's advantage stems from its 800 GB/s memory bandwidth and 76-core GPU, compared to M1 Max's 400 GB/s / 32-core GPU and M2 Pro's 200 GB/s / 19-core GPU. Memory bandwidth, not compute, is the binding constraint for LLM inference.

Context length insensitivity. TPS is nearly identical between long and medium context lengths across all configurations, indicating that our workload (personal data summaries, typically 2–4K tokens) falls well within the efficient operating range of these models.

Practical implications. Even at the lowest tier (M2 Pro, 12.7 TPS), a 200-token insight is generated in under 16 seconds. At Tier 2 (M2 Ultra with 35B-A3B, 49.9 TPS), the same insight takes 4 seconds. Both are well within acceptable latency for a personal insight system that generates reports on demand rather than in real-time conversation.

5.4 Experiment D: Federation Protocol Evaluation

We evaluate the federation protocol (Section 3.4) in a real multi-device deployment across a local area network.

Topology. The M2 Ultra serves as the panoramic query coordinator. The M1 Max hosts data for Mealens (diet) and Ururu (mood). The M2 Pro hosts data for Narrus (reading) and Dailyn (finance). All three machines communicate over a standard home WiFi network.

Protocol. The panoramic node sends summary requests to both data nodes in parallel, collects JSON summaries, constructs a unified prompt, and runs inference on Qwen3.5-35B-A3B-Q8.

Results from 5 runs are summarized in Table 8.

Table 8: Federation protocol performance over 5 runs. Raw data stays on origin devices; only summaries are transmitted.

Metric	Value
Raw data on devices (total)	108,850 bytes
Transmitted summaries (total)	867 bytes
Compression ratio	125.5×
Federation latency (avg)	469.5 ms
Federation latency (min)	103.2 ms
Federation latency (max)	1,459.5 ms
LLM inference time (avg)	40.3 s
End-to-end latency (avg)	40.8 s
Raw data leaked across devices	**0 bytes (all runs)**

Compression analysis. The 125.5× compression ratio reflects the fundamental design of the summary interface: rather than transmitting 90 days of individual transactions, meal logs, mood entries, and reading records, each data node computes domain-specific aggregates (e.g., weekly spending by category, daily mood trend slopes, top reading topics) and returns a compact JSON summary. The semantic loss from this compression is deliberately minimal: the panoramic model needs patterns and trends, not individual data points.

Latency decomposition. Federation latency (469.5 ms average) is ∼1.2% of the total end-to-end latency (40.8 s), which is dominated by LLM inference (40.3 s). This validates the architectural decision to separate data collection from inference: the network overhead of federated data collection is negligible. The max-latency outlier (1,459.5 ms) likely reflects WiFi contention and is not architecturally significant.

Privacy verification. Across all 5 runs, the audit trail confirms zero raw-data leakage. The data nodes' summary endpoints are stateless: they receive a time-window parameter and return pre-computed aggregates. No mechanism exists for the panoramic node to request or receive raw records, as the API simply does not expose them. This is privacy by *architecture*, not by access control.

6 Discussion

6.1 IIR and the 1.5× Target

Our initial target for the Insight Increment Ratio was 1.5×. The observed average of 1.48× approaches but does not fully reach this target in aggregate, though four out of ten users (user_02, user_07, user_08, user_09) individually exceed 1.5×. We analyze the structural factors that shape the IIR:

Scoring saturation. Single-domain configurations already achieve moderate scores (55.4–66.3 out of 100) because the relevance, specificity, and actionability dimensions can be partially satisfied without cross-domain data. A user's spending patterns alone are enough to generate relevant and somewhat actionable financial advice. The IIR formula divides by this non-trivial baseline, compressing the ratio.

Cross-domain dimension ceiling. While the cross-domain dimension is the primary driver of the IIR gain (jumping from 2–5 points in single-domain to 21–24 in panoramic), this single dimension can contribute at most 25 out of 100 points. The other three dimensions show more modest improvements from data integration (∼2–5 points each), limiting the overall score amplification.

Crisis users exceed the target. Notably, the four users who exceed 1.5× (user_02: 1.53, user_07: 1.51, user_08: 1.52, user_09: 1.51) all experienced severe cascading life crises—a traffic accident leading to hospitalization, a medical error resulting in suspension, a layoff triggering divorce, and a bad debt causing cash flow collapse. These users exhibit the strongest cross-domain signal because their crises simultaneously perturb all four data dimensions, creating correlations that single-domain analysis cannot detect. This supports a key finding: **the value of cross-domain integration is greatest precisely when users need it most—during life crises.**

Synthetic data limitations. Our synthetic data, while designed with realistic correlations, may underrepresent the richness of real user data. In particular, the textual content of mood journal entries and reading annotations—which would provide the richest signals for cross-domain reasoning in real data—is necessarily simpler in synthetic generation. We hypothesize that real user data would yield higher IIR values.

The 1.48× average IIR represents a substantive improvement. A 48% increase in insight quality from data integration—with no change to the model—is practically significant. Moreover, the qualitative examples (Section 5.1) demonstrate that the nature of the improvement is categorical: single-domain analysis produces observations, while cross-domain analysis produces explanations and interventions.

6.2 The Diminishing Returns Plateau

The model-scale experiment (Section 5.2) reveals that the practical value curve flattens above 9B parameters. The 9B model achieves 93.4% of the 35B model's score while running at 84% of its speed on the same hardware (41.3 vs. 49.9 TPS, noting that the 35B MoE is faster due to sparse activation).

This finding has a concrete practical implication: **a user with only a 32 GB device (M1 Max or M2 Pro) can run a 9B model and achieve near-optimal insight quality.** The 35B MoE model on a 192 GB machine provides marginal improvement. The bottleneck for personal AI quality is not model scale—it is data availability and integration.

This aligns with the broader PRISM thesis: in the personal data regime, data richness matters more than model size.

6.3 Social Impact: Early Warning and Preventive Intervention

Beyond personal convenience, cross-domain data integration enables a class of interventions with significant social value. Our synthetic user population includes several cases where the system's cross-domain analysis could function as an *early warning system* for cascading life crises.

Cascading crisis detection. Consider user_02 (delivery rider, severe traffic accident at Day 46). The single-domain mood analysis (Config C) detects "mood has declined." But the full panoramic analysis (Config H) detects a synchronized collapse across all four domains—income drops to zero, calories crash from 1,800 to 200/day, sleep collapses from 6.2h to 1.5h, and reading topics shift from casual browsing to legal aid and accident compensation—identifying a *life crisis cascade* rather than an isolated mood dip. Similarly, user_09 (business owner) shows the system detecting a debt-induced health crisis trajectory: financial stress → insomnia → cardiac symptoms → hospitalization, visible only when financial, emotional, dietary, and reading data are analyzed together.

Pre-crisis intervention window. Several users exhibit *foreshadowing signals* in the days before their crisis events. User_08 (P7 engineer) shows rising anxiety in mood data and shifts in reading topics (from technical articles to layoff-related content) 10 days before the actual layoff. User_07 (hospital resident) shows accumulating sleep deprivation and stress escalation across 3 consecutive night shifts before the medication error. A deployed system could detect these multi-domain warning patterns and prompt early intervention—suggesting rest, financial planning, or professional support—before the crisis materializes.

Psychological support and recovery monitoring. Post-crisis, the system can track recovery trajectories across domains. User_02's data shows a slow but measurable recovery arc: reading topics gradually shift from legal aid to career transition, calories slowly increase from

hospital levels, and mood stabilizes at a new (lower) baseline. This multi-dimensional recovery tracking could complement professional mental health support, providing therapists with objective behavioral data rather than relying solely on self-reporting.

Societal implications. The ability to detect cascading life crises through passive data collection has implications for public health infrastructure. In populations where mental health services are scarce or stigmatized, an on-device system that can detect distress patterns without requiring the user to actively seek help—and without exposing their data to any external party—could serve as a crucial safety net. The privacy-preserving architecture is essential here: users are more likely to maintain honest records (mood journals, spending logs) when they trust that no one else will see the data.

6.4 Practical Deployment Considerations

Hardware accessibility. The entire experimental stack uses consumer Apple hardware. The M2 Pro (the least powerful Tier 1 device) is available in MacBook Pro configurations starting at approximately $2,000 USD. The M2 Ultra Mac Studio (Tier 2) costs approximately $4,000 USD. While not inexpensive, these are consumer-grade devices that many knowledge workers already own or could justify purchasing.

Latency budget. The end-to-end panoramic query latency of 40.8 s may seem high for a conversational assistant, but it is appropriate for the intended use case: periodic (daily or weekly) personal insight reports, not real-time chat. A user would request a "weekly life summary" and receive a comprehensive cross-domain analysis within a minute—a fundamentally different interaction pattern from cloud chatbots.

Cold start. The vertical app design addresses cold start through utility-driven adoption. Users download Dailyn because they need an accounting app, not because they want to train a personal AI. Data accumulation is a byproduct of normal usage. After 30–60 days of natural usage across multiple apps, the system has sufficient data for meaningful cross-domain insights.

6.5 Limitations

Synthetic data. All results are based on synthetic users. While we have designed the data generation process to be realistic, synthetic data inevitably underrepresents the complexity, noise, and idiosyncrasy of real human behavior. A user study with real participants is the critical next step.

LLM-as-Judge. Our scoring methodology relies on Claude Opus 4.6 as an evaluator. While LLM-as-Judge has been validated in prior work (Zheng et al., 2023), it introduces potential biases—particularly the risk that the judge systematically favors longer or more structured outputs, which the panoramic configuration naturally produces. Future work should complement LLM scoring with human evaluation.

Single model family. All experiments use Qwen3.5, eliminating architecture as a variable but limiting generalizability. Different model families (Llama, Mistral, Gemma) may exhibit different model-scale vs. insight-quality curves.

Security model. Our privacy analysis focuses on data *leakage* (raw data leaving devices). We do not address adversarial scenarios such as prompt injection attacks on the federation protocol, or side-channel attacks on the inference process. A production deployment would require a more comprehensive threat model.

Quantization. All models use Q8 quantization. While this provides a good balance of quality and efficiency, the interaction between quantization level and personal-data reasoning quality has not been explored. Aggressive quantization (Q4, Q2) may disproportionately degrade the nuanced reasoning required for cross-domain insight generation.

7 Conclusion

We have presented PRISM, a system that demonstrates how cross-domain personal data integration on consumer hardware produces emergent insights unreachable by single-domain analysis. Our ablation study across 10 synthetic users and 8 data configurations establishes an average Insight Increment Ratio of 1.48×, with the cross-domain dimension driving nearly all of the gain. The model-scale analysis reveals that a 9B model captures 93.4% of the 35B model's quality, placing the practical sweet spot well within the reach of consumer 32 GB devices. The federation protocol achieves 125.5× data compression with zero raw-data leakage, proving that multi-device personal AI can operate without any data leaving the user's premises.

The broader implication is a paradigm inversion. The dominant cloud AI model of *large model + small data* reaches an inherent ceiling because users rationally withhold their most valuable data from remote servers. The PRISM paradigm of *medium model + rich data* exploits an orthogonal axis of improvement: not larger parameters, but deeper and more authentic data. As on-device inference capabilities continue to improve with each hardware generation, this paradigm becomes increasingly viable.

The critical next step is a longitudinal user study with real participants using real vertical applications. The synthetic data results presented here are a necessary proof of concept; the true test of the PRISM thesis requires real users generating real data over real time.

Reproducibility

All code, synthetic data, and experiment scripts are available at https://github.com/AtomGradient/Prism. The experimental results and interactive visualizations are hosted at https://atomgradient.github.io/Prism.

References

Chen, Z., Wang, Y., Liu, Z., et al. (2025). A survey of personalized large language models. *arXiv preprint arXiv:2502.11528*.

Chuang, Y.-S., Xie, S., Luo, H., Kim, Y., Glass, J., & He, P. (2024). On the way to LLM personalization: Learning to remember user conversations. *arXiv preprint arXiv:2411.13405*.

Li, T., et al. (2025). PersonalLLM: Tailoring LLMs to individual preferences. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*.

Wang, T., Zhang, Z., Liang, J., et al. (2024). Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Huang, Y., Cao, J., Zhang, X., et al. (2024). LIBER: Lifelong user behavior modeling based on large language models. *arXiv preprint arXiv:2411.14713*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282.

Ramage, D. & Mazzocchi, S. (2020). Federated analytics: Collaborative data science without data collection. *Google AI Blog*.

Hannun, A., et al. (2023). MLX: An efficient machine learning framework for Apple Silicon. *Apple Machine Learning Research*.

Gerganov, G. (2023). llama.cpp: LLM inference in C/C++. https://github.com/ggerganov/llama.cpp.

Anthropic (2025). Claude Opus 4.6 model card. https://www.anthropic.com.

Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*.

Qwen Team (2025). Qwen3.5 technical report. *Alibaba Cloud*.