Data-Centric Model Adaptation

- **Name**: Jianqing Zhang
- **2-year Ph.D.**: Shanghai Jiao Tong & Tsinghua University
- **Visiting**: Hong Kong Polytechnic University (PolyU)
- **Collaborations**:
  - Qiang Yang, HKUST & PolyU, China
  - Yang Hua, Queen's University Belfast, UK
  - Yang Liu, PolyU, China
  - Marco Canini, KAUST, Saudi Arabia
  - Han Yu, NTU, Singapore
  - Yuyin Zhou, UC Santa Cruz, USA

Overview

- **Research interests**: **Data-Centric Model Adaptation**
- **Research fields**: *Code LLM RL, Synthetic Data Generation, Federated Learning, Recommender System*
- **Open-sourced projects** (initiator):
  - EvolveGen, PFLlib (**1,900**+ stars, **300**+ forks), HtFLlib, HtFLlib on Device, FL-IoT, etc.
- **Publications**: **9** first-author top-tier papers
  - AAAI'23 (**oral, Top 4%**), KDD'23, ICCV'23, NeurIPS'23, JMLR'25 (**WAIC Outstanding Paper, Top 40**)
  - AAAI'24, CVPR'24, KDD'25 (**Best Paper, Top 3**)
  - EMNLP'24, ICML'25, ICML'25 (**spotlight, Top 2%**)
- **Strengths: passion, research–application synergy, field insight, and sociability**
- **Awards**: Youth Talent of China Association for Science and Technology (China Association for Artificial Intelligence, CAAI), Wenjun Wu Honorary Doctorate in AI, PhD National Scholarship, etc.
- **Leadership**: ① Led 9-member team in building a Federated ML platform for data centers, ② Cross-hospital cancer recognition model, ③ Cross-province intelligent 12345 hotline model, etc.
- **Impact**: **900**+ citations, **30K**+ views across major media, well-recognized by IEEE/ACM Fellows
- **Intern**: ByteDance AML, Tsinghua AIR, KAUST SANDs lab, Tencent CodeBuddy

Systematical Research Trace

Data-Centric Model Adaptation: Balancing generalization and specialization on new data from a data-driven perspective

① **Recommender:** **Exploiting** long- and short-term user behavior data

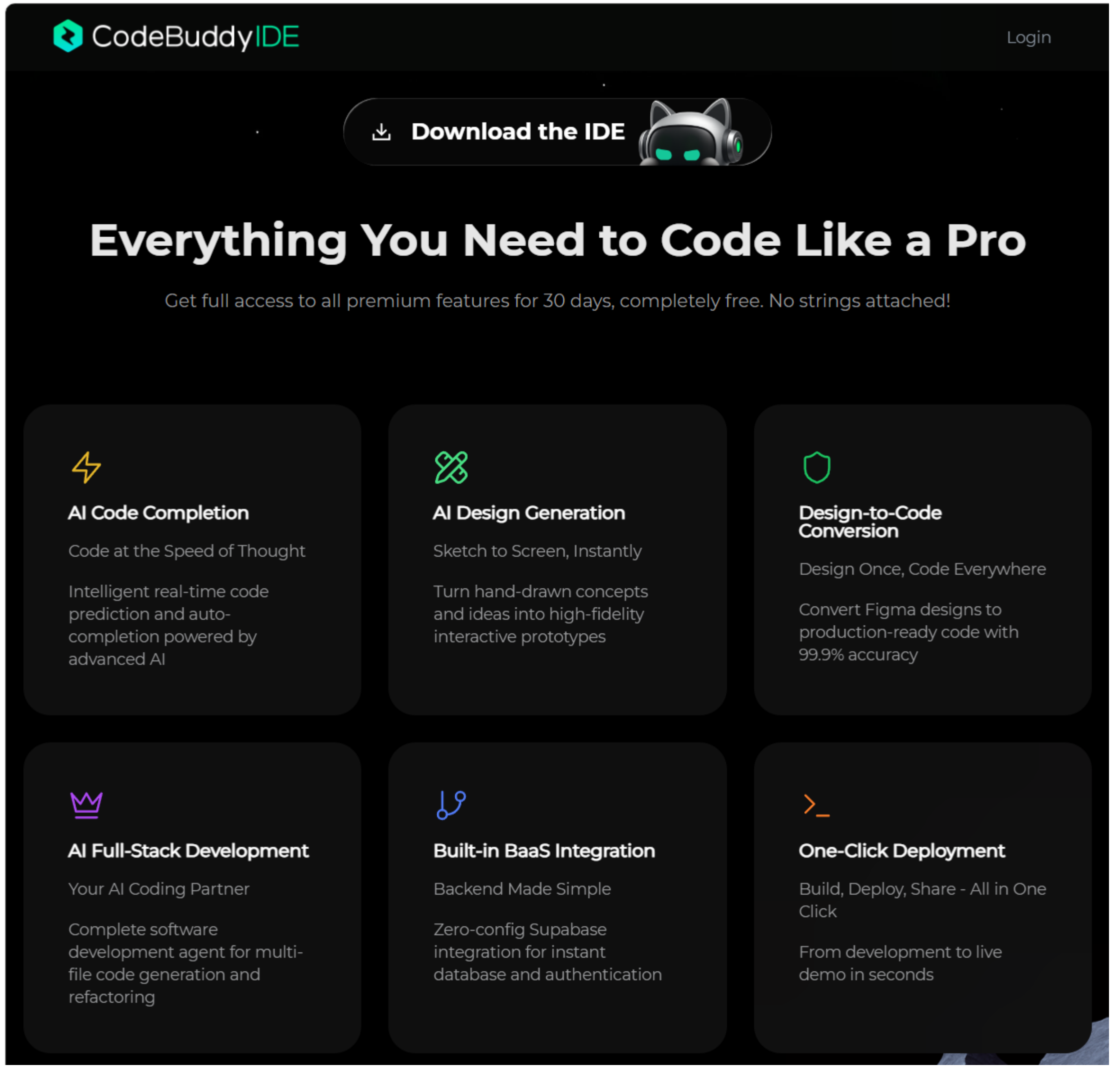
②**Federated learning:** **Exploiting** generalization and specialization in heterogeneous data across devices

③**Synthetic dataset generation:** **Generating** vertical domain data based on few private data and large models

④**Code LLM reinforcement learning:** **Exploiting** self-**generated** data to reduce coding errors

Code LLM

- **CodeBuddy,** Cursor, Claude Code, Trae, etc.
- Core lies in the <u>**large model**</u> behind it
- Object:
  - Code Edit Model
  - Code Agent
- Target:
  - Efficiency (less token/data)
  - Quality (less error)
- Internship work:
  - Code Generation
  - Code Edit
- Achievement: **patent +3, paper +3**
  - RLVR, RLHF

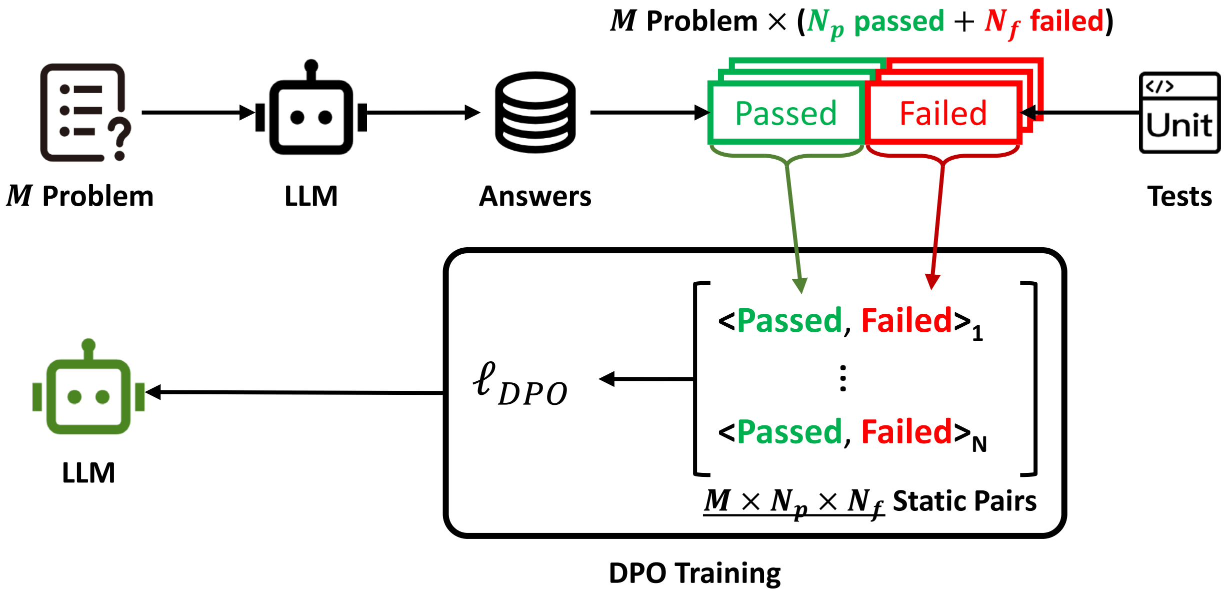


[Code]: AP2O

- AP2O: Adaptive Progressive Preference Optimization
  - **Problem:** LLM-generated code has **compilation or runtime errors**
  - **Goal:** Reduce code errors, improve **code quality** & **pass rate** & **accept rate**
  - **Solution:** Progressive preference optimization + adaptive error replay
  - **Results**: Improve **pass@k** by up to **3%** for **0.5B~34B LLM** on *EvalPlus*, *LiveCodeBench*, etc.
    - Qwen2.5-Coder, CodeLlama, DeepSeek-Coder, Qwen2.5, Llama3, Qwen3
  - **Bonus**: Reduce **data requirement** greatly (only using 4%~60% data)
  - **Findings**:
    - **Poor models**: Progressing from **low to high** error frequency
    - **Strong models**: Progressing from **high to low** error frequency

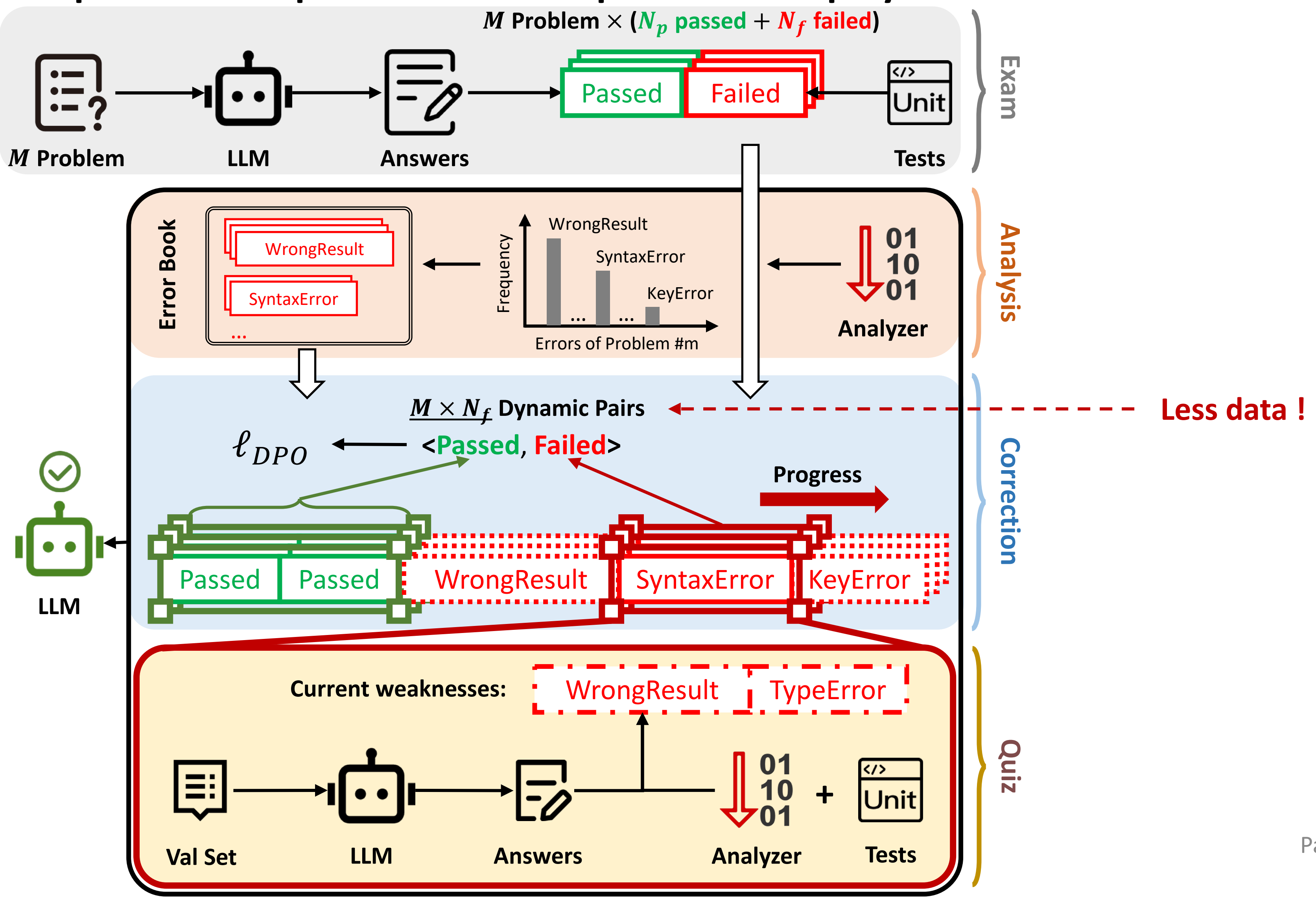
[Code]: Existing problems

- Traditional DPO training for code generation
  - **Problem 1:** Unawareness of **code errors**
  - **Problem 2:** Inability to **focus** on specific error types **(SyntaxError, TypeError, etc.)**



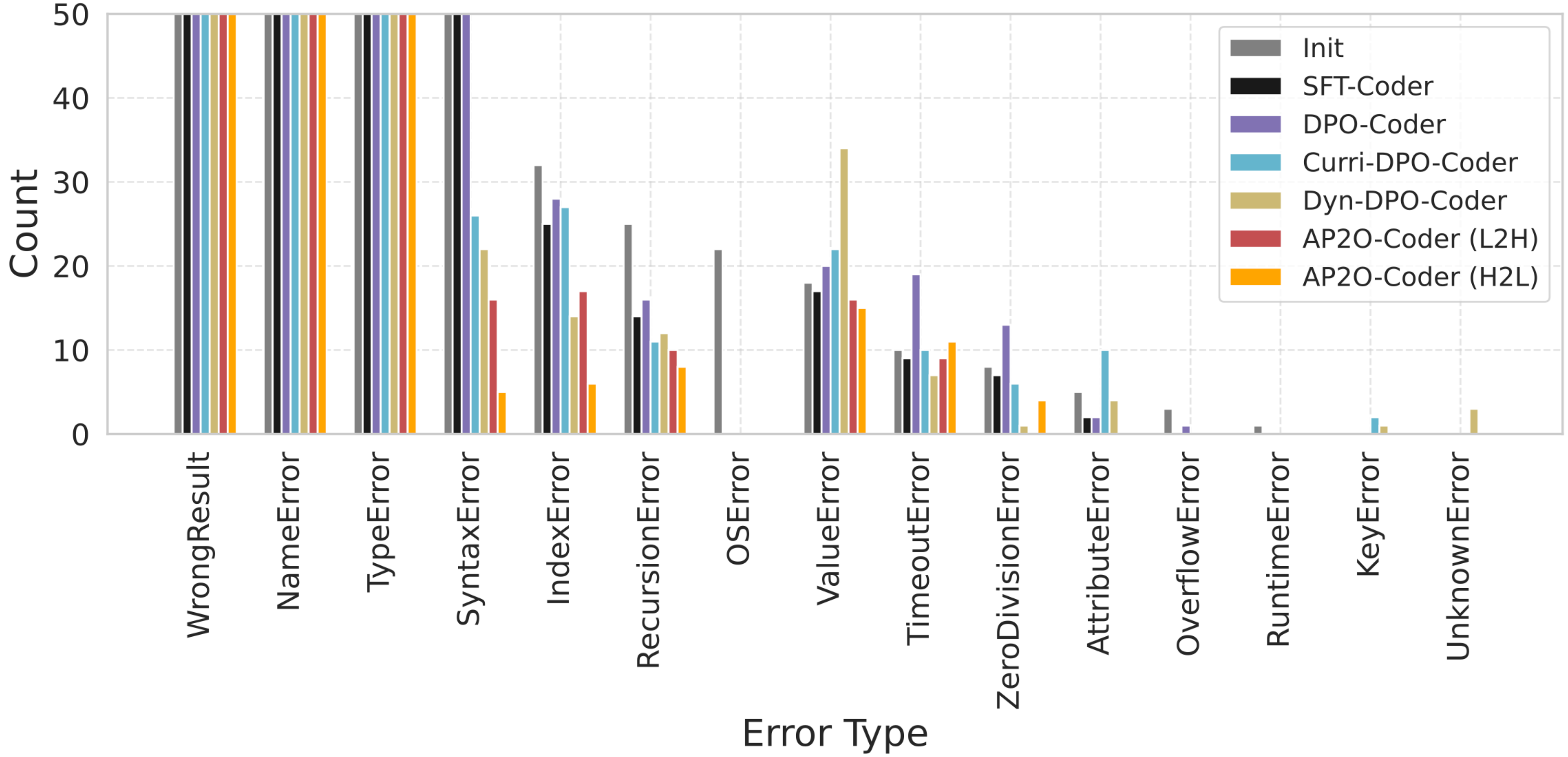
[Code]: AP2O

- **Solution: Progressive preference optimization + adaptive error replay**



[Code]: AP2O

- AP2O reduces coding **errors**



Outperform baselines by up to 3% in pass@k

[Code]: AP2O

- AP2O (**H2L**) is better for large models.
  - **Concentration** then **exploration** fits **mature LLMs**
  - H: LLM sees **identical error types** in adjacent updating steps - **specification**
  - L: LLM sees **various error types** in adjacent updating steps - **generalization**

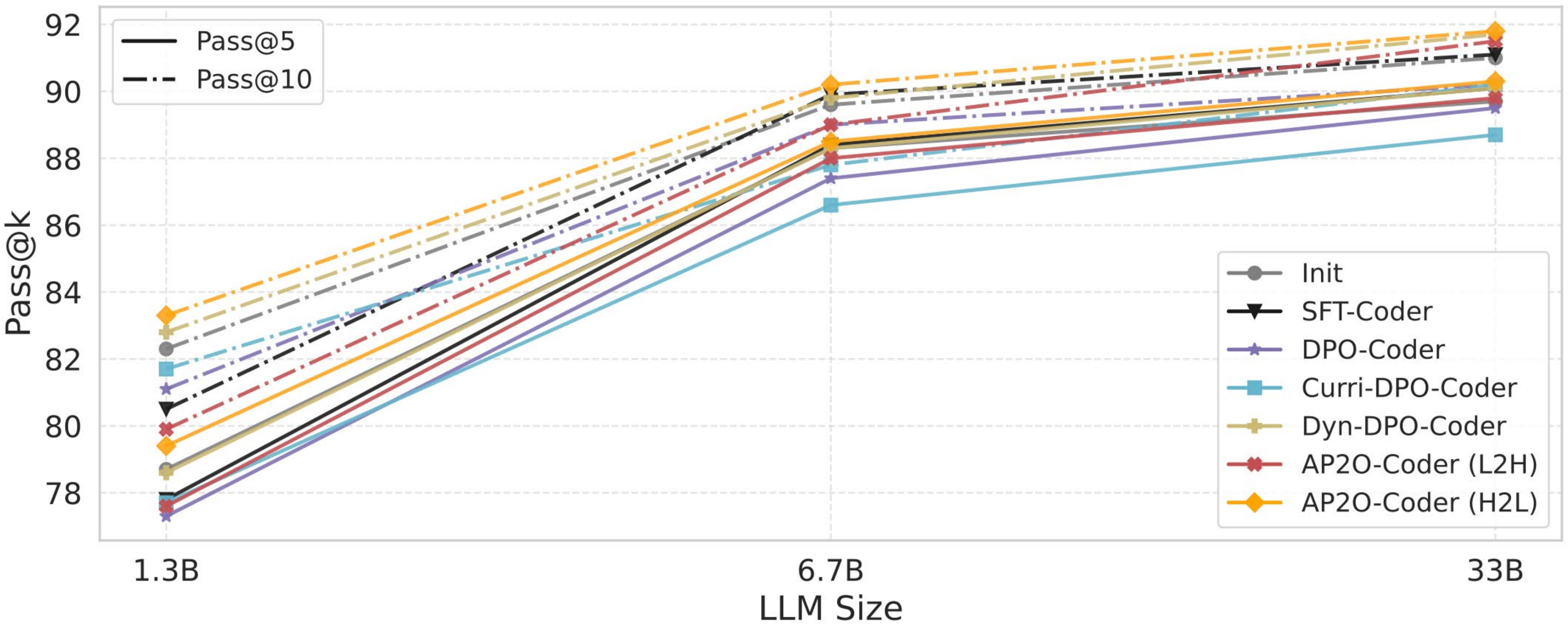
LLM Type	CodeLlama			DeepSeek-Coder			Qwen2.5-Coder					
LLM Size	7B	13B	34B	1.3B	6.7B	33B	0.5B	1.5B	3B	7B	14B	32B
Init	36.8	41.3	46.2	64.6	77.4	78.4	53.0	69.3	83.5	87.1	90.4	91.5
SFT-Coder	37.9	43.2	46.8	64.8	75.9	78.9	60.1	70.4	85.1	87.4	90.7	90.9
DPO-Coder	38.3	42.3	45.2	63.5	77.2	78.7	56.8	73.2	84.5	87.9	90.8	91.0
Curri-DPO-Coder	38.7	42.4	46.5	63.8	76.6	79.2	53.3	73.1	83.7	87.2	90.2	90.8
Dyn-DPO-Coder	38.6	42.3	44.9	63.4	76.2	78.8	57.1	71.5	84.7	87.6	90.7	91.6
`AP2O-Coder` (L2H)	**39.8**	43.1	47.9	**65.9**	77.6	79.1	**61.5**	**76.3**	85.7	88.1	90.8	91.8
`AP2O-Coder` (H2L)	38.9	**44.5**	**49.6**	64.7	**78.8**	**80.1**	56.5	71.7	**86.3**	**88.9**	**91.4**	**92.2**

Table 1: The $pass@1$ on EvalPlus (HumanEval) across various types and sizes of code LLMs.

Outperform baselines by up to 3% in pass@k

[Code]: AP2O

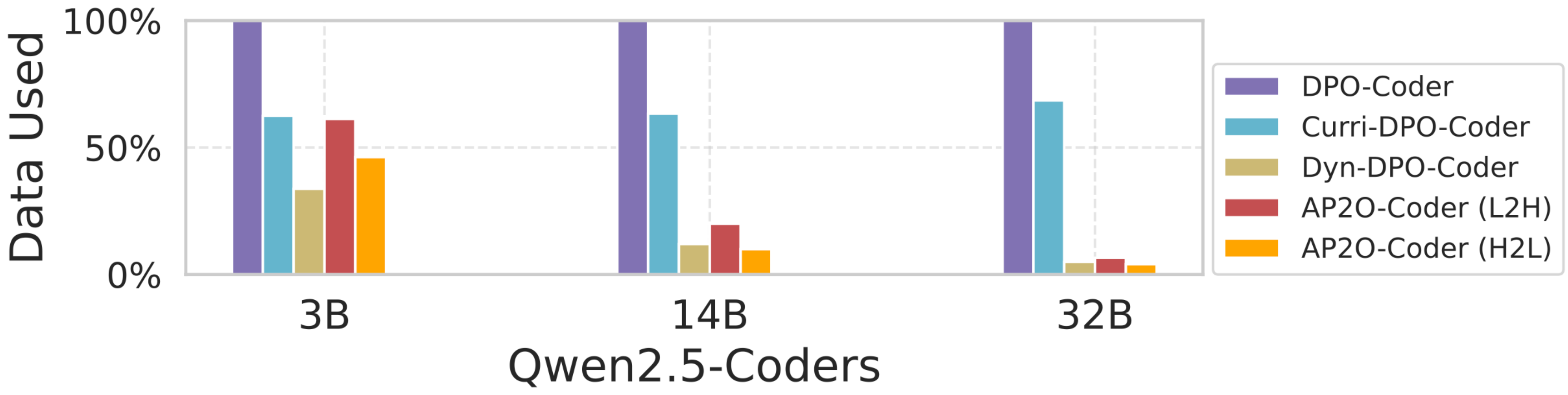
- AP2O (**H2L**) also improves **generalization**



Outperform baselines by up to 3% in pass@k

[Code]: AP2O

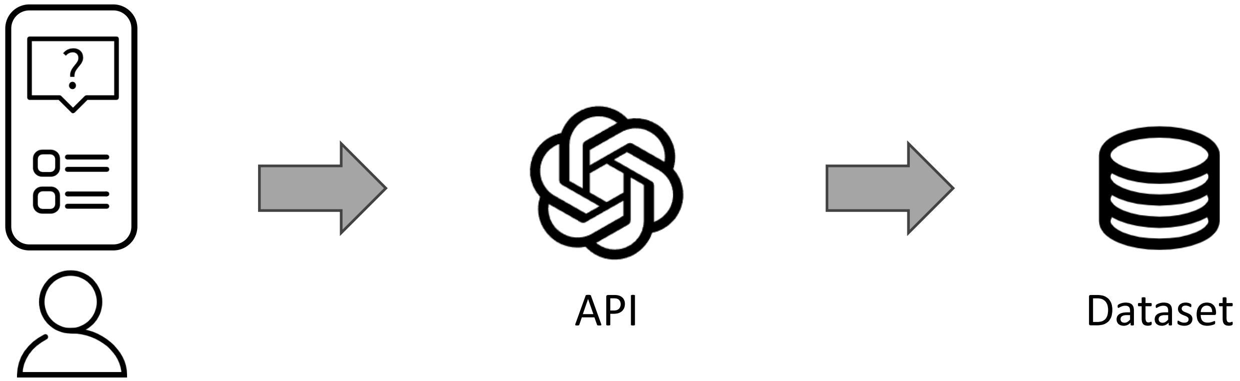
- AP2O is also sample **efficient**



Outperform baselines by up to 3% in pass@k

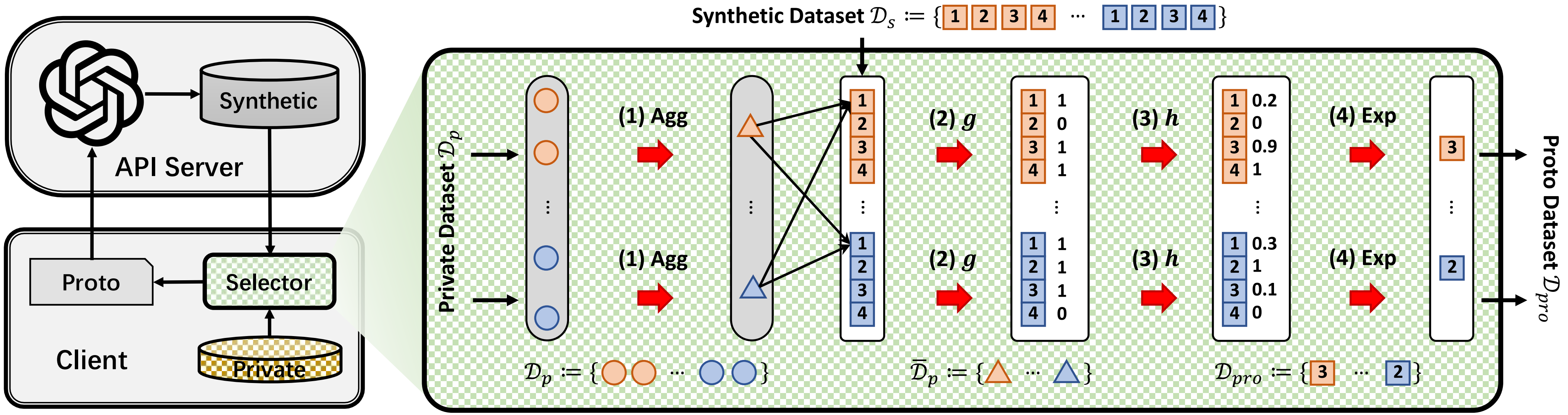
Synthetic Data Generation (SDG)

- Given a **prompt**, with or without **data examples**,
- AI generates a dataset that **aligns with the user's request**.
    - Focusing on special domains (e.g., code, medicine, industry, etc.)



[SDG]: PCEvolve

- **Solution: Given a few samples** — we'll **evolve** an entire dataset for you, **used in Microsoft**
- While **protecting privacy** via differential privacy (DP) using Exponential mechanism



Accepted by ICML 2025 (spotlight)

[SDG]: PCEvolve

- **COVIDx**: chest X-ray images for COVID-19
- **Came17**: tumor tissue patches from breast cancer metastases
- **KVASIR-f**: endoscopic images for gastrointestinal abnormal findings detection
- **MVAD-l**: leather surface anomaly detection

Top-1 accuracy (%) on four specialized datasets

	COVIDx	Came17	KVASIR-f	MVAD-l
Init	49.34	50.47	33.43	33.33
RF	50.01	54.82	34.66	48.17
GCap	50.86	55.77	32.66	27.33
B	50.42	54.41	32.57	43.21
LE	50.02	55.44	35.51	27.93
DPImg	49.14	61.06	33.35	37.03
PE	59.63	63.66	48.88	57.41
PE-EM	57.60	63.34	43.01	50.06
PCEvolve-GM	56.91	62.63	43.55	55.56
PCEvolve	**64.04**	**69.10**	**50.95**	**59.26**

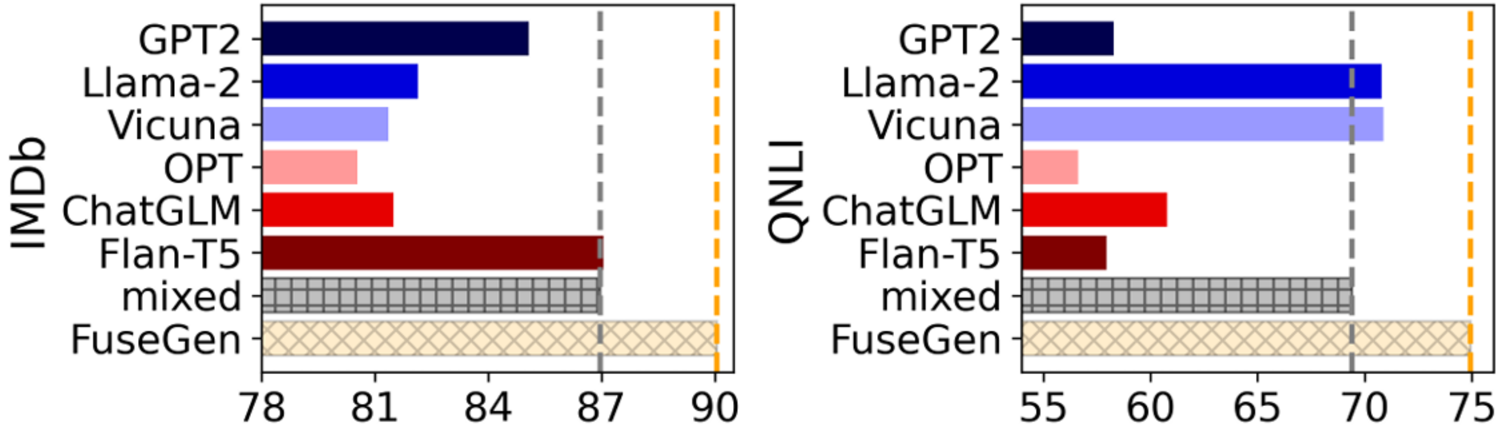
Accepted by ICML 2025 (spotlight)

[SDG]: PCEvolve

Generated leather surface images w.r.t. MVAD-l for industry anomaly detection. The three rows show normal images, cut defects, and droplet defects. “Initial” denotes API-generated images using just the prompt. “Private” denotes the real images from MVAD-l.

[SDG]: FuseGen

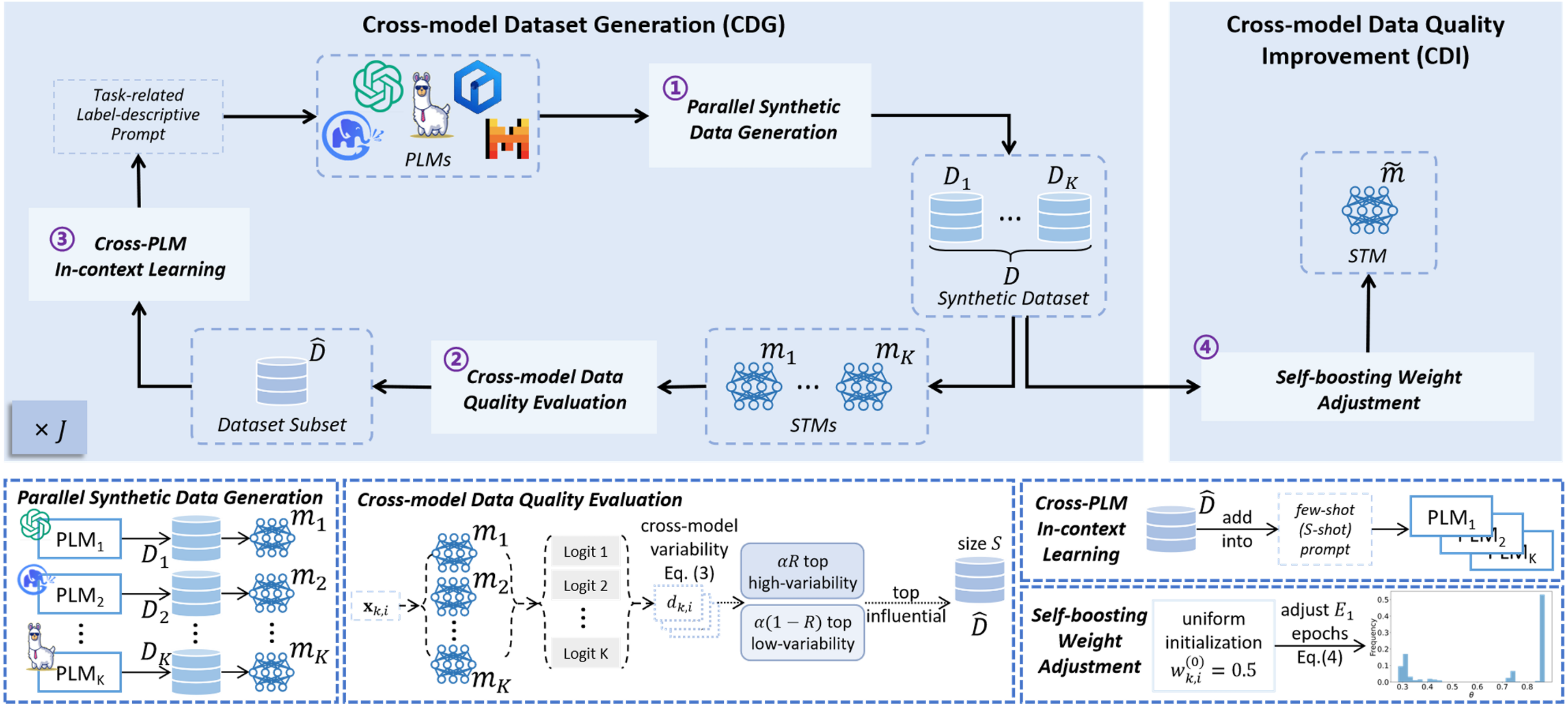
- **Problem:** Pre-trained Language Models (PLMs) have **different tastes** for specific domains
- **Solution:** We **merge models' outputs** to create **diverse datasets** through **evolution**



Accepted by EMNLP 2024

[SDG]: FuseGen

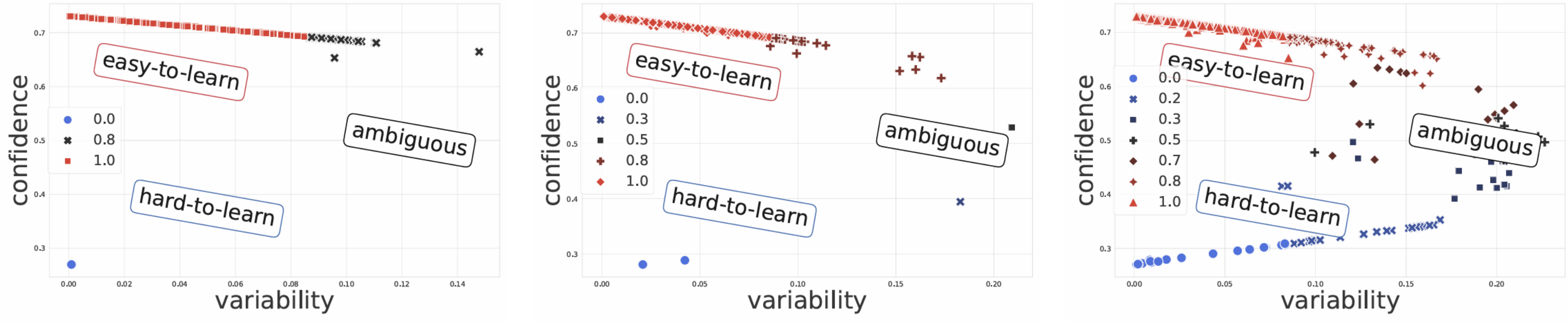
- We consider downstream models' **feedback as reward signals for evolution**



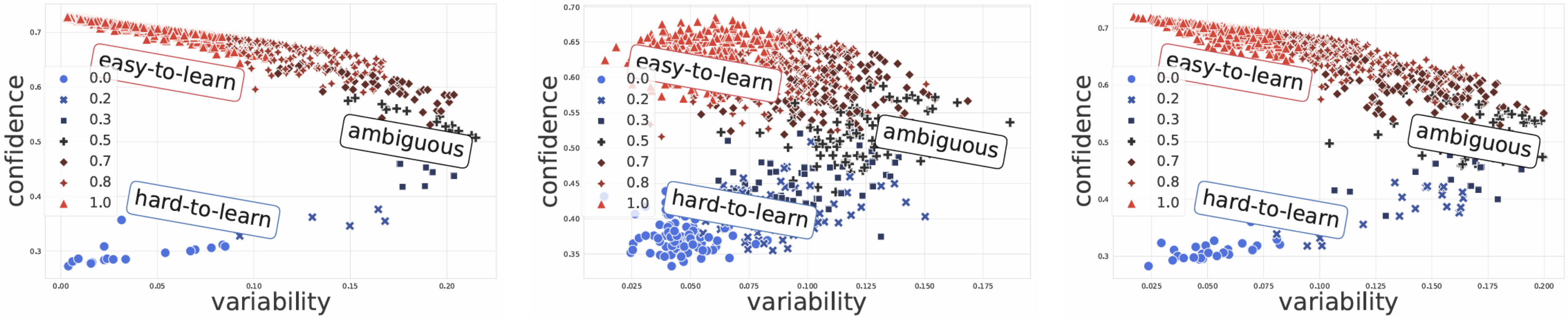
Accepted by EMNLP 2024

[SDG]: FuseGen

- Synthetic **dataset cartography**



(a) Llama-2 ZeroGen $K = 1$ (84.23) (b) Llama-2 ProGen $K = 1$ (84.24) (c) Llama-2 Ours $K = 6$ (86.60)

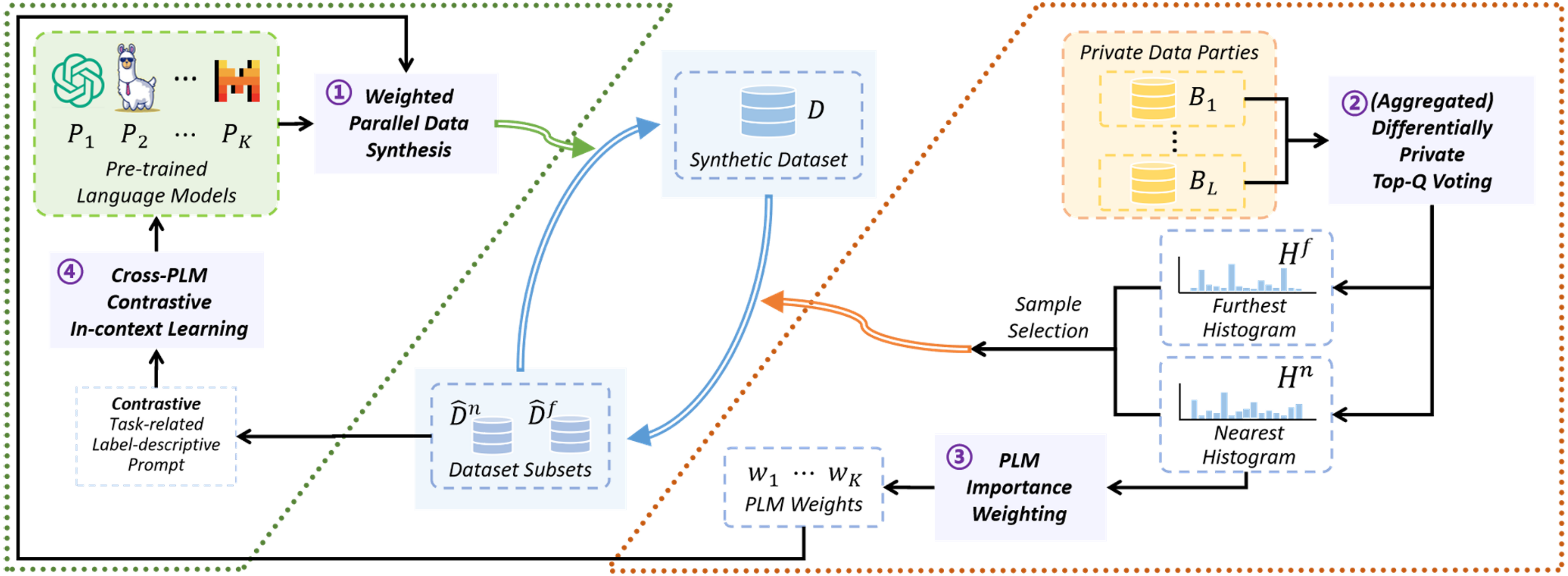


(d) Flan-T5 ZeroGen $K = 1$ (88.18) (e) Flan-T5 ProGen $K = 1$ (85.80) (f) Flan-T5 Ours $K = 6$ (88.73)

Accepted by EMNLP 2024

[SDG]: WASP

- **Problem:** Using only positive examples for evolution lacks diversity for domain dataset generation
- **Solution:** **Contrastive voting**-based positive and negative sample selection for **evolution**



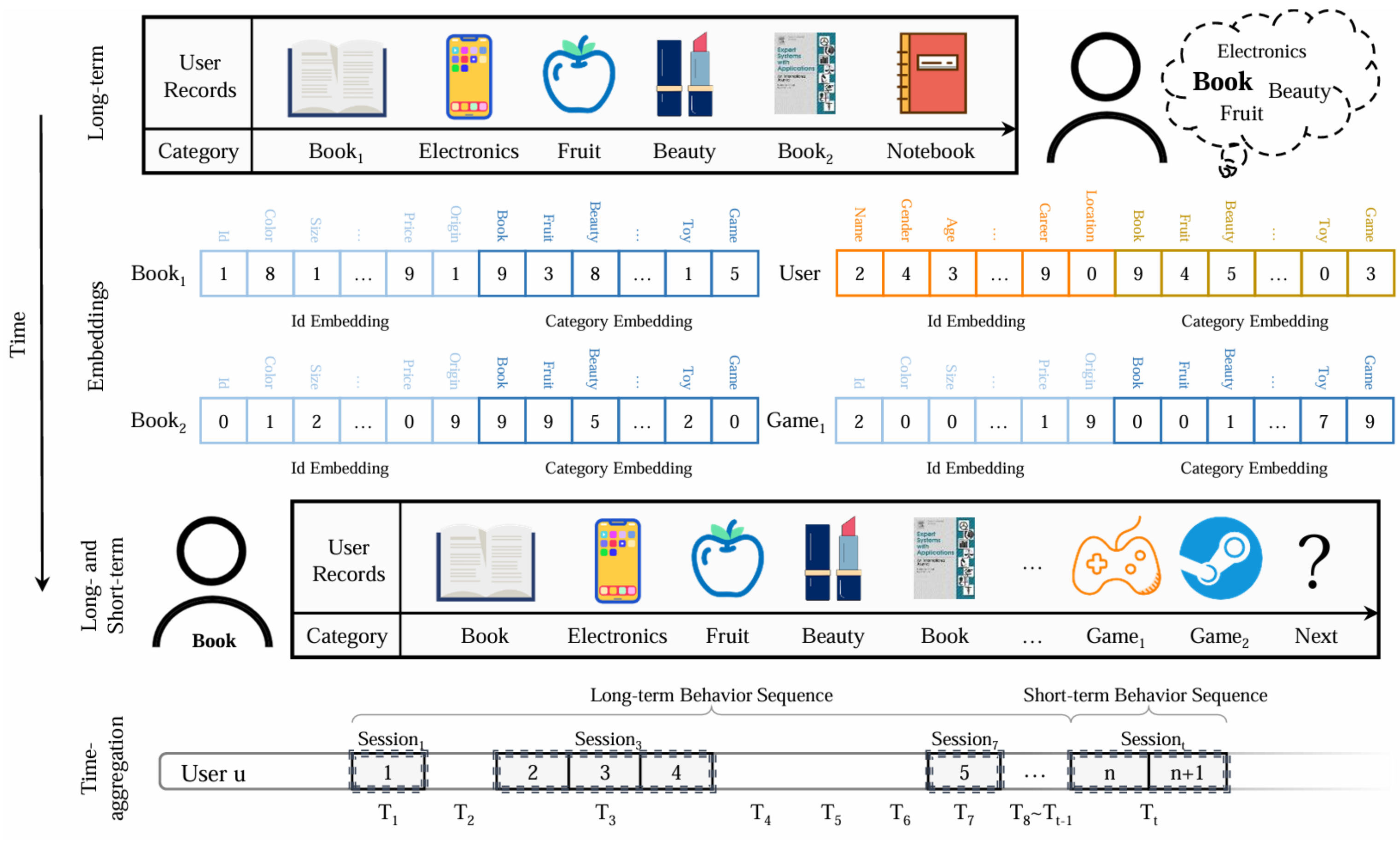
Accepted by ICML 2025

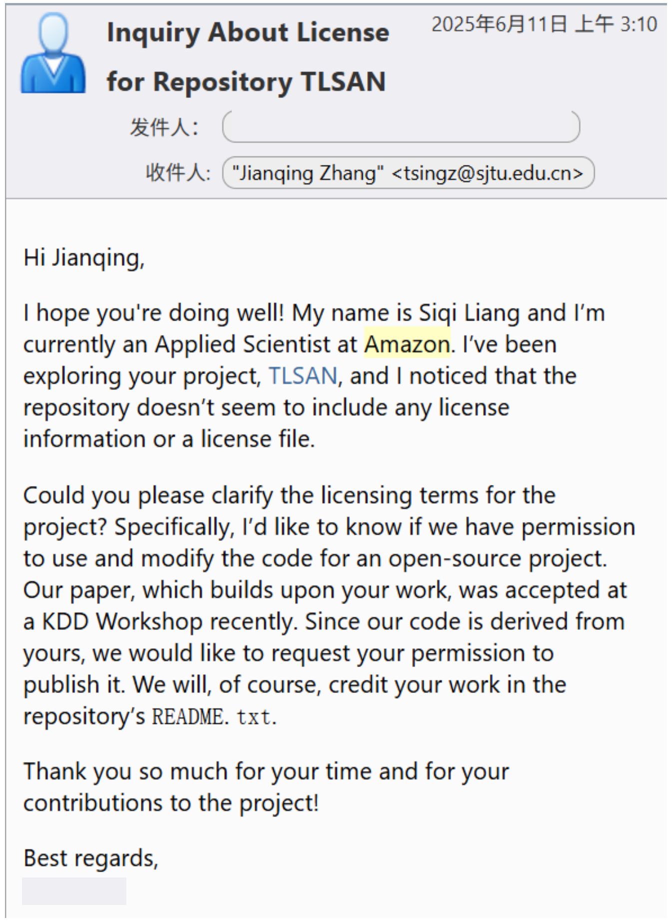
[SDG]: WASP

Model	High-quality Sample Text (Examples)	Label
GPT-2	"This movie isn't for everyone, but it's worth a look. But if you have to be in the dark to enjoy it, just buy the soundtrack."	positive
GPT-2	"is that it has some semi-interesting world-building, but the information it offers is grossly insufficient to allow one to progress in the story. The central characters have little to do with the world beyond playing around with their toys and talking to themselves. "	negative
Llama-2	"Incredible film! I laughed, I cried, and I left the theater feeling inspired. Every performance was top-notch and the direction was flawless. A true masterpiece that will be remembered for years to come. 10/10."	positive
Llama-2	"This film was a colossal disappointment. I found myself glancing at my wrist every 10 minutes, eagerly anticipating its conclusion. The narrative was intricate and challenging to comprehend, and the personas were underdeveloped. I had no emotional investment in their plight. The performances were lacking, and the dialogue was trite and unconvincing. In summary, I would not advise viewing this motion picture to any individual. "	negative
Vicuna	"A movie for people who want to see an action-packed adventure that will keep them on the edge of their seats."	positive
Vicuna	"Felt like a missed opportunity. Glacial pacing and shallow characters. The premise was intriguing but the execution fell flat. Avoid this one."	negative
OPT	"-This film is enjoyable to watch because of it's well-developed storyline and most of all its storyline that i can relate to and would like to have. This film also has good acting performances and the songs help the film a lot. It's not a typical love story kind of a film, but it does have lots of interesting messages to give. Very well-done! Definitely recommend this film!"	positive
OPT	"Quote: The first sentence means nothing to me. Without context the first two sentences also mean nothing."	negative
ChatGLM3	"Attention getter, visually interesting and outstanding acting, the story of an American citizen that is in Mexico and gets involved in a murder is a good movie. "	positive
ChatGLM3	"This movie is an unwatchable mess, basically a rip-off of that scene in Dr. Strangelove where the U.S. Air Force gets into	negative

Recommender System (RS)

- **Problem:** Users have **personalized taste for time** in the behavior data besides others
- **Solution:** Time-aware Long- and Short-term Attention Network, **<u>memory</u>, used in Amazon AWS**





Inquiry About License for Repository TLSAN

2025年6月11日 上午 3:10

发件人:

收件人: "Jianqing Zhang" <tsingz@sjtu.edu.cn>

Hi Jianqing,

I hope you're doing well! My name is Siqi Liang and I'm currently an Applied Scientist at Amazon. I've been exploring your project, TLSAN, and I noticed that the repository doesn't seem to include any license information or a license file.

Could you please clarify the licensing terms for the project? Specifically, I'd like to know if we have permission to use and modify the code for an open-source project. Our paper, which builds upon your work, was accepted at a KDD Workshop recently. Since our code is derived from yours, we would like to request your permission to publish it. We will, of course, credit your work in the repository's README. txt.

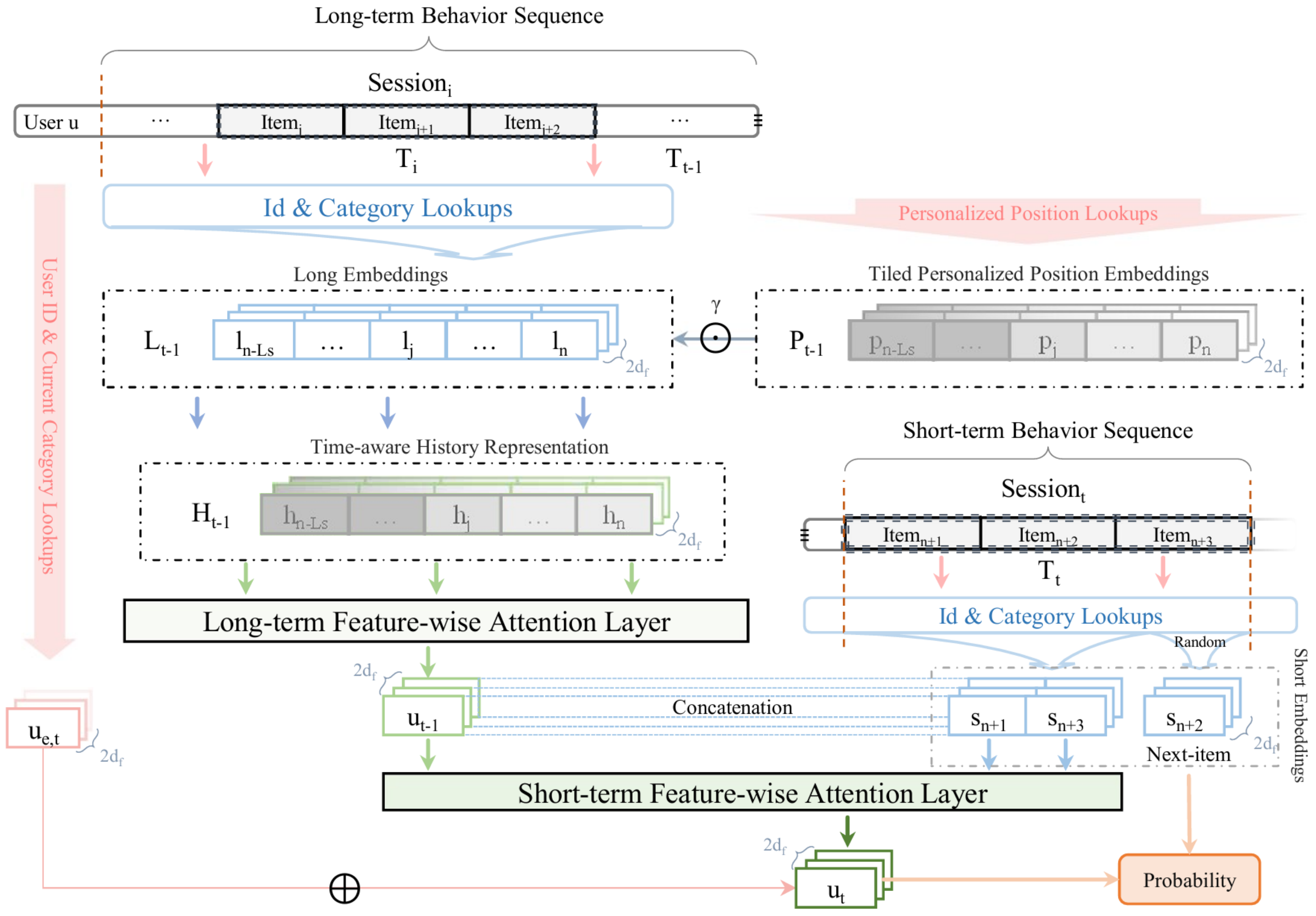
Thank you so much for your time and for your contributions to the project!

Best regards,

Accepted by Neurocomputing 2021 **Outperform baselines by up to 20.79%**

[RS]: TLSAN

- Next-item recommendation, useful for **AI Coding's Next Edit Suggestion (NES)**

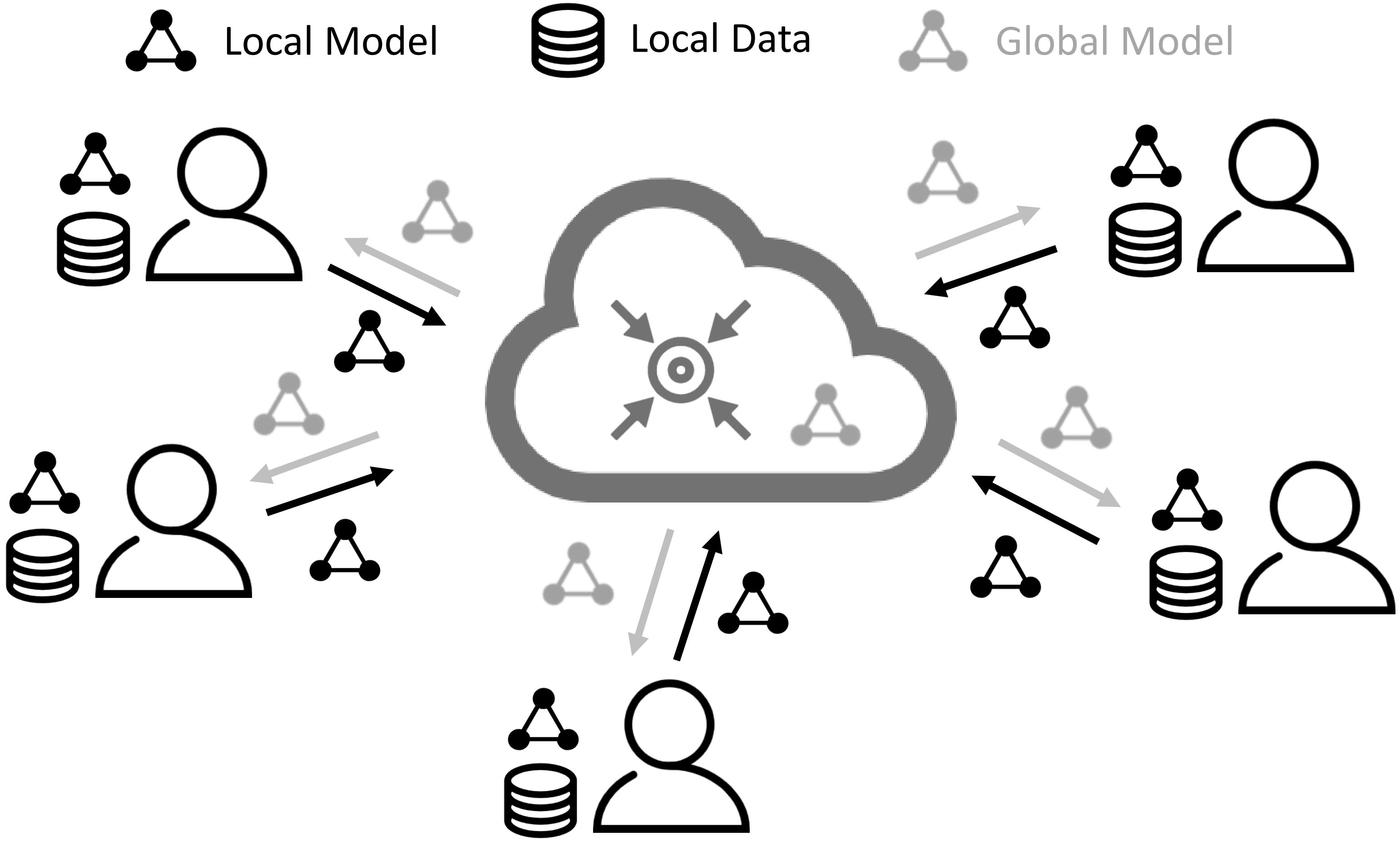


Accepted by Neurocomputing 2021

Outperform baselines by up to 20.79%

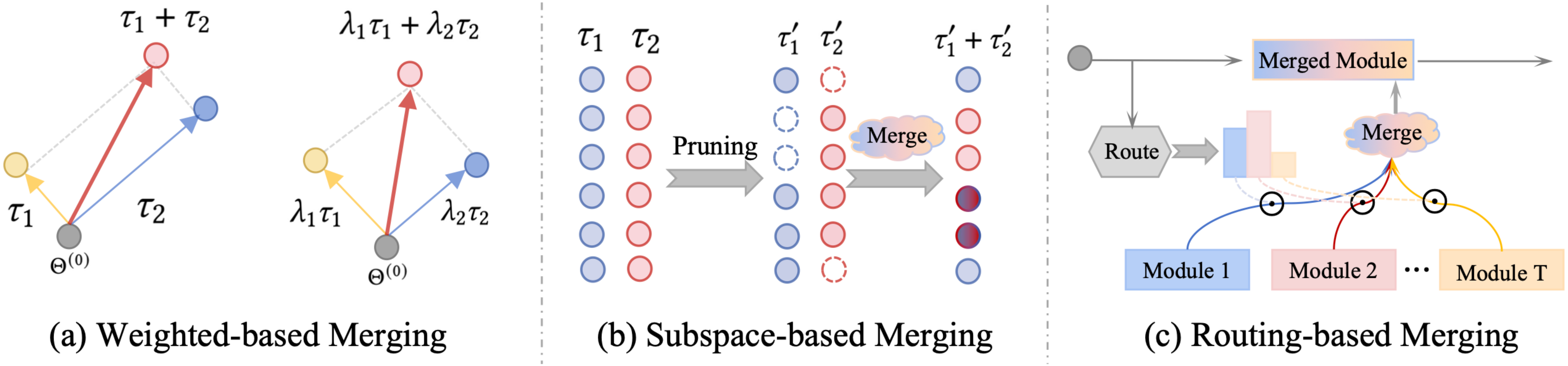
Distributed Learning (DL) (Model Merging)

- A **collaborative** and **privacy-preserving** technique for AI model training
- Finally output **one global model**



Distributed Learning (DL) (Model Merging)

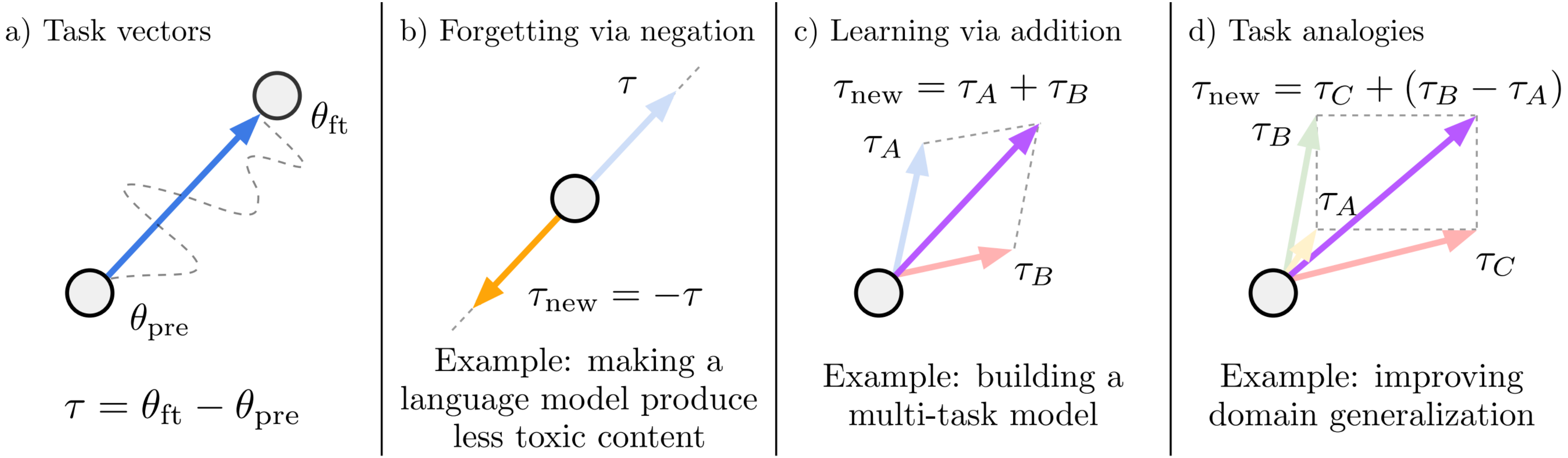
- I mainly focus on **model merging** in DL, which is also **popular** in **large model training** by
  - merging parameters, merging intermediate features, merge parameter-efficient LoRA modules, etc.
- **Multi-modal model**: obtain a single, effective, and parameter-efficient modality-agnostic model
- **RL**: DogeRM [1] merges the reward model with LLMs fine-tuned on different downstream domains to create domain-private reward models directly



[1] Lin, Tzu-Han, et al. "DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging." *EMNLP* 2024.

Distributed Learning (DL) (Model Merging)

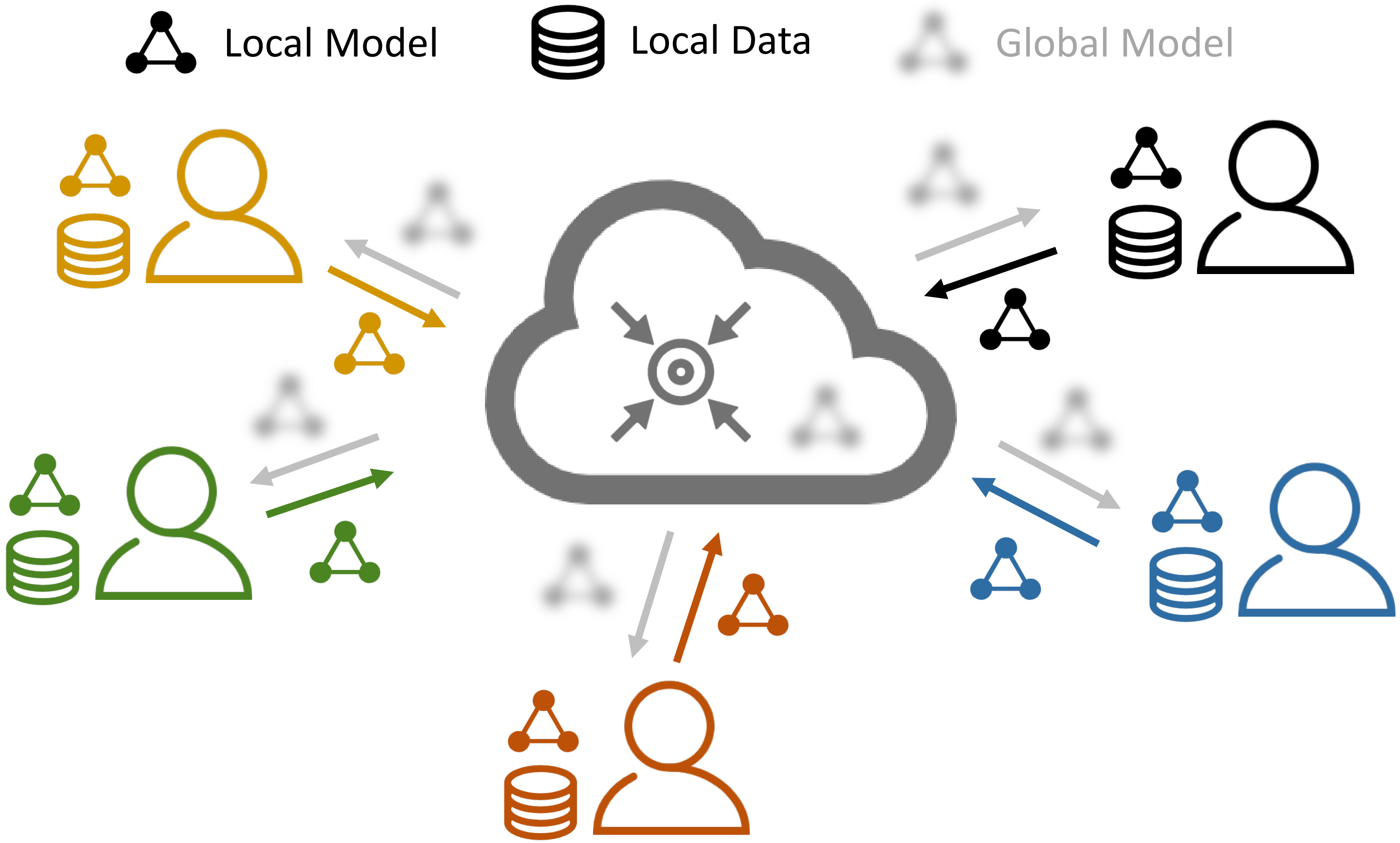
- I mainly focus on **model merging** in DL, which is also **popular** in **large model training** by
  - merging parameters, merging intermediate features, merge parameter-efficient LoRA modules, etc.
- **Model editing**: [1] shows that task vectors can be modified and combined together, and the behavior of the resulting model is steered accordingly. AlphaEdit [2] (ICLR’25 best paper)



[1] Ilharco, Gabriel, et al. "Editing models with task arithmetic." *ICLR* 2023.
[2] Fang, Junfeng, et al. "Alphaedit: Null-space constrained knowledge editing for language models." *ICLR* 2025.

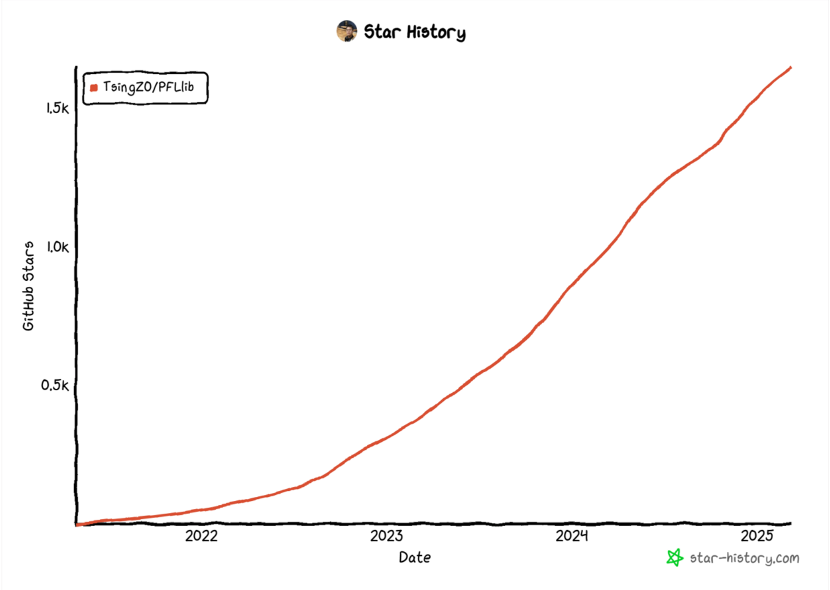
[DL]: Data Heterogeneity (Model Merging)

- **Problem:** Different clients have different data distributions, resulting in a **poor global model**



[DL]: PFLlib: pFL algorithm library and benchmark

- **Beginner-friendly**
- **39 FL&pFL, 3 scenarios, 24 datasets**
- **Popular (1800+ stars)**
- **500 clients: 5GB GPU memory**
- **Rapidly developing:**



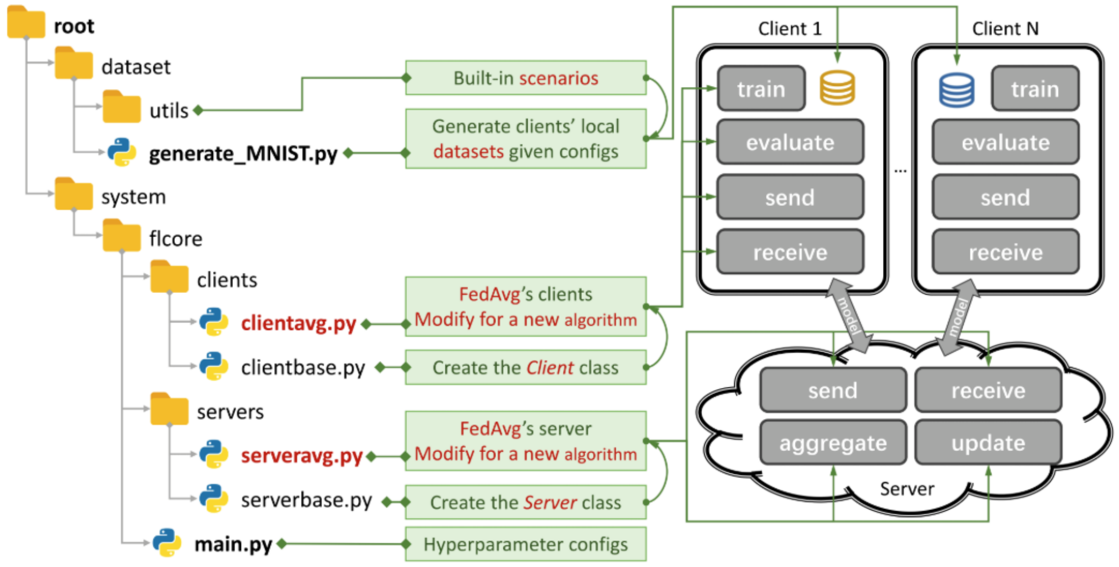
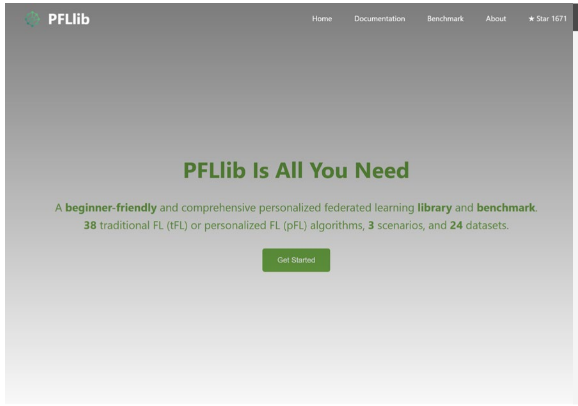
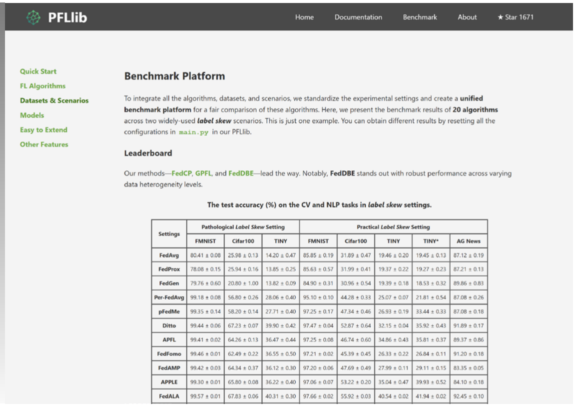


Figure 1: An Example for FedAvg. You can create a scenario using `generate_DATA.py` and run an algorithm using `main.py`, `clientNAME.py`, and `serverNAME.py`. For a new algorithm, you only need to add new features in `clientNAME.py` and `serverNAME.py`.



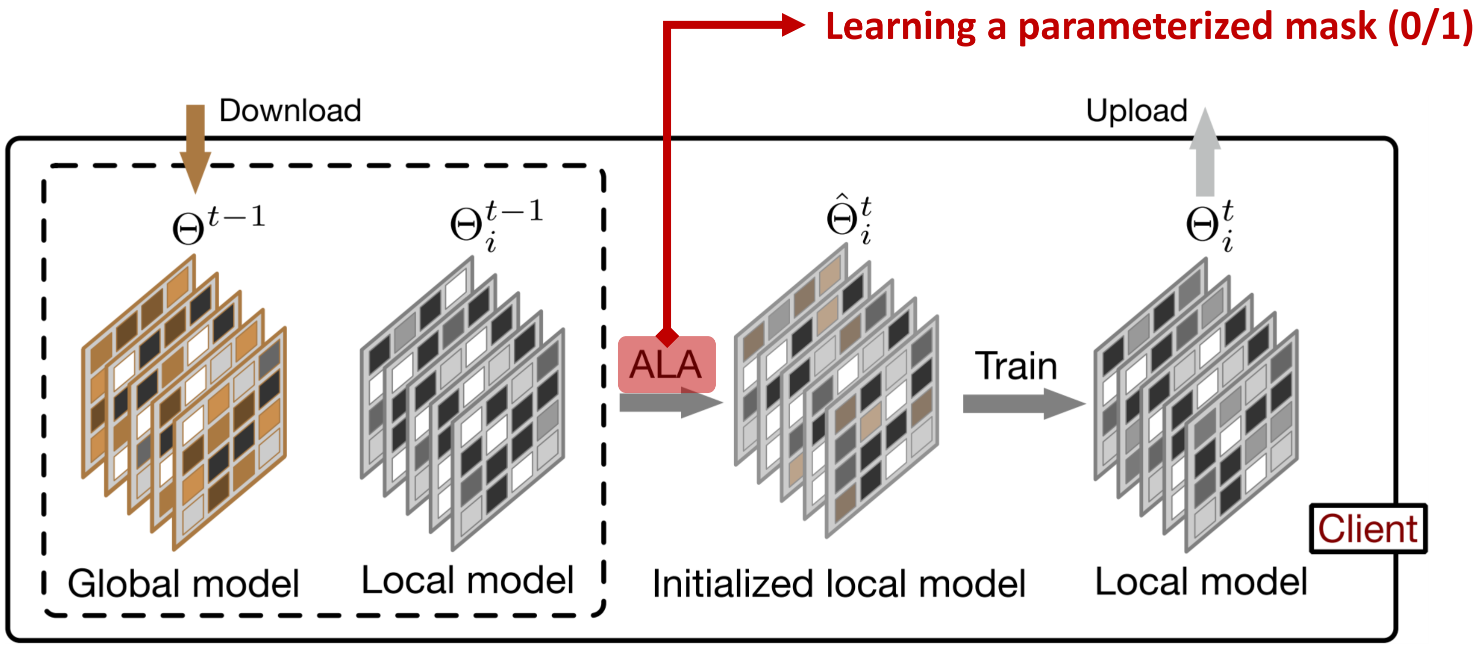


Accepted by JMLR 2025 (WAIC Outstanding Paper Top40)

https://github.com/TsingZ0/PFLlib and https://www.pfllib.com/

[DL]: FedALA (Model Merging)

- Extract each client's **desired parameters** from the global model that facilitates local training
- **Adaptively aggregate** the parameters in the global and local model in each communication round



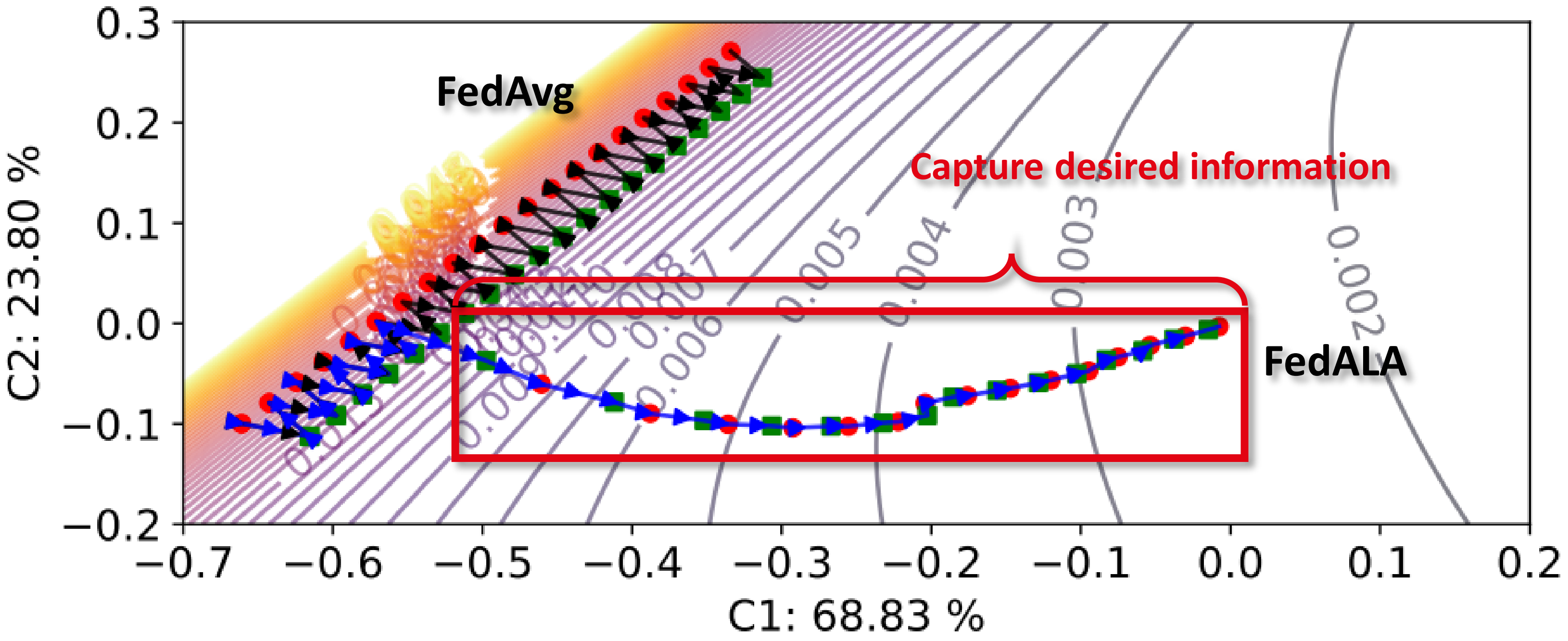
Workflow on the client in one iteration

Accepted by AAAI 2023 (oral)

Outperform baselines by up to 3.27%

[DL]: <u>FedALA</u> (Model Merging)

- Learning trajectory on one client: **FedAvg** vs. **FedALA**
- <u>Activate</u> **ALA** in the subsequent iterations



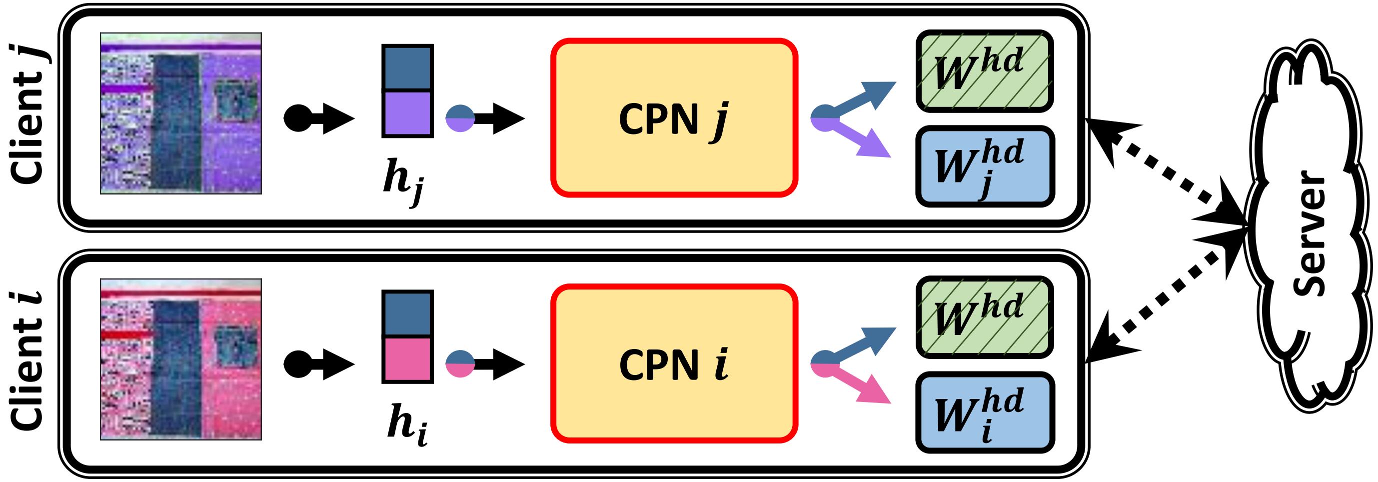
2D visualization of local learning trajectory

Accepted by AAAI 2023 (oral)

Outperform baselines by up to 3.27%

[DL]: FedCP (Model Routing)

- We **separate feature information** via an auxiliary **Conditional Policy Network (CPN).**
  - Detach generic and personalized information
  - Sample-specific separation
  - Lightweight (e.g., 4.67% parameters of ResNet-18)



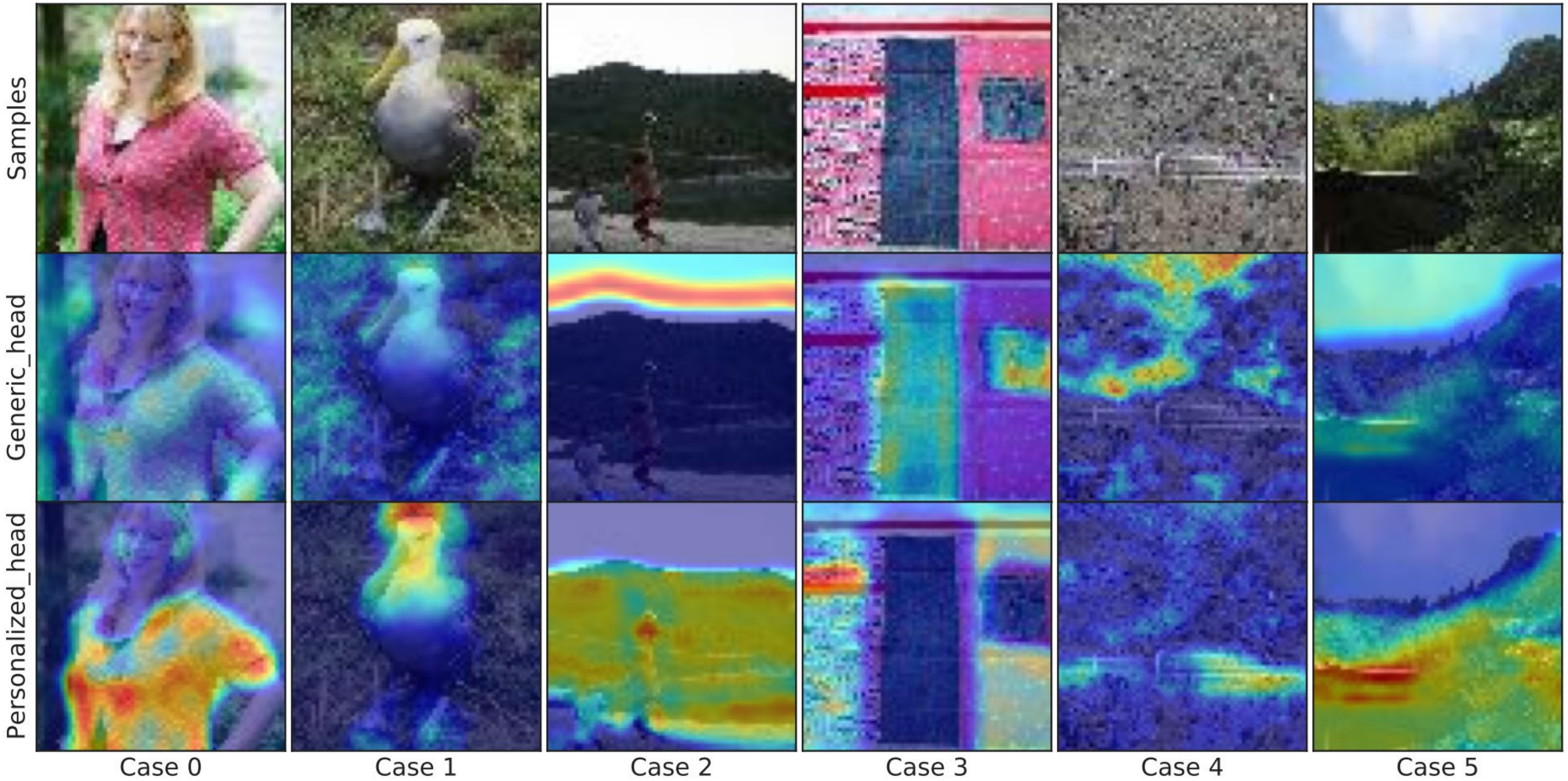
- Then, we **utilize global and personalized information** via global and personalized heads.

Accepted by KDD 2023

Outperform baselines by up to 6.69%

[DL]: <u>FedCP</u> (Model Routing)

- **Effect of feature information separation**



Six samples from the Tiny-ImageNet dataset

Accepted by KDD 2023

Outperform baselines by up to 6.69%

[DL]: <u>GPFL</u> (Model Routing)

- GCE introduces **generic and personalized routes** like MoE
- CoV **eliminates the interaction** between global and personalized feature learning

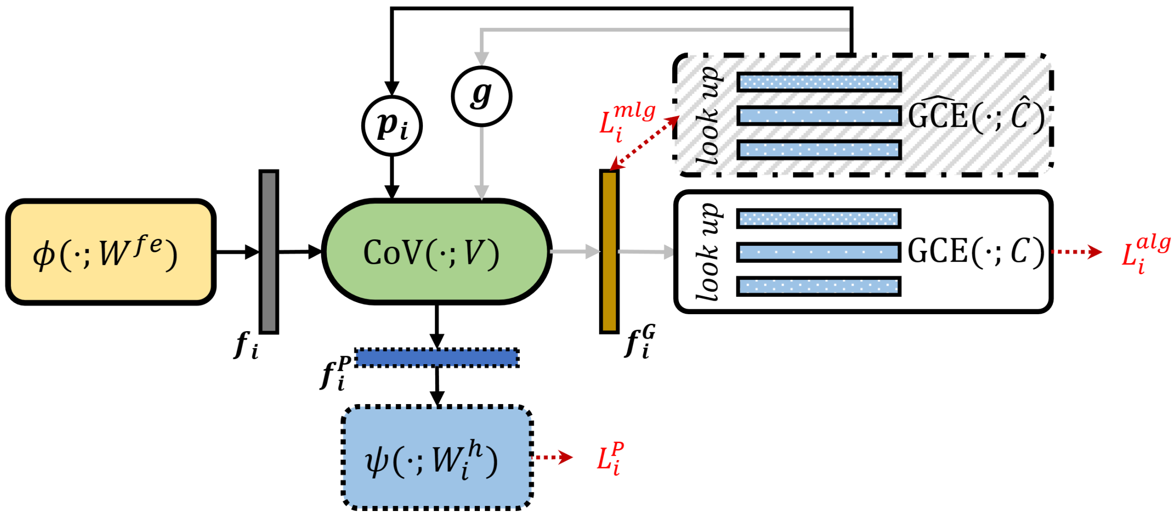
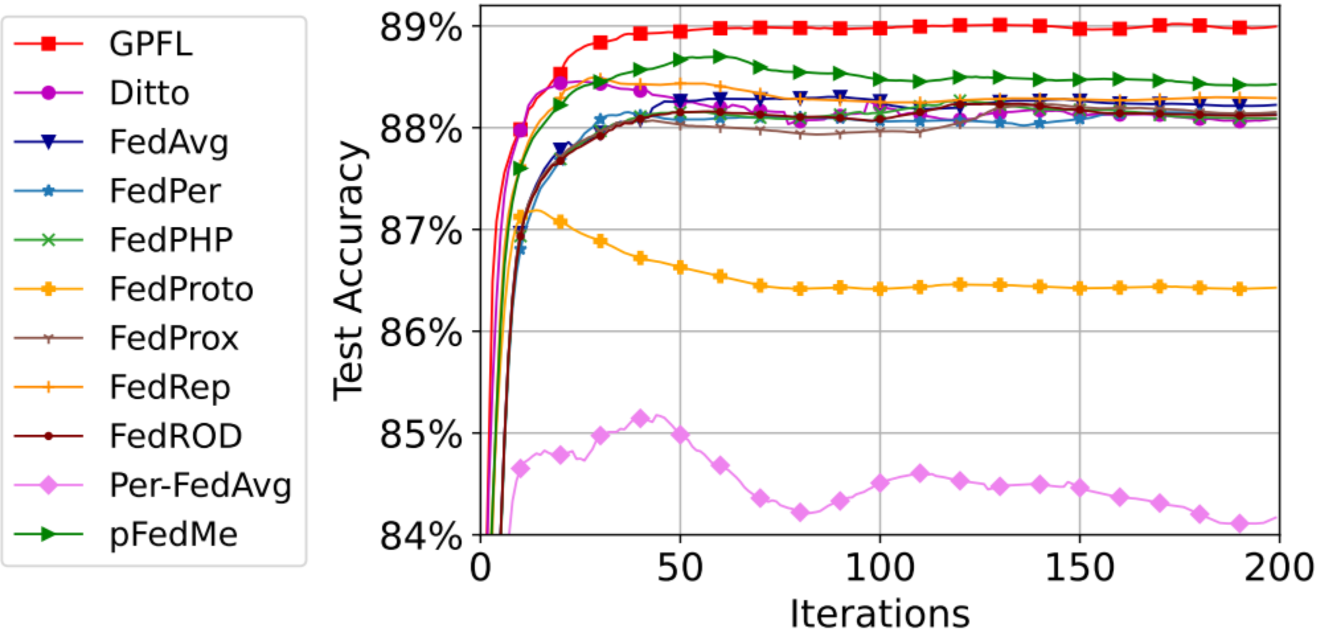


Illustration of client modules and data flow between them

[DL]: <u>GPFL</u> (Model Routing)

- Relieve the **widely existed** overfitting issue in pFL



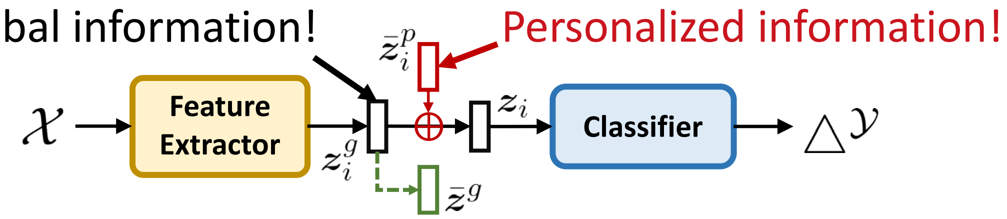
Test accuracy curves in the feature shift setting

Accepted by ICCV 2023

Outperform baselines by up to 8.99%

[DL]: DBE (Feature Decoupling)

- Eliminate domain bias by store **personalized information** in **PRBM**
- Enhance **information disentanglement** by guiding feature extractor with **MR**



Local model (with PRBM and MR)

Accepted by NeurIPS 2023

Outperform baselines by up to 11.36%

[DL]: DBE (Feature Decoupling)

- Improve **bi-directional knowledge transfer**

Corollary 1. *Consider a local data domain* $\mathcal{D}_i$ *and a virtual global data domain* $\mathcal{D}$ *for client* i *and the server, respectively. Let* $\mathcal{D}_i = \langle \mathcal{U}_i, c^* \rangle$ *and* $\mathcal{D} = \langle \mathcal{U}, c^* \rangle$*, where* $c^* : \mathcal{X} \mapsto \mathcal{Y}$ *is a ground-truth labeling function. Let* $\mathcal{H}$ *be a hypothesis space of VC dimension* d *and* $h : \mathcal{Z} \mapsto \mathcal{Y}, \forall\, h \in \mathcal{H}$*. When using* `DBE`*, given a feature extraction function* $\mathcal{F}^g : \mathcal{X} \mapsto \mathcal{Z}$ *that shared between* $\mathcal{D}_i$ *and* $\mathcal{D}$*, a random labeled sample of size* m *generated by applying* $\mathcal{F}^g$ *to a random sample from* $\mathcal{U}_i$ *labeled according to* c^**, then for every* $h^g \in \mathcal{H}$*, with probability at least* $1 - \delta$*:*

$$\mathcal{L}_{\mathcal{D}}(h^g) \leq \mathcal{L}_{\hat{\mathcal{D}}_i}(h^g) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)} + \boxed{d_{\mathcal{H}}(\tilde{\mathcal{U}}_i^g, \tilde{\mathcal{U}}^g)} + \lambda_i,$$

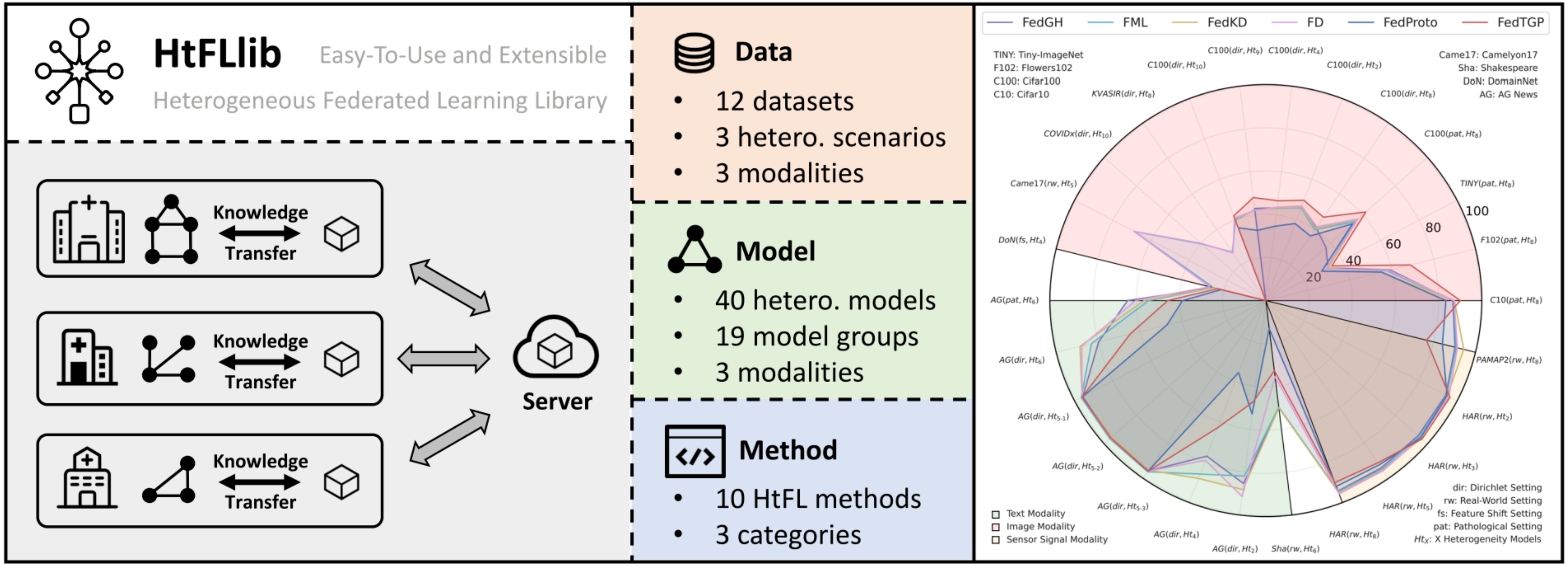
where $\mathcal{L}_{\hat{\mathcal{D}}_i}$ *is the empirical loss on* $\mathcal{D}_i$*,* e *is the base of the natural logarithm, and* $d_{\mathcal{H}}(\cdot,\cdot)$ *is the* $\mathcal{H}$*-divergence between two distributions.* $\lambda_i := \min_{h^g} \mathcal{L}_{\mathcal{D}}(h^g) + \mathcal{L}_{\mathcal{D}_i}(h^g)$*,* $\tilde{\mathcal{U}}_i^g \subseteq \mathcal{Z}$*,* $\tilde{\mathcal{U}}^g \subseteq \mathcal{Z}$*, and* $\boxed{d_{\mathcal{H}}(\tilde{\mathcal{U}}_i^g, \tilde{\mathcal{U}}^g) \leq d_{\mathcal{H}}(\tilde{\mathcal{U}}_i, \tilde{\mathcal{U}})}$*.* $\tilde{\mathcal{U}}_i^g$ *and* $\tilde{\mathcal{U}}^g$ *are the induced distributions of* $\mathcal{U}_i$ *and* $\mathcal{U}$ *under* $\mathcal{F}^g$*, respectively.* $\tilde{\mathcal{U}}_i$ *and* $\tilde{\mathcal{U}}$ *are the induced distributions of* $\mathcal{U}_i$ *and* $\mathcal{U}$ *under* $\mathcal{F}$*, respectively.* $\mathcal{F}$ *is the feature extraction function in the original FedAvg without* `DBE`*.*

Accepted by NeurIPS 2023

Outperform baselines by up to 11.36%

[DL]: **HtFLlib**: HtFL Library and Benchmark

- **Easy-to-use** and **extensible:** modify only two files to add a new algorithm
- **PFLlib compatible:** support all PFLlib's scenarios, datasets, tools, etc.
- **First & comprehensive:** 40 heterogeneous models, 3 modalities, 10 HtFL methods, etc.



Accepted by KDD 2025 (Best Paper Runner UP)

https://github.com/TsingZ0/HtFLlib

[DL]: DL on Real-World Devices

- **Real-world deployment**
  - **+ CoLExT, + real-world datasets, + systematical metrics**

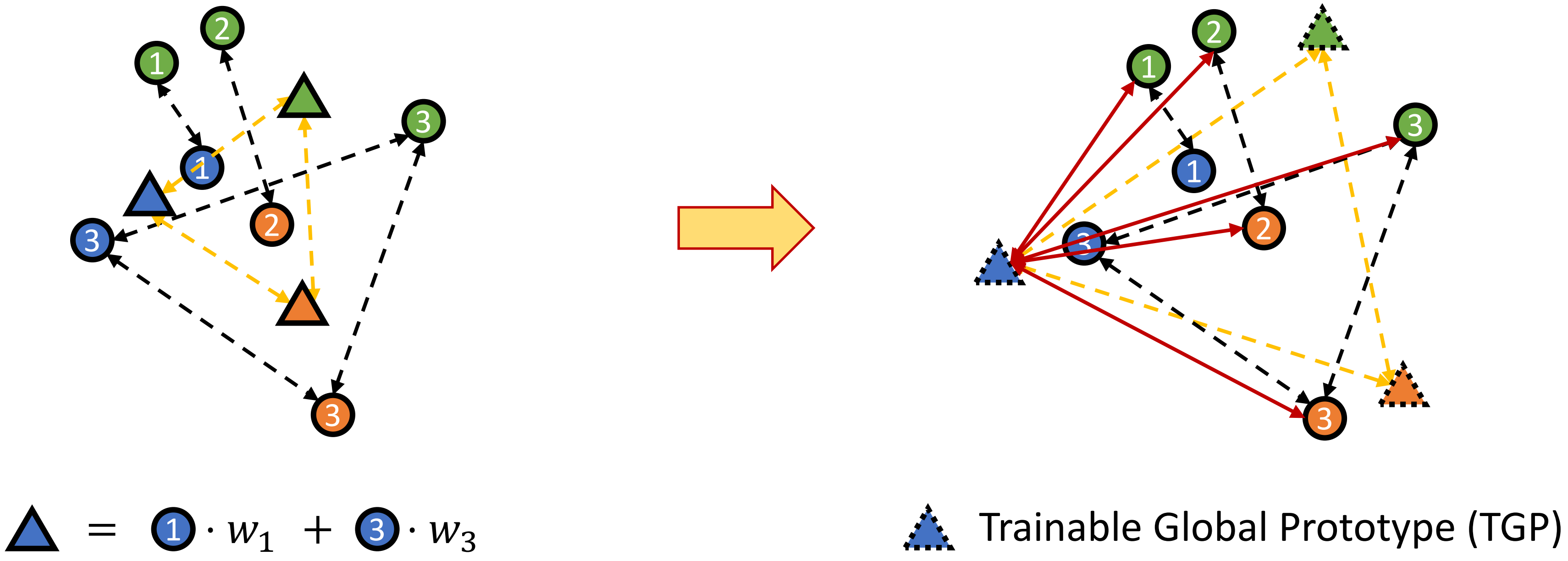
- 28 Single Board Computers (SBC)
  - Orange Pi, LattePanda, Nvidia Jetson
- 20 Smartphones
  - Samsung, Xiaomi, Google Pixel, Asus ROG, One Plus
- High Voltage Power Meter
- Wired and wireless networking
- Workstation - FL Server

Collaborating with KAUST

https://github.com/TsingZ0/HtFL-OnDevice

[DL]: FedTGP (Knowledge Distillation)

- **Enlarge** the global prototype margin
- Ensure optimal feature quality across clients



Accepted by AAAI 2024

Outperform baselines by up to 9.08%, and achieve up to an 18.96% improvement over FedProto

[DL]: FedTGP (Knowledge Distillation)

- **Adaptive-margin-enhanced Contrastive Learning (ACL)**
- ACL is **universal** and can be applied to other tasks

$$\min_{\hat{\mathcal{P}}} \sum_{c=1}^{C} \mathcal{L}_P^c,$$

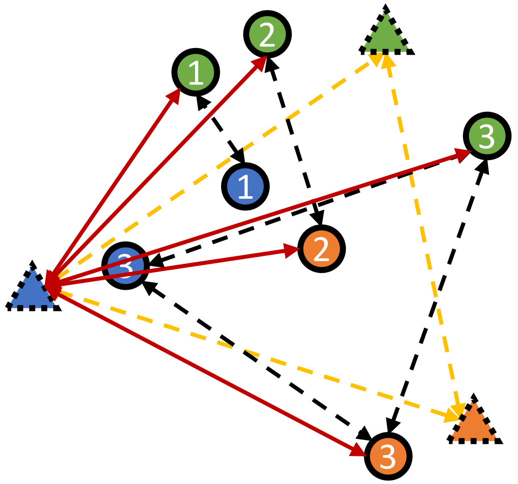
$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} -\log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))}}{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}}$$

$$\delta(t) = \min(\max_{c \in [C], c' \in [C], c \neq c'} \phi(Q_t^c, Q_t^{c'}), \tau),$$

$$Q_t^c = \frac{1}{|\mathcal{P}_t^c|} \sum_{i \in \mathcal{I}^t} P_i^c, \forall c \in [C]$$

maximum cluster margin

τ is a margin threshold



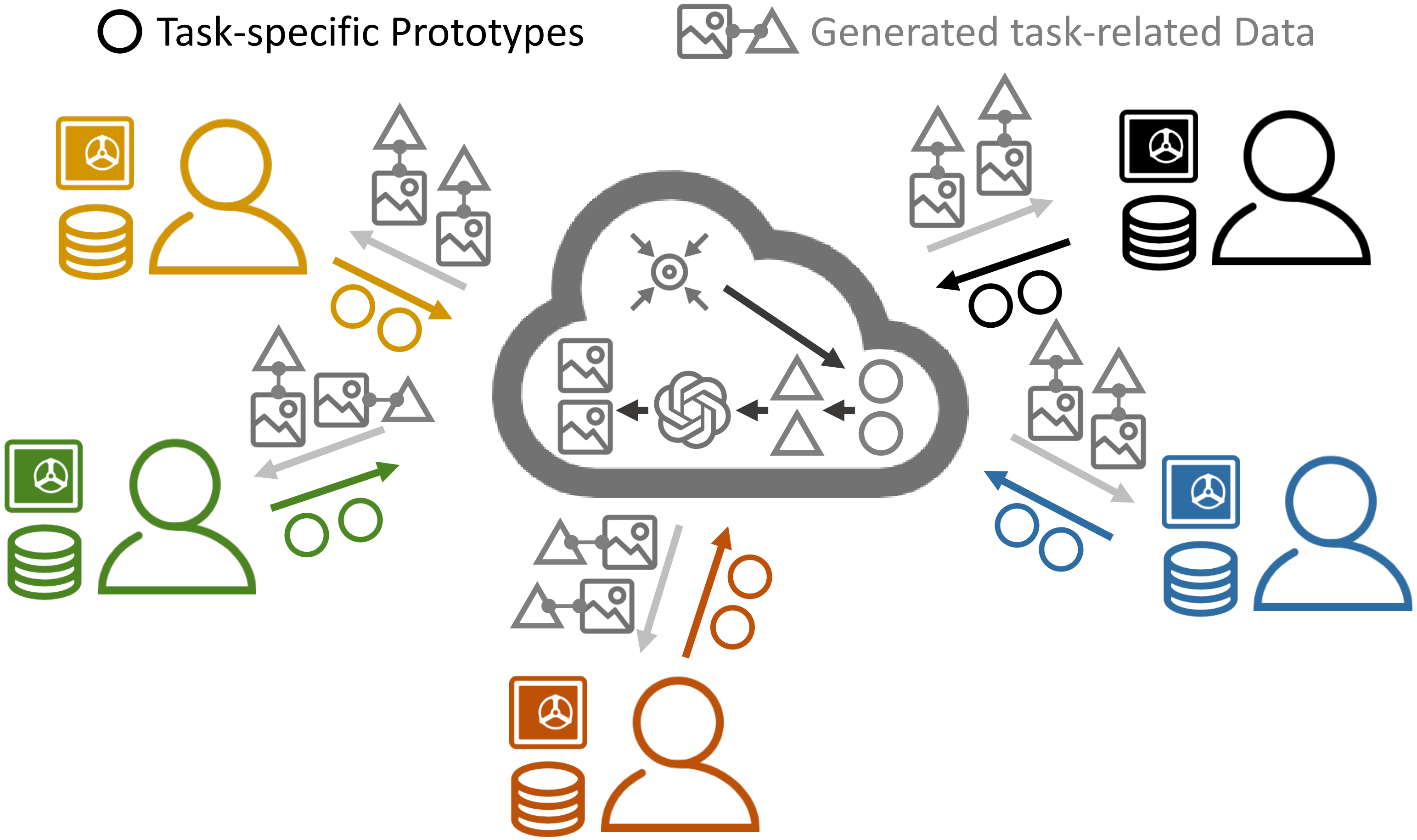
$\hat{P}^c$: A TGP of class c

$\hat{\mathcal{P}}$: All TGP

P_i^c: A prototype of class c from client i

[DL]: FedKTL (Knowledge Distillation)

- Transfer **common knowledge** from the pre-trained generators to clients
- Obtain **task-specific knowledge** from other clients

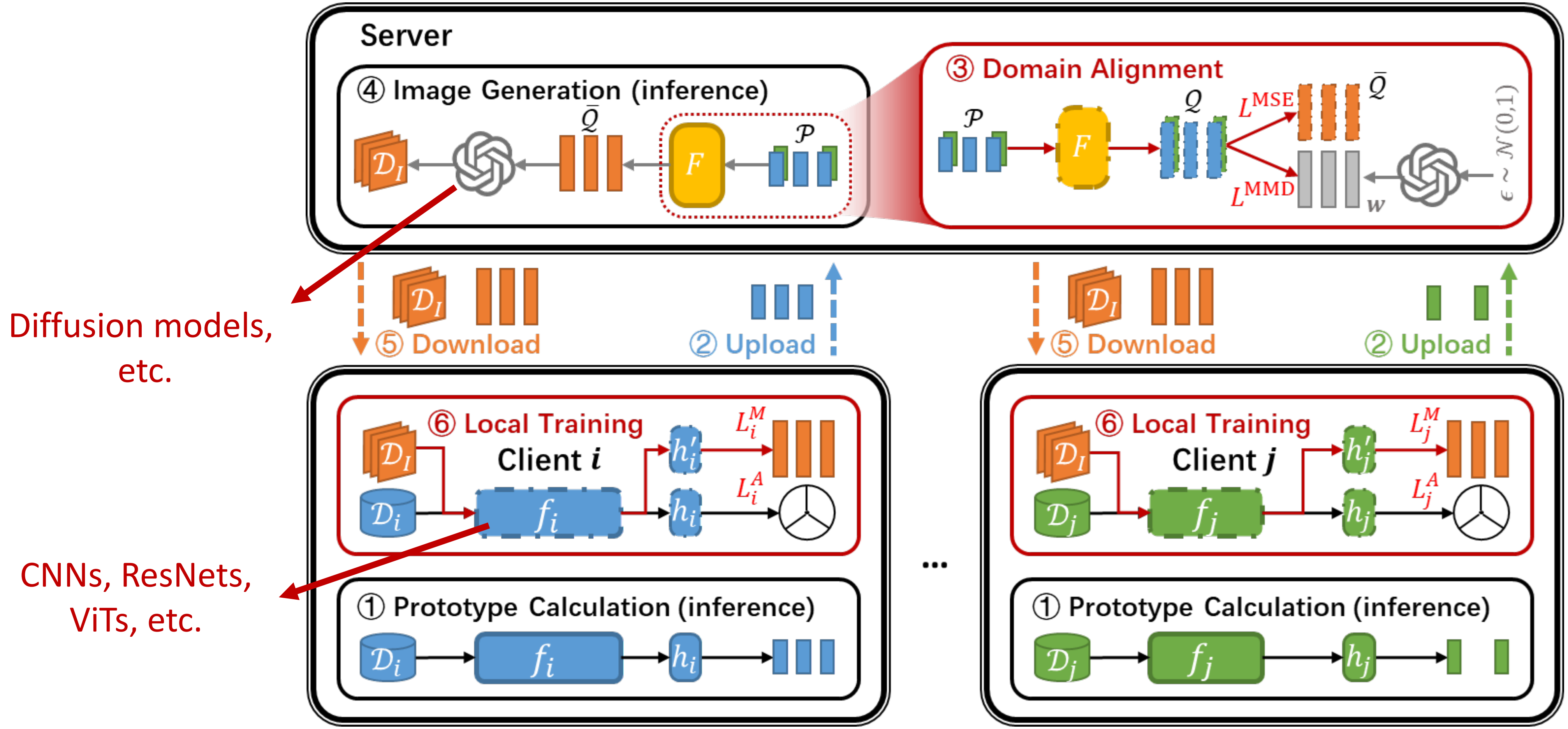


Accepted by CVPR 2024

Outperform baselines by up to 7.31%

[DL]: FedKTL (Knowledge Distillation)

- **We need to Align** small models' feature space with the generative model's for the transfer
- Transfer global knowledge using an **additional supervised local task**



Accepted by CVPR 2024

Outperform baselines by up to 7.31%

[DL]: FedKTL (Knowledge Distillation)

- FedKTL can **adapt to various generators** that were pre-trained using various datasets
- The **semantics of the generated images** can be different from clients' data

(a) Client #1 (b) AFHQv2 (c) Benches (d) FFHQ-U (e) WikiArt

Generators pre-trained on different image datasets

	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
AFHQv2	26.82±0.32	**27.05±0.26**	26.32±0.52
Bench	27.71±0.25	**28.36±0.42**	27.56±0.50
FFHQ-U	**27.28±0.23**	27.21±0.35	26.59±0.47
WikiArt	27.37±0.51	**27.48±0.33**	27.30±0.15

Accepted by CVPR 2024

Outperform baselines by up to 7.31%

Industry Insights & Future Outlook

Code Assistant
- **Code edit (NES),** user intent personalization
- **Agent evolution,** environment perception
- **Code-defined agent,** Real-time agent generation

Past work: intention modeling, personalized module, synthetic data pipeline, RLVR, etc.

Voice Assistants
- **Lightweight** model on the edge side
- **Private data** on the edge, GDPR, etc.

Future AGI Outlook
- **Vertical domain** to general domain
- **Pluggable data module**, suitable for small and medium-sized enterprises

Feel free to contact me!

Home page: https://github.com/TsingZ0

Thanks!