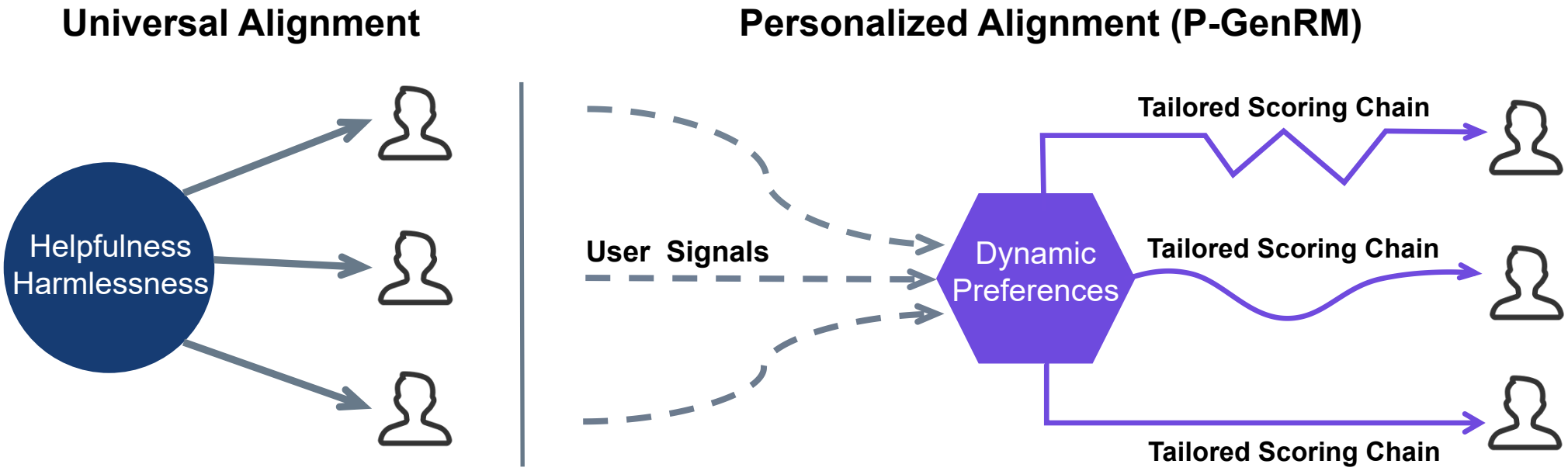
P-GenRM: Personalized Generative Reward Model With Test-time User-based Scaling

Pinyi Zhang, Ting-En Lin, Yuchuan Wu, et al.
Tongyi Lab, Alibaba Group



In open-ended dialogues, how do we define 'helpfulness'?

Tongyi

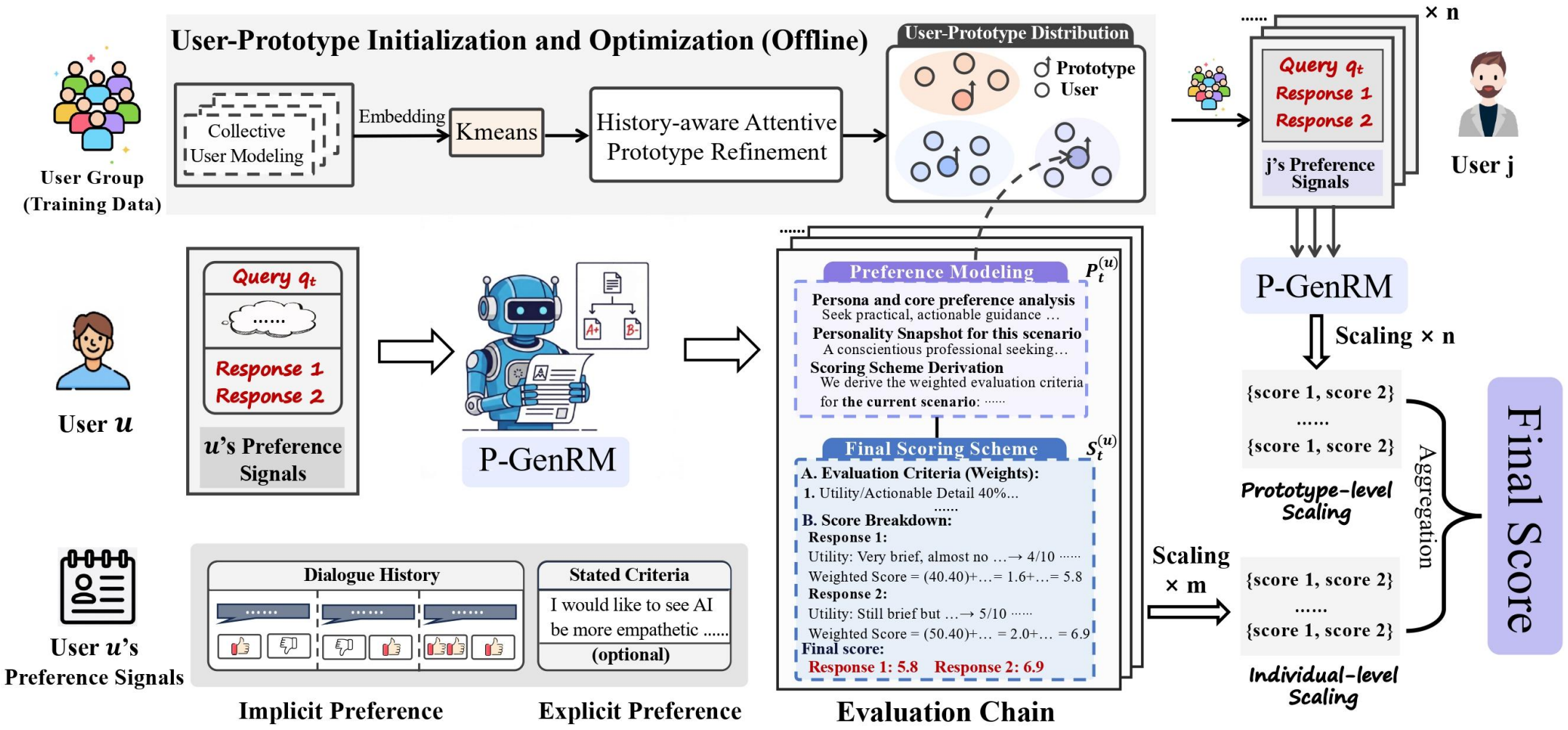
Helpfulness = ??? & ???

Helpfulness = Personality & Scenario

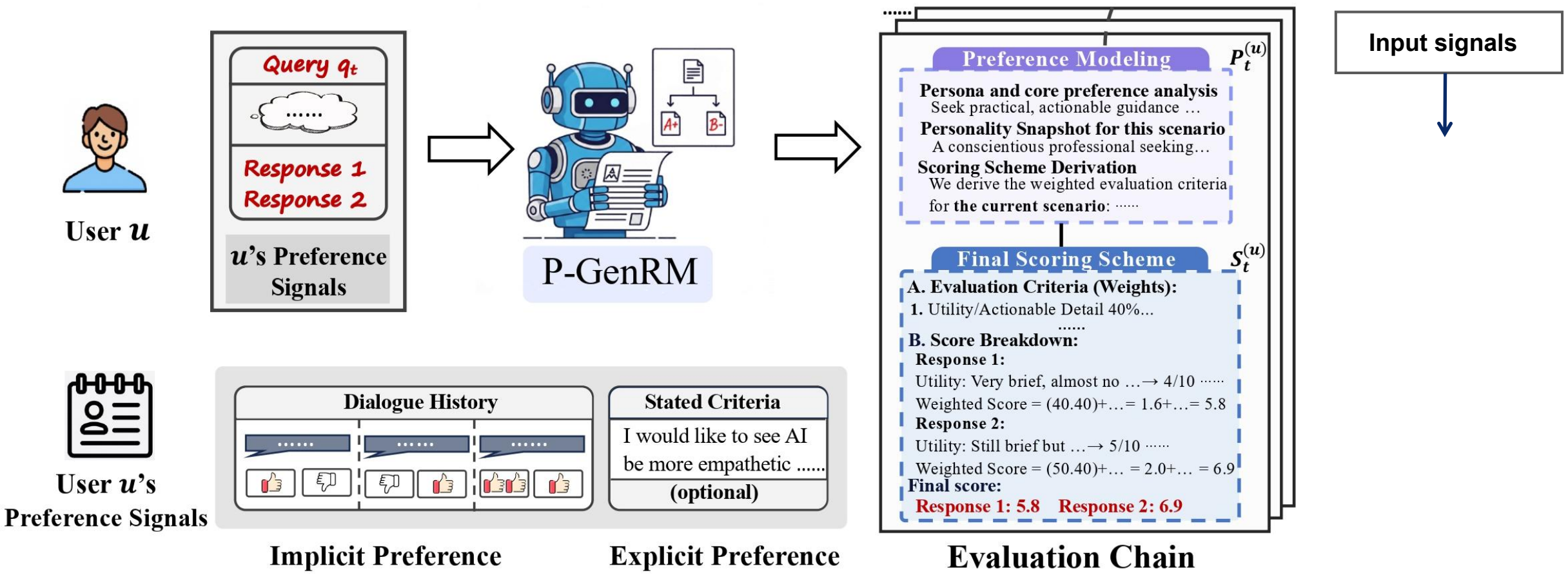
Challenges: Dynamic reward modeling and Generalization

	Current paradigm	P-GenRM
The Modeling Problem	**Static Rules** Reduces diverse preferences into limited, fixed rules.	**Adaptive rewarding** Infers dynamic, user-specific rubrics for different scenarios.
The Generalization Problem	**Weak Generalization** Performance declines for users with limited interaction history.	**Prototype transfer** Leverages collaborative signals to infer preferences with sparse data.

How P-GenRM works

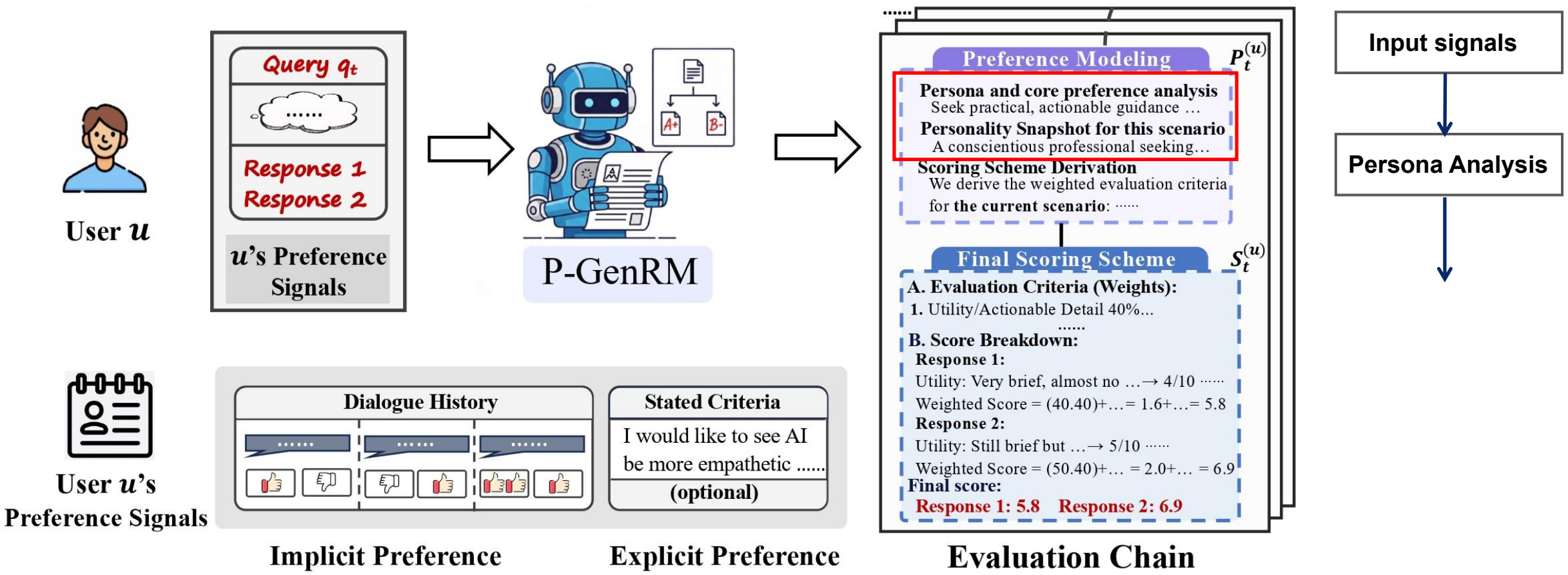


Structured, Dynamic Evaluation Chain



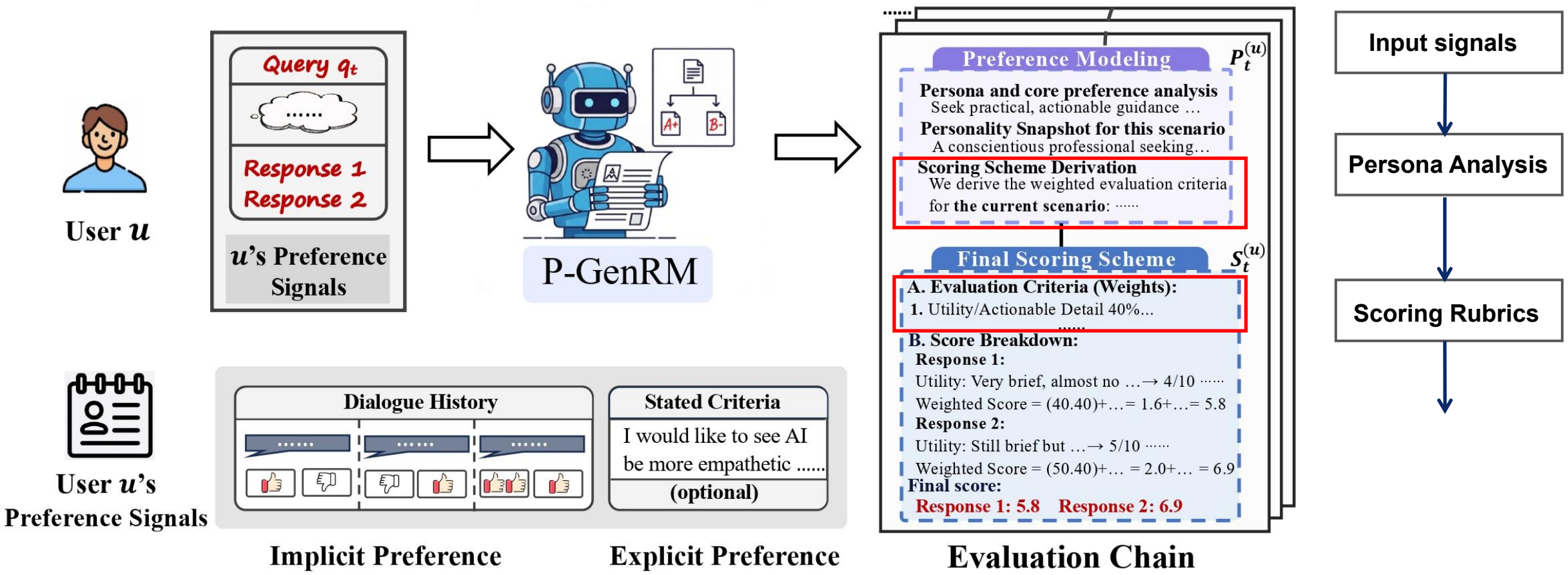
- From hybrid preference signals to structured, dynamic evaluation chain
- From opaque scalars to interpretable, auditable rewards

Structured, Dynamic Evaluation Chain



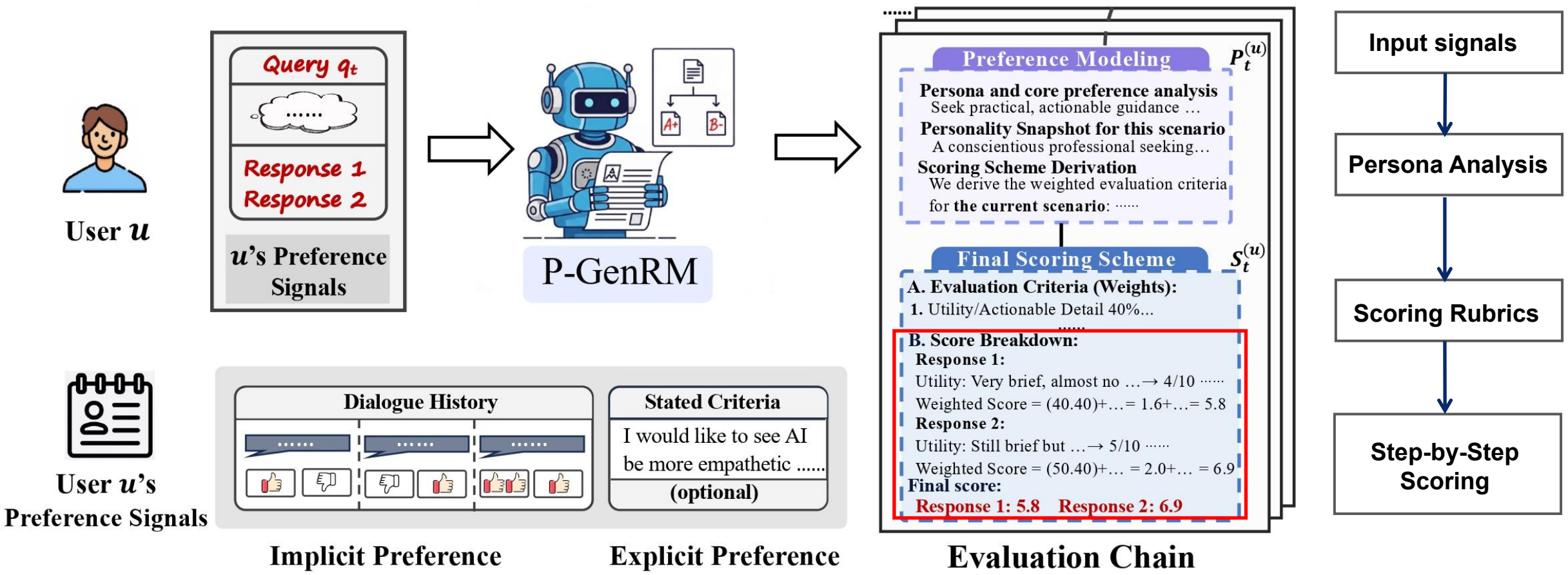
- From hybrid preference signals to structured, dynamic evaluation chain
- From opaque scalars to interpretable, auditable rewards

Structured, Dynamic Evaluation Chain



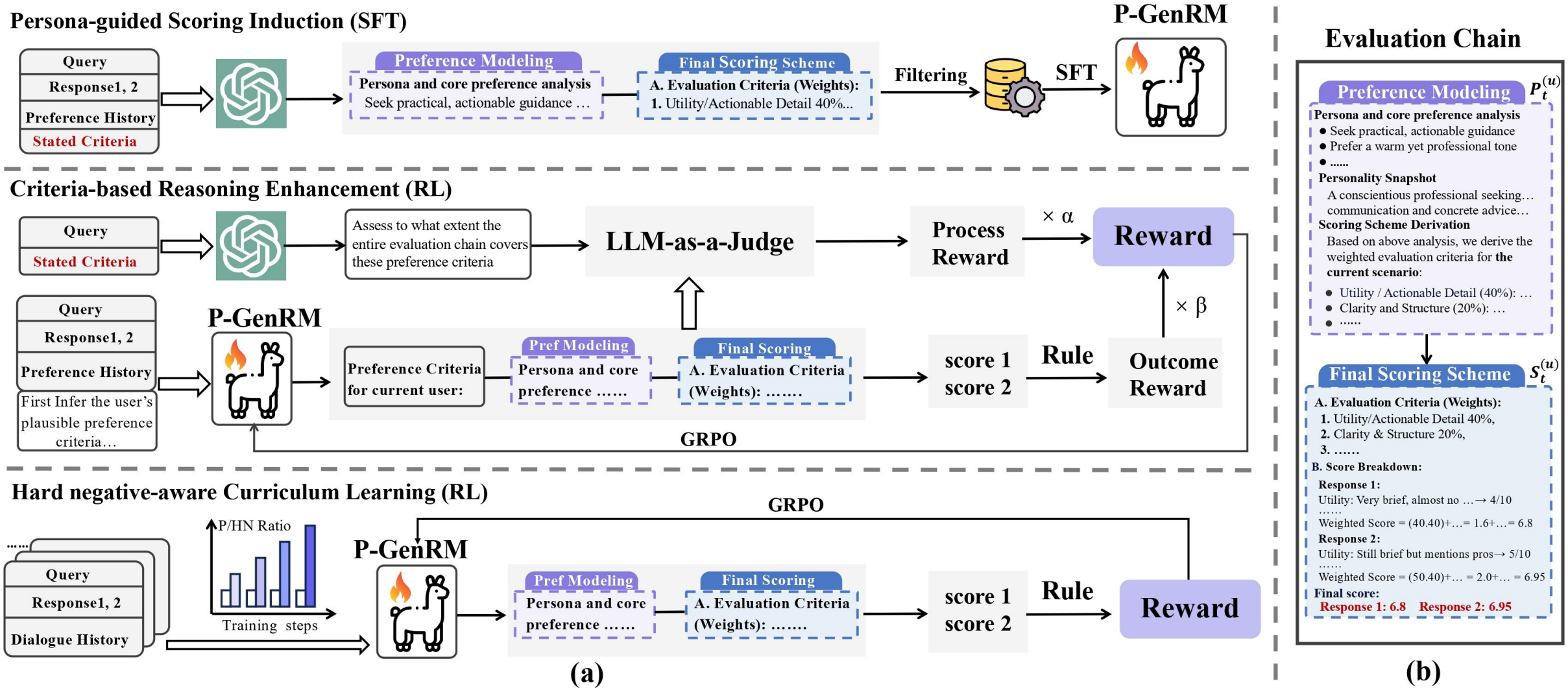
- From hybrid preference signals to structured, dynamic evaluation chain
- From opaque scalars to interpretable, auditable rewards

Structured, Dynamic Evaluation Chain

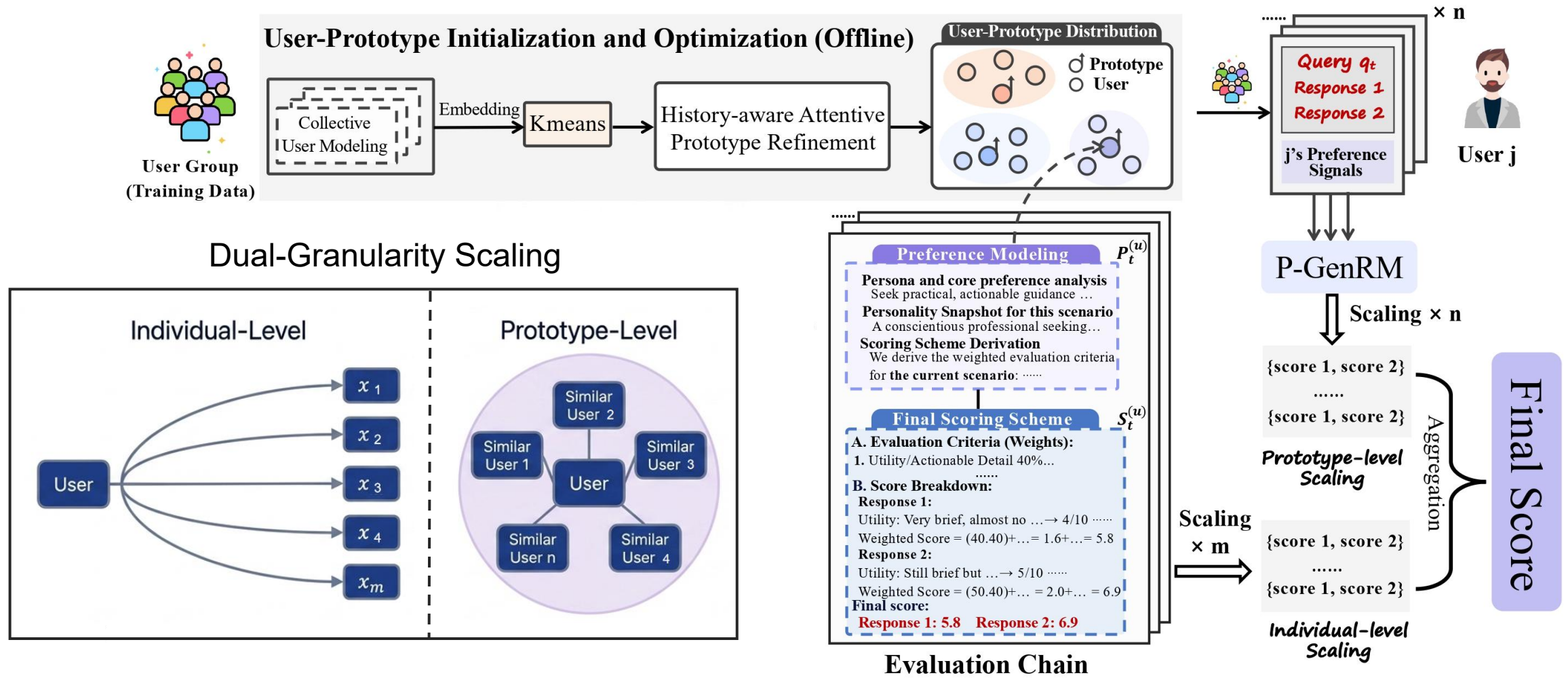


- From hybrid preference signals to structured, dynamic evaluation chain
- From opaque scalars to interpretable, auditable rewards

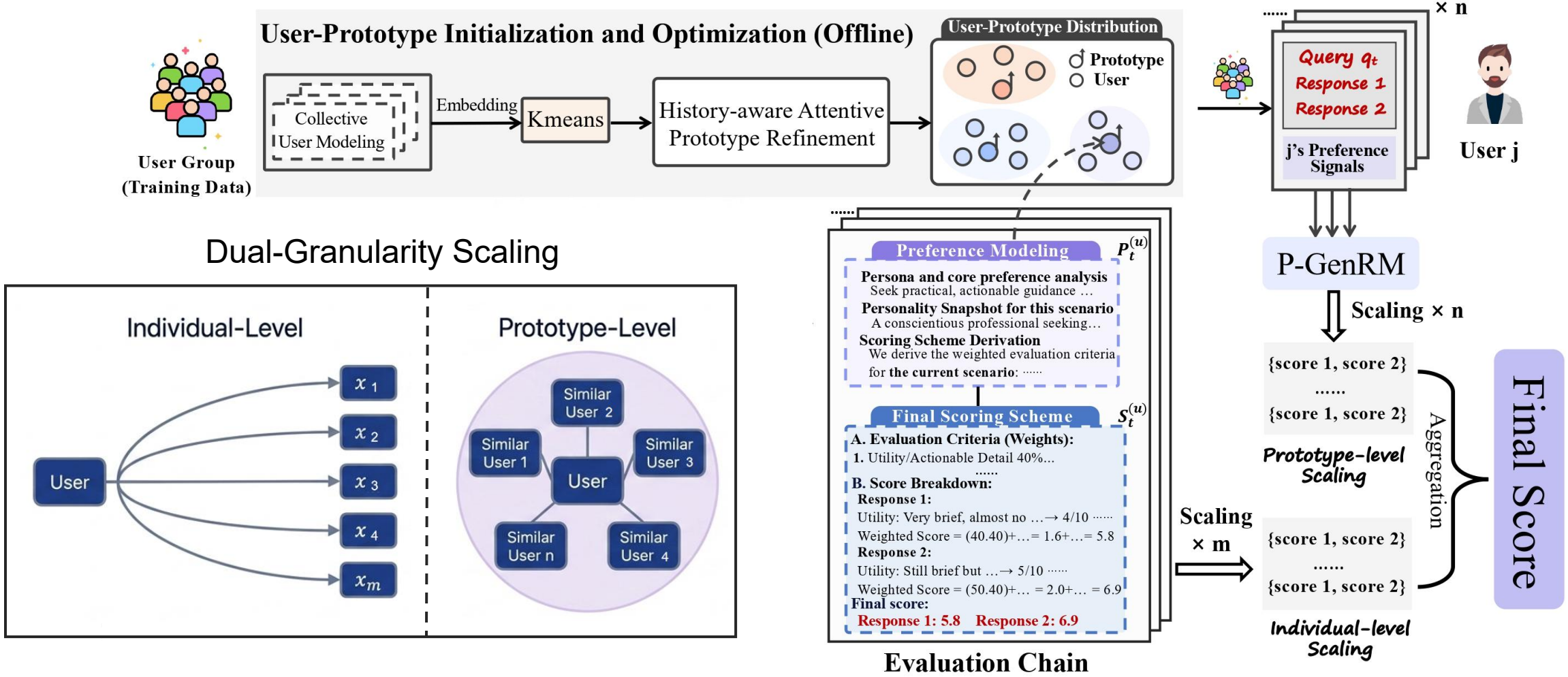
Three stage training framework



Test-time User-based Scaling



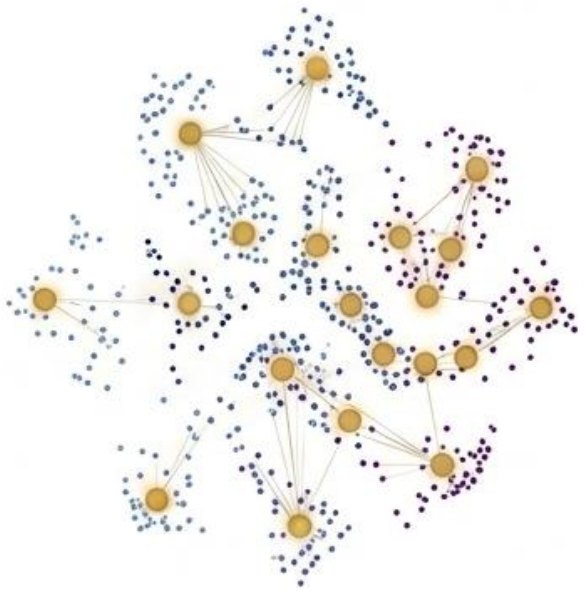
Test-time User-based Scaling



- Parallel sampling to explore multiple hypothesis of a user's preference
- Incorporating similar users' scoring schemes to enhance both accuracy and generalization

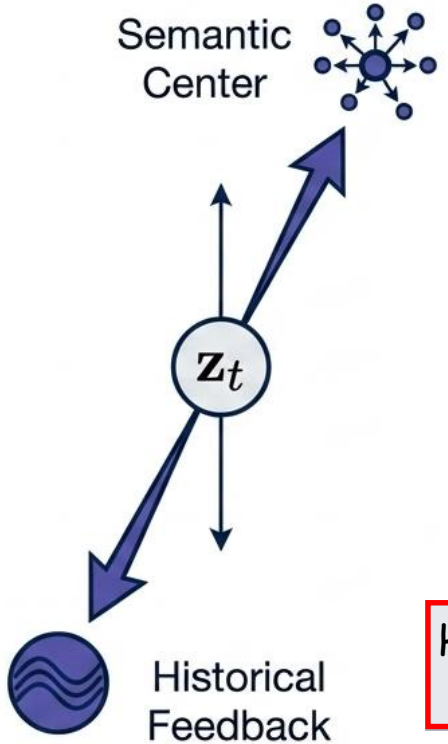
Prototype Initialization and Optimization

Initialization



K-means clustering to establish base user prototypes.

Optimization



Transforming semantic centers into discriminative priors

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \lambda_{\text{cent}} \|\mathbf{a}_j - \boldsymbol{\mu}_j\|_2^2 + \lambda_{\text{tr}} \|\mathbf{a}_j - \mathbf{p}_j\|_2^2$$

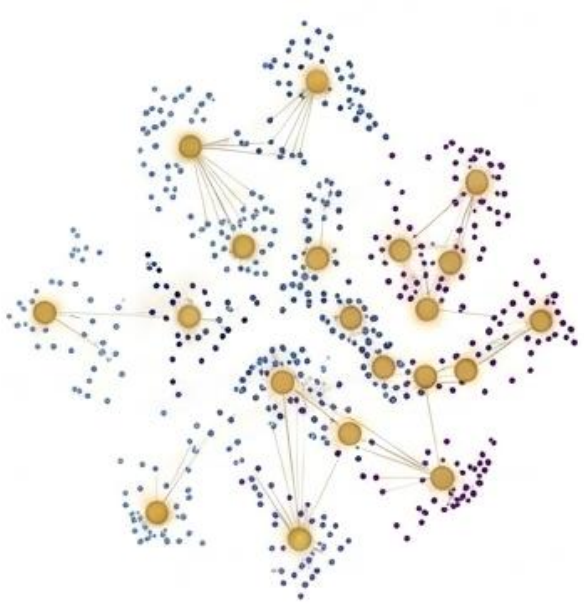
History-aware Loss

Center Regularization

Transition smoothness

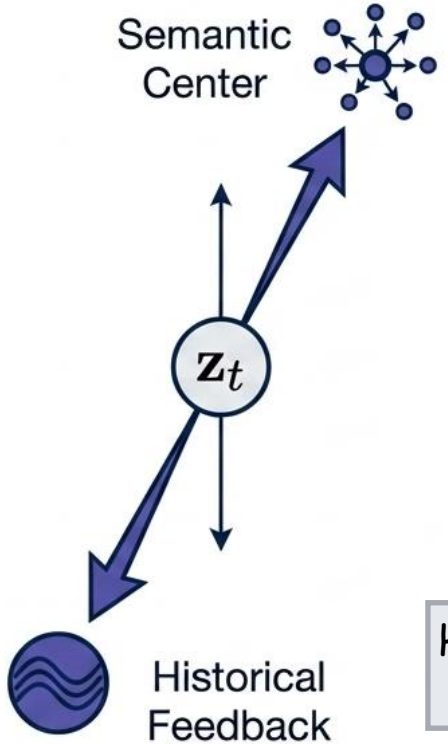
Prototype Initialization and Optimization

Initialization



K-means clustering to establish base user prototypes.

Optimization



Transforming semantic centers into discriminative priors

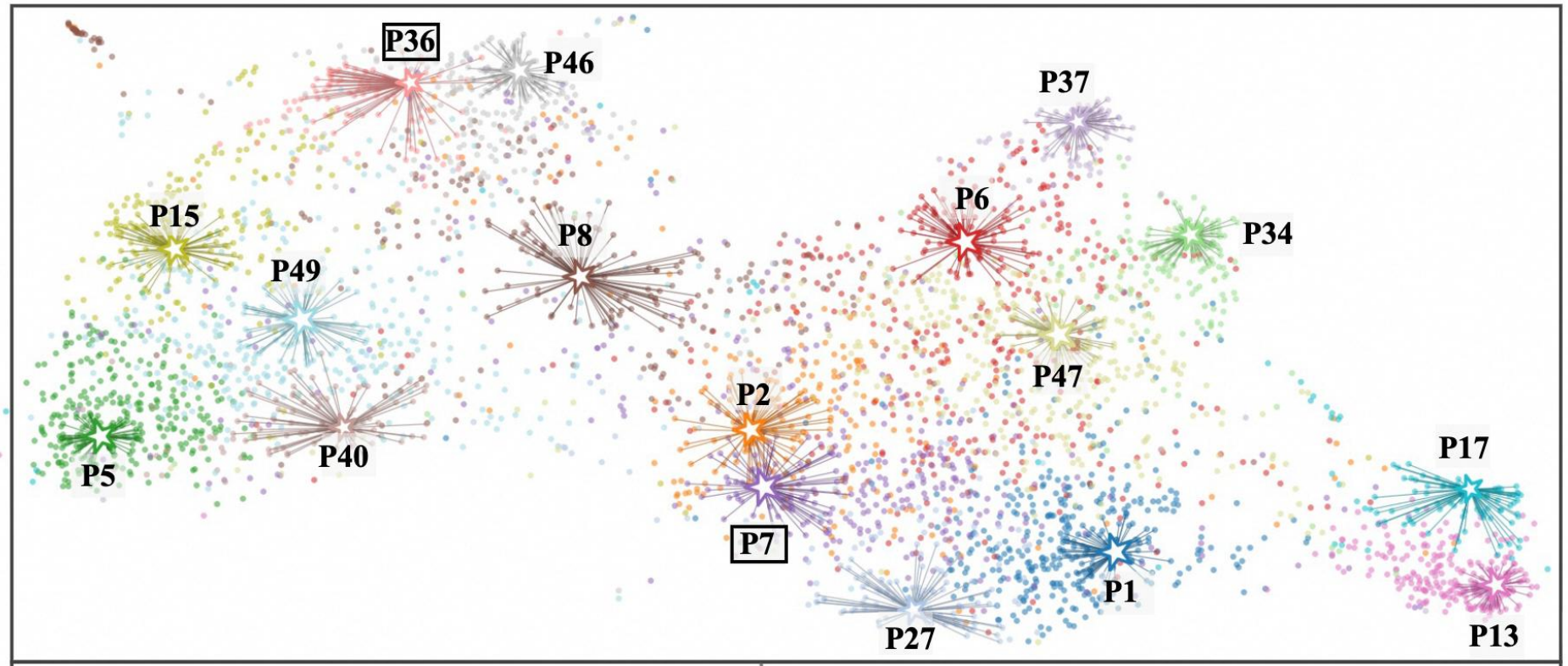
$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \lambda_{\text{cent}} \|\mathbf{a}_j - \boldsymbol{\mu}_j\|_2^2 + \lambda_{\text{tr}} \|\mathbf{a}_j - \mathbf{p}_j\|_2^2$$

History-aware Loss

Center Regularization

Transition smoothness

User-Prototype Distribution



P7

factually correct… and well-develop…
Fluent and easy to read…
Respectful and balanced in values language; dislike forced cultural slang ...
Ethically safe (contain disclaimers when violence/illegality is discussed)
Nuance & Creativity…

---

immediately deliver substantive, factual information…
Fluency & Clarity – writing quality is expected…
go beyond a mere disclaimer and actually outline key points or …
appreciate neutrality ….
Safety/harmlessness must be preserved…

P36

Accuracy & relevance are paramount
Conciseness is valued; chosen answers tend to be tighter …
Clear structure (often bullet points) and smooth flow are liked;
Courteous, moderately formal, and empathetic…
A small amount of engagement (asking follow-up questions) is a plus but …

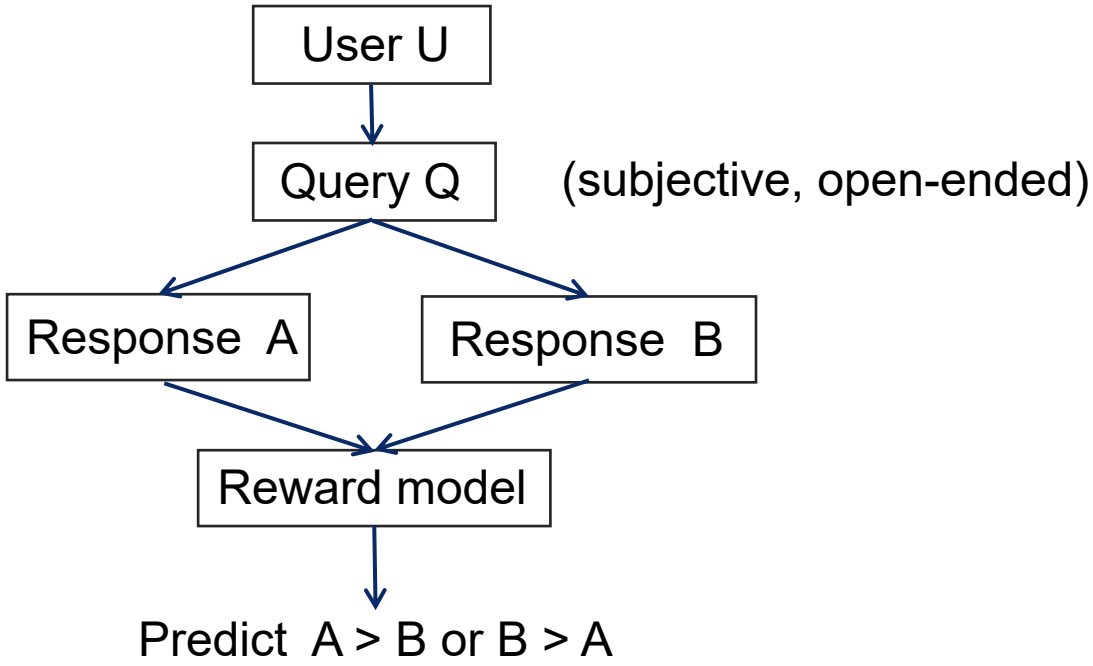
Helpfulness & Factual Accuracy
They value brevity and clarity… consistently pick the more concise answer
Bullet-point structure or crisp sentences are preferred
friendly… a light, informal touch is acceptable and even welcome
Professional/Friendly Tone – respectful, mildly upbeat

Datasets

PersonaRewardBench:

It filters and processes high-quality pairwise preference data (under presonalized scenarios) from Chatbot Arena and PRISM.

The task is to predict which of the two candidate responses better aligns with the individual personal preference of a specific user.



Experiments

Base Model	Chatbot Arena-Personalized		PRISM-Personalized	
	Llama 3.1 8B	Llama 3.1 70B	Llama 3.1 8B	Llama 3.1 70B
In-Context LLM as a Judge				
Default	56.37 ± 2.16%	57.02 ± 2.06%	52.04 ± 0.54%	54.02 ± 0.83%
+ CoT	57.05 ± 2.05%	57.61 ± 1.82%	52.58 ± 0.85%	54.09 ± 0.64%
+ Demographics	—	—	52.96 ± 0.59%	54.21 ± 0.56%
+ Preference History	58.53 ± 1.45%	58.65 ± 2.10%	56.24 ± 0.48%	57.11 ± 0.76%
+ SynthesizeMe	61.07 ± 2.01%	63.14 ± 1.91%	54.70 ± 0.87%	58.19 ± 0.61%
+ Persona-guided Scoring Induction (Ours)	62.20 ± 1.41%	65.55 ± 1.64%	58.33 ± 0.51%	61.61 ± 0.72%
Finetuned Reward Models				
Bradley-Terry				
Finetuned Reward Model	67.21 ± 2.07%	71.12 ± 1.71%	63.27 ± 0.62%	63.44 ± 0.77%
Existing Personalized Reward Model				
GPO	57.87 ± 2.20%	58.50 ± 2.37%	57.29 ± 1.06%	59.16 ± 1.25%
VPL	58.12 ± 2.64%	59.02 ± 2.08%	58.25 ± 0.68%	59.70 ± 1.10%
PAL	57.31 ± 2.49%	59.40 ± 2.69%	56.74 ± 1.18%	57.75 ± 0.83%
FT RM + SynthesizeMe	69.78 ± 1.98%	72.05 ± 2.24%	62.84 ± 0.85%	63.74 ± 0.66%
Personalized Generative Reward Model				
P–GenRM	**72.68** ± 1.85%	**73.42** ± 1.74%	**65.32** ± 0.56%	**66.21** ± 0.76%
Test-time User-based Scaling				
+ Ind-8, Pro-4	74.30 ± 1.60%	—	67.54 ± 0.58%	—
+ Ind-16, Pro-8	**75.92** ± 1.70%	—	**68.06** ± 0.69%	—

- Even without scaling, P-GenRM-8B outperforms 70B-sized baselines.
- Test-time User-based Scaling yields an additional improvement of ~3%.

Experiments

Model	Chatbot Arena	PRISM
Proprietary Model		
o3	64.47 ± 1.62%	56.34 ± 0.64%
o3 + PSI	69.14 ± 1.46%	63.87 ± 0.85%
P–GenRM (8B)	72.68 ± 1.85%	65.32 ± 0.56%
+ Ind-8	73.61 ± 1.54%	65.79 ± 0.68%
+ Ind-4 , Pro-4	73.66 ± 1.39%	66.20 ± 0.75%
+ Ind-16	73.87 ± 1.69%	66.66 ± 0.82%
+ Ind-8 , Pro-4	74.30 ± 1.60%	67.54 ± 0.58%
+ Ind-8 , Pro-8	74.89 ± 1.75%	67.44 ± 0.84%
+ Ind-32	75.59 ± 1.64%	67.65 ± 0.66%
+ Ind-16 , Pro-8	**75.92** ± 1.70%	**68.06** ± 0.69%
+ Ind-0 , Pro-8	*66.90* ± 1.54%	*57.65* ± 0.86%
+ Ind-16 , Pro-16	*72.59* ± 1.61%	*64.61* ± 0.72%

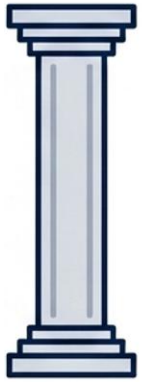
Policy Model	Mean	SE	95% CI
Llama3.1-8B-Instruct	2.954	0.0074	[2.939 , 2.969]
Qwen2.5-7B-Instruct	2.970	0.0089	[2.952 , 2.988]
Llama3.1-70B-Instruct	3.156	0.0093	[3.138 , 3.174]
Qwen2.5-72B-Instruct	3.214	0.0089	[3.192 , 3.228]
Llama3.1-8B-Instruct-DPO	3.316	0.0068	[3.303 , 3.329]
Llama3.1-8B-Instruct-GRPO	3.354	0.0102	[3.334 , 3.374]

Reward Model	Arts	Pers.	Soc.	Avg.
Qwen3–8B	0.486	0.543	0.600	0.543
Qwen3–32B	0.543	0.600	0.543	0.562
Qwen3–235B–A22B	0.600	0.657	0.600	0.619
LLaMA3.1–8B	0.486	0.543	0.543	0.524
SynthMe–8B	0.486	0.657	0.600	0.581
LLaMA3.1–70B	0.543	0.657	0.600	0.600
P-GenRM (8B) + Ind-8, Pro-4	0.543	0.714	0.657	**0.638**

Left : Efficiency of dual-granularity scaling: Improving P-GenRM accuracy with modest increases in scaling operations.

Right : P-GenRM's effectiveness in policy model training and generalization on the OOD dataset, LaMP-QA.

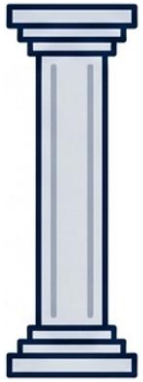
Conclusion



Interpretable

From black-box rewards to

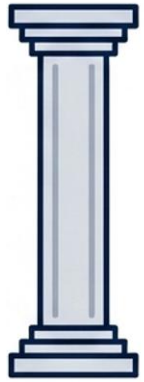
Structured evaluation chains



Adaptable

From static rules to

Dynamic reward modeling



Transferable

Test-time User-based Scaling

alleviates the cold-start problem.

Building AI that understands not just what people say, but what they uniquely value.

Tongyi