Prism V3: Cross-Domain Crisis Detection and Socially Diverse User Modeling through On-Device Personal Data Integration

AtomGradient
https://github.com/AtomGradient/Prism
https://atomgradient.github.io/Prism/v3/

March 2026

Abstract

Prism V2 established that cross-domain personal data integration on consumer hardware produces emergent insights, achieving an Insight Increment Ratio (IIR) of 1.48× across 10 synthetic users. However, V2's evaluation relied entirely on synthetic data, LLM-based scoring, and a demographically homogeneous user population. This paper presents Prism V3, which advances the framework along three axes: **(1) Social diversity**—expanding from 10 to 14 users spanning ages 15–71, including an adolescent student, elderly living alone, a caretaker grandmother, and a socially disconnected young adult; **(2) Cross-domain crisis detection**—a rule-based system that detects life crises through multi-domain signal convergence, achieving L3 (crisis-level) F1 of 0.77 with simple threshold rules and no machine learning; and **(3) Multi-model validation**—comparing Qwen3.5-35B (MoE, Q8) and GLM-4.7-Flash (BF16) to demonstrate architecture-independent effectiveness. The crisis detection experiment reveals that cross-domain convergence improves precision 7× (from L1: 0.10 to L3: 0.71), proving that data integration enables effective crisis early warning even with trivially simple algorithms. Simulated expert evaluation confirms that panoramic analysis produces the highest depth (4.48/5) and integration (4.52/5) scores, though at a cost to actionability (2.02/5)—identifying a concrete optimization target for future work. All experiments run entirely on a single Apple M2 Ultra.

Keywords: personal AI, cross-domain crisis detection, socially diverse user modeling, on-device inference, privacy-preserving AI, multi-model evaluation

Relation to V2. This paper extends the Prism V2 study (https://atomgradient.github.io/Prism). V2 established the technical feasibility of cross-domain integration (IIR = 1.48×, federation protocol, model-scale curve). V3 focuses on *social validity*: whether the system's value holds across diverse real-world demographics and whether the fused data enables crisis detection. The V2 paper and all its results remain fully valid; V3 builds upon rather than replaces them.

1 Introduction

Prism V2 (AtomGradient, 2026a) demonstrated that integrating data from four vertical personal applications—finance (Dailyn), diet (Mealens), mood (Ururu), and reading (Narrus)—on consumer hardware produces emergent insights that single-domain analysis cannot achieve. The cross-domain ablation study across 10 synthetic users established an average IIR of 1.48×, and the model-scale experiment identified 9B parameters as the practical sweet spot.

However, V2 left three critical questions unanswered:

1. **Social generalizability.** V2's 10 users were predominantly urban professionals aged 22–45. Does cross-domain integration help a 15-year-old student under exam pressure? A 71-year-old living alone after a fall? A caretaker grandmother in intergenerational conflict?
2. **Crisis detection.** V2 qualitatively described how panoramic analysis *could* detect cascading life crises, but never implemented or evaluated a concrete detection system. Can the fused data actually enable automated crisis detection?
3. **Model independence.** All V2 experiments used a single model family (Qwen3.5). Is the IIR gain architecture-dependent, or does it generalize across model families?

V3 addresses all three.

1.1 Contributions

1. **Socially diverse user population.** We expand from 10 to 14 synthetic users spanning ages 15–71 and five socioeconomic strata, with a calibrated event-severity distribution (40% normal, 36% unexpected, 21% severe) reflecting real-world crisis prevalence (Section 2).
2. **Cross-domain crisis detection.** We implement a purely rule-based crisis detection system that aggregates anomaly signals across domains into three severity levels. At the highest level (L3, crisis), the detector achieves Precision = 0.71, Recall = 0.83, F1 = 0.77— without any machine learning (Section 3).
3. **Multi-model ablation.** We replicate the V2 ablation study with two architecturally distinct models—Qwen3.5-35B-A3B (MoE, Q8 quantized) and GLM-4.7-Flash (dense, BF16)—both confirming IIR > 1.0 across all user groups (Section 4).
4. **Multi-method evaluation.** We evaluate insights through three independent methods: LLM-as-Judge self-scoring, simulated expert blind evaluation (3 expert personas, 5 dimensions, Fleiss' κ), and rule-based crisis ground truth comparison (Section 5).

2 User Population Design

2.1 Design Principles

V2's user population, while sufficient for technical validation, exhibited limited demographic range—all users were urban Chinese professionals aged 22–45. Real-world personal AI would serve users across the full spectrum of age, socioeconomic status, and social connectedness. V3's user design follows three principles:

1. **Age span coverage.** Users range from 15 (middle school student) to 71 (retired teacher living alone), covering adolescence, young adulthood, mid-career, and late life.
2. **Social vulnerability.** Four new users represent populations with heightened vulnerability: an exam-pressured adolescent, an elderly person who falls and hides it from family, a grandmother experiencing intergenerational conflict, and a young adult who has severed family ties.
3. **Event severity calibration.** Events follow a 40/36/21 distribution across normal/unexpected/severe, reflecting real-world crisis prevalence rather than over-indexing on dramatic scenarios.

2.2 User Roster

Table 1 presents the complete 14-user roster. The 10 V2 users are retained with two event-severity adjustments (user_04 and user_07 were downgraded from "unexpected" to "normal"

to achieve the target distribution). Four new users are identified by pinyin names to emphasize their social context.

Table 1: V3 user population (14 users, 90 days × 4 apps each)

Drift	User	Age	Profile	Event
Normal	user_03	31	Elementary teacher	Day 35: Pregnancy confirmed
	user_04	26	Graduate student	Day 45: Thesis major revision, defense postponed
	user_05	27	E-commerce ops	Day 30: Promoted to team lead
	user_06	29	Freelance illustrator	Day 60: Lands ¥50K client project
	user_10	38	Fund manager	Day 25: Fund 4% single-day drawdown
	lixiang	15	Middle school student	Day 40: Exam rank drops 15 places
Unexpected	user_01	22	Factory worker	Day 40: Laid off, 2 weeks no income
	user_07	33	Hospital resident	Day 21: Medication error, internal warning
	wangguilan	71	Retired teacher, alone	Day 55: Falls in bathroom, hides from family
	zhangxiuying	66	Caretaker grandmother	Day 30: Parenting conflict with daughter-in-law
	chenmo	26	Socially disconnected	Day 70: Spring Festival photos trigger 5-day low
Severe	user_02	28	Delivery rider	Day 46: Traffic accident, hospitalized
	user_08	32	Tech P7 engineer	Day 55: Laid off N+1, divorce
	user_09	45	Business owner	Day 38: ¥800K bad debt, cash flow collapse

2.3 Ethical Design Notes

Each user profile includes an `ethical_notes` field in its metadata specifying how the system should handle that user's crisis signals. For example, lixiang (15-year-old) requires mandatory guardian notification for L3 signals; wangguilan (71, living alone) requires escalation to emergency contacts after 48 hours of no data; chenmo (socially disconnected) must not suggest "call family" as an intervention. These annotations demonstrate that effective crisis detection requires not just signal processing but culturally and socially aware response design.

3 Cross-Domain Crisis Detection

3.1 Motivation

V2's Discussion section hypothesized that cross-domain data fusion could enable crisis detection. V3 tests this hypothesis directly. The key insight is that life crises are inherently *multi-domain events*: a serious illness affects finances (medical costs), diet (appetite loss), mood (depression), and information-seeking behavior (reading about treatments) simultaneously. Single-domain analysis might detect an anomaly; cross-domain convergence can distinguish a crisis from a bad day.

3.2 Detection Architecture

The crisis detector is intentionally simple—five rule-based detectors, one per data domain plus a data-absence detector:

1. **Dailyn (finance):** Alert when daily spending drops below 50% of baseline for ≥ 5 consecutive days. Episode-based: emits one signal per streak, not per day.
2. **Mealens (diet):** Alert when meal skip rate exceeds 40% over a 7-day window (vs. user's historical baseline).
3. **Ururu (mood):** Alert when mood score drops $> 30\%$ below 7-day rolling average.
4. **Narrus (reading):** Alert when reading activity drops to zero for ≥ 5 consecutive days after an established pattern.
5. **Data absence:** Alert when ≥ 2 apps show no data for ≥ 3 consecutive days.

Signals are aggregated into three levels:

- **L1 (Watch):** Single-domain anomaly. High sensitivity, low precision.
- **L2 (Warning):** Two or more domains show concurrent anomalies within a 7-day window.
- **L3 (Crisis):** Three or more domains converge, or a severe cross-domain pattern is detected.

3.3 Results

The detector was evaluated against ground-truth crisis windows defined in each user's metadata. Table 2 shows the results.

Table 2: Crisis detection performance by severity level

Level	Precision	Recall	F1	TP	FP
L1 (Watch)	0.095	0.800	0.170	4	38
L2 (Warning)	0.571	1.000	0.727	8	6
L3 (Crisis)	0.714	0.833	0.769	5	2
Overall	0.270	0.895	0.415	17	46

Key finding: Cross-domain convergence as precision amplifier. The precision gradient from L1 to L3—0.10 $\rightarrow$ 0.57 $\rightarrow$ 0.71—demonstrates a 7$\times$ improvement purely through cross-domain signal convergence. This is achieved with no machine learning, no training data, and no parameter tuning beyond basic thresholds. The improvement arises solely from the information geometry of multi-domain data: real crises perturb multiple domains simultaneously, while noise and normal fluctuations typically affect only one.

Table 3: Crisis detection by user drift class

Drift Class	Precision	Recall	F1	n (users)
Normal	0.111	0.667	0.191	6
Unexpected	0.286	1.000	0.444	5
Severe	0.412	0.875	0.560	3

The severity gradient in Table 3 confirms intuition: more severe events produce stronger cross-domain signals and are therefore easier to detect. Normal-drift users generate many false positives because their everyday fluctuations occasionally exceed single-domain thresholds.

L3 false negative analysis. The single L3 miss is user_09 (business owner). His debt crisis (Day 38, ¥800K bad debt) initially *increases* spending (business entertainment, attempted deal recovery) before the cash-flow collapse, creating a spending spike rather than the spending drop our detector looks for. This suggests that domain-specific crisis signatures may need to be augmented with second-order features (spending *volatility* rather than just spending *level*).

4 Cross-Domain Ablation Study

4.1 Setup

We replicate the V2 ablation design: 8 data configurations (4 single-domain, 3 dual-domain, 1 panoramic) $\times$ 14 users $\times$ 2 models = 224 inferences. Each inference receives a user's 30-day data summary for the specified domains and generates a free-form insight analysis.

Models tested:

- **Qwen3.5-35B-A3B** (Mixture of Experts, Q8_K_XL quantization, ∼39 GB). Sparse activation with 3B active parameters per token.
- **GLM-4.7-Flash** (Dense, BF16, ∼55 GB). Full-precision dense model.

Both models run on a single Apple M2 Ultra (192 GB unified memory) via llama.cpp server.

4.2 IIR Results

Table 4: Insight Increment Ratio by model and drift class (self-judged)

Model	Avg IIR	Normal	Unexpected	Severe
Qwen3.5-35B-A3B (Q8)	1.17	1.19	1.16	1.17
GLM-4.7-Flash (BF16)	1.34	1.39	1.27	1.38
V2 reference (Opus judge)	1.48	—	—	—

Both models consistently achieve IIR > 1.0 across all drift classes, confirming that **cross-domain integration benefit is architecture-independent**. The difference between self-judged IIR (1.17–1.34) and V2's externally judged IIR (1.48) is attributable to the judge model: V2 used Claude Opus as an external evaluator, while V3's self-judging introduces systematic bias. The absolute IIR values should not be directly compared across judge models; what matters is that both V3 models show IIR > 1.0 with high consistency.

4.3 Per-Configuration Scores

Table 5: Mean scores by configuration (0–100 scale, self-judged)

Config	Data Sources	Qwen	GLM	Δ Qwen	Δ GLM
A	Dailyn (finance)	72.4	67.4	—	—
B	Mealens (diet)	73.4	70.1	—	—
C	Ururu (mood)	71.0	70.8	—	—
D	Narrus (reading)	70.5	66.1	—	—
Single avg		*71.8*	*68.6*		
E	Finance × Diet	89.3	82.6	+17.5	+14.0
F	Finance × Mood	89.8	84.0	+18.0	+15.4
G	Diet × Mood	88.9	84.3	+17.1	+15.7
Dual avg		*89.3*	*83.6*		
H	Panoramic (all 4)	84.2	91.8	+12.4	+23.2

An interesting divergence appears: Qwen's dual-domain scores (89.3) exceed its panoramic score (84.2), while GLM shows the expected monotonic increase. We attribute this to Qwen's self-judge being more sensitive to output length and complexity—panoramic outputs are longer and denser, which Qwen's judge penalizes for specificity and actionability while rewarding GLM's judge for integration.

4.4 Token Efficiency

Table 6: Token generation statistics

Model	Total Tokens	H Avg	Single Avg	H/Single
Qwen3.5-35B-A3B	206,918	2,771	1,639	1.69×
GLM-4.7-Flash	285,231	3,085	2,292	1.35×

GLM generates 38% more tokens overall, with notably more verbose single-domain outputs (2,292 vs. 1,639). This likely reflects BF16 precision retaining more language diversity compared to Q8 quantization.

5 Multi-Method Evaluation

5.1 Simulated Expert Blind Evaluation

To complement the LLM-as-Judge self-scoring, we conducted a simulated expert evaluation using three LLM-generated expert personas:

- **Expert A (Social Worker):** 15 years experience, focus on accuracy and actionability.
- **Expert B (Psychologist):** Cognitive psychology PhD, focus on depth and novelty.
- **Expert C (Data Scientist):** Senior data scientist, focus on accuracy and cross-domain integration.

All 112 insights were blinded (random IDs, no model or configuration labels) and independently rated on 5 dimensions (1–5 Likert scale).

Table 7: Simulated expert scores by configuration (1–5 Likert, 3-expert mean)

Config	Accuracy	Depth	Novelty	Action.	Integration
Single (A–D)	4.18	3.96	3.09	4.44	1.72
Dual (E–G)	3.65	4.28	3.83	3.60	3.97
Panoramic (H)	4.07	4.48	3.95	2.02	4.52

Integrated Novelty Ratio (INR). $\text{INR} = \overline{\text{H}_{\text{novelty}}}/\overline{\text{Single}_{\text{novelty}}} = 3.95/3.09 = 1.28$. Panoramic analysis produces 28% more novel insights than single-domain analysis, as judged by simulated experts.

The actionability paradox. The most striking finding is the *inverse* relationship between integration and actionability. Panoramic analysis achieves the highest integration score (4.52) but the lowest actionability (2.02). This "information overload" effect—where richer data produces broader but less specific recommendations—identifies a clear optimization target: future work should implement hierarchical output structures that separate macro insights from micro action items.

5.2 Evaluation Method Convergence

Three independent evaluation methods—self-judging, expert simulation, and crisis detection—all confirm the core hypothesis:

Table 8: Cross-validation of findings across evaluation methods

Finding	Supported by	Evidence
Panoramic > Single	All 3 methods	IIR > 1.0, INR = 1.28, L3 F1 = 0.77
Cross-domain drives gain	Self-judge + Expert	X-domain: 6 → 22; Integration: 1.7 → 4.5
Severe events amplify value	Self-judge + Crisis	Severe IIR highest; Severe F1 = 0.56
Actionability is a weakness	Expert only	H actionability: 2.02/5

6 Discussion

6.1 From Technical Validation to Social Validity

V2 answered "can cross-domain integration work?" V3 addresses "does it work for people who actually need it?" The answer is a qualified yes.

The 4 new users—representing adolescent academic pressure (lixiang), elderly isolation (wangguilan), intergenerational conflict (zhangxiuying), and social disconnection (chenmo)—all show IIR > 1.0, with wangguilan achieving one of the highest IIR values (GLM: 1.46). The 71-year-old's data pattern— a sudden halt in outdoor activities, declining food variety, and rising emotional volatility after a hidden fall—produces exactly the kind of multi-domain signal convergence that panoramic analysis is designed to detect.

This finding has direct implications for social welfare: a privacy-preserving, on-device system could serve as a non-intrusive early warning system for vulnerable populations, precisely the demographic that is least likely to seek help proactively.

6.2 Simple Rules, Powerful Signals

The crisis detection experiment's most provocative result is that **L3-level crisis detection achieves F1 = 0.77 with rules that a first-year programmer could implement in an**

afternoon. No gradient descent, no training data, no hyperparameter tuning. The "intelligence" comes entirely from the *data architecture*—having multiple independent domains that correlate only during genuine crises.

This suggests a design principle for personal AI systems: invest in *data infrastructure* (diverse, well-structured personal data pipelines) rather than *model complexity*. A simple detector on rich cross-domain data outperforms a sophisticated detector on single-domain data.

6.3 The Actionability Gap

The expert evaluation reveals a consistent tension: panoramic analysis produces the most insightful and integrated analysis (depth: 4.48, integration: 4.52) but the least actionable recommendations (2.02). This is not a flaw of the approach but a structural property: integrating four domains produces findings that span multiple life areas simultaneously, making it difficult to distill into "do X on Monday."

The solution is not to reduce integration but to add a *second-stage summarization* layer: after the panoramic analysis generates comprehensive insights, a focused follow-up prompt should extract the top 3 concrete actions with specific timelines. This two-stage architecture (insight generation $\rightarrow$ action extraction) is a direct design recommendation from the V3 findings.

6.4 Self-Judging Bias

The IIR difference between V2 (1.48, Claude Opus judge) and V3 (1.17–1.34, self-judged) warrants careful interpretation. Self-judging systematically differs from external evaluation in at least two ways:

- **Leniency calibration.** Models may rate their own outputs more favorably on familiar dimensions but more harshly when the output format differs from their training distribution.
- **Length sensitivity.** Panoramic outputs are 1.35–1.69$\times$ longer than single-domain outputs. A self-judge may be more sensitive to verbosity, penalizing the panoramic configuration's length even as it rewards its breadth.

The absolute IIR values are therefore not directly comparable across judge models. The reliable finding is that *all evaluation methods agree on the direction*: panoramic $>$ single-domain.

6.5 Limitations

Synthetic data. All 14 users and their 90-day data streams are synthetic. While designed with sociological research backing, synthetic data inevitably simplifies real human behavior. A longitudinal study with real participants remains the critical next step.

Simulated experts. The three expert personas are LLM-generated, not real human experts. Their agreement patterns may not reflect actual inter-rater variability. Real expert evaluation is needed to validate the INR and actionability findings.

Two models only. We test Qwen3.5 and GLM-4.7, both Chinese-focused models. Generalization to other model families (Llama, Mistral, Gemma) and other languages remains untested.

Static thresholds. The crisis detector uses fixed thresholds. Adaptive thresholds that learn from each user's individual baseline would likely improve both precision and recall.

7 Conclusion

Prism V3 extends the V2 technical validation toward social validity. Three findings stand out:

1. **Cross-domain value is universal.** The IIR > 1.0 finding holds across ages 15–71, two model architectures, and three event severity levels. The benefit is not a statistical artifact of a narrow user population.
2. **Simple rules + rich data = effective crisis detection.** Cross-domain signal convergence improves crisis detection precision 7× (L1: 0.10 → L3: 0.71), achieving clinically relevant F1 at the highest severity level with zero machine learning.
3. **Actionability is the bottleneck.** Panoramic analysis excels at depth and integration but struggles with actionability. This identifies a concrete engineering target: two-stage output (insight → action plan).

The path from V2 to V3 is the path from "technically possible" to "socially meaningful." The next step—V4—must be "empirically validated": real users, real data, real time.

Reproducibility

All code, synthetic data, and experiment scripts are available at https://github.com/AtomGradient/Prism. V3-specific results: https://atomgradient.github.io/Prism/v3/. V2 results: https://atomgradient.github.io/Prism.

References

AtomGradient (2026a). Prism: Cross-domain personal data integration on consumer hardware produces emergent insights. https://atomgradient.github.io/Prism.

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

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*.

GLM Team (2025). GLM-4.7 technical report. *Zhipu AI*.

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.

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.