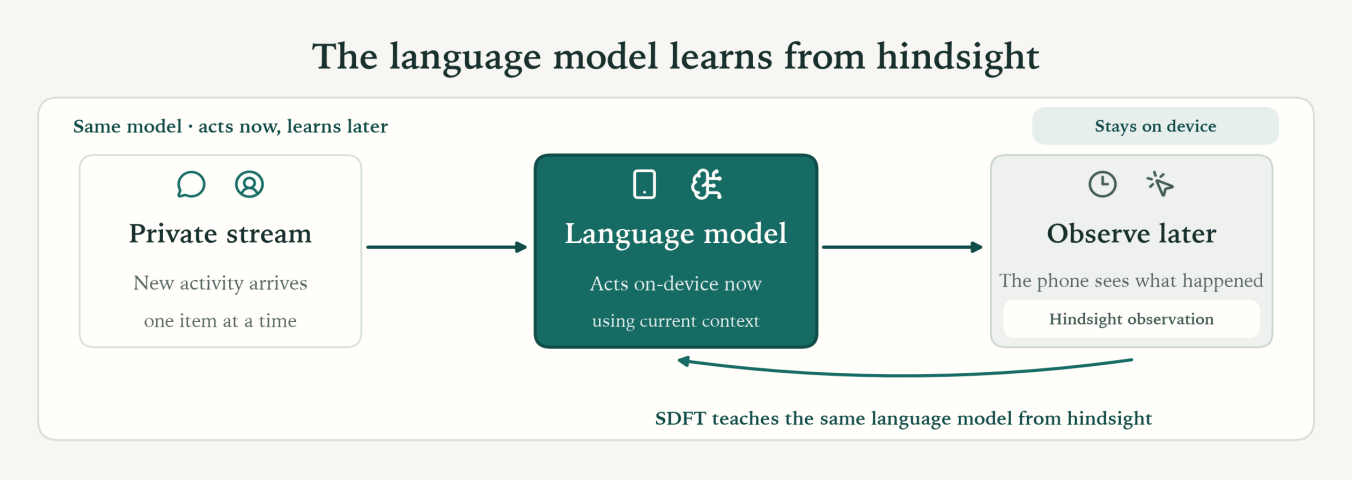
# Fine-Tuning Small Language Models for Continual Learning On-Device

## Learning from hindsight with online self-distillation




I-Ju Lin & Zhang-Wei Hong

Adapted from the project website • Website updated August 15, 2026

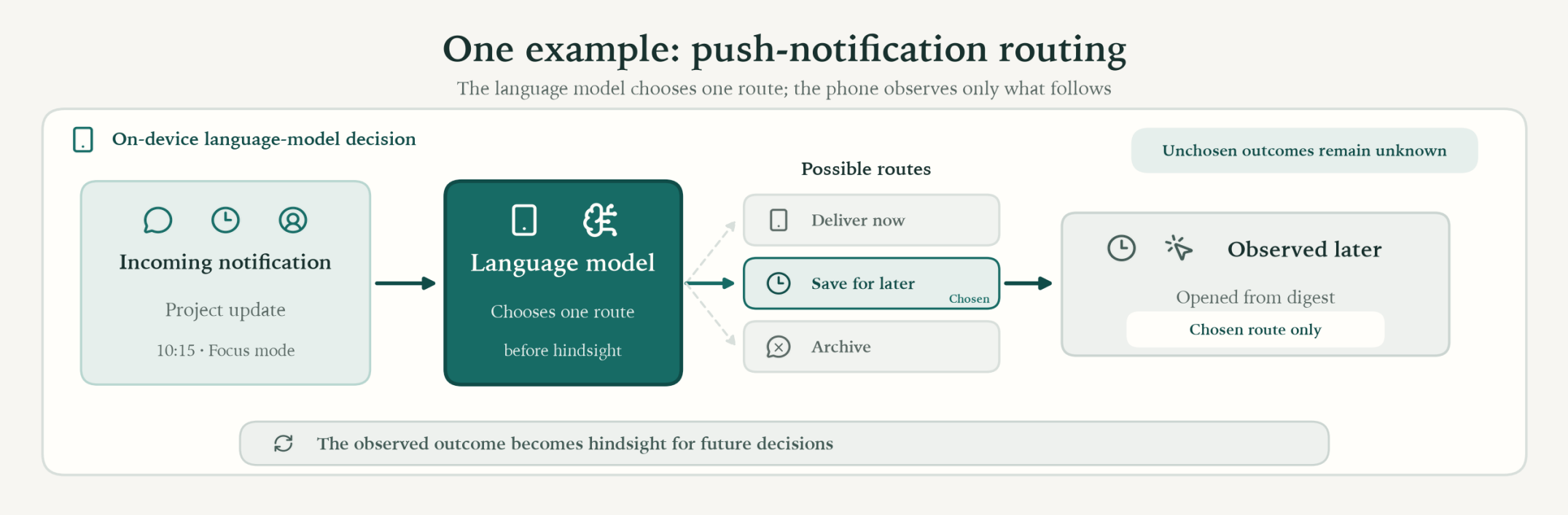
# Local inference is only half the problem

A personal assistant should adapt to a user’s writing, choices, and daily routines.

Phone interactions are sparse, asynchronous, and often reveal no correct action.

Can a small model keep learning while personal context and updates stay on-device?

# The model acts before useful evidence arrives




Only the chosen route exposes an outcome. Unchosen outcomes remain unknown.

# Neither a correct label nor a reliable reward

Supervised fine-tuning: needs the desired action, which is usually unavailable.

Reinforcement learning: needs a useful reward; opening or ignoring a notification is ambiguous.

Memory: can retrieve prior interactions, but each decision depends on finding the right example.

Goal: turn delayed factual evidence into updates that improve future decisions.

# Use hindsight to teach the original decision

**Student: $p(y) = \pi(y \mid x)$**

**Teacher: $q(y) = \pi(y \mid x, z)$**

$$L = KL(\text{stop_gradient}[q] \,||\, \pi\theta(\cdot \mid x))$$

x: original context z: observed hindsight

Conceptual objective: the teacher sees later evidence; the student learns from the original input.

# The released target preserves uncertainty

Reliable single-route callback: 90% supported routes + 5% adapter-disabled model review + 5% saved decision-time beliefs.

Ambiguous digest open: redistribute saved beliefs across actions that remain plausible.

UNKNOWN outcome: create no target and perform no update.

The implemented target adds causal support and reliability gating to the conceptual SDFT objective.

# Explore early; replay a bounded set of lessons

## Controlled exploration

Epsilon-greedy behavior avoids early lock-in. An uncertainty-gated probe seeks observable feedback. Exploration tapers toward greedy serving.

## Balanced replay

Retain recent usable lessons, balance feedback groups, and weight by recency. Every batch includes the newest lesson.

Replay is training-only. Delayed outcomes remain attached to their original context and action.

# A controlled benchmark with delayed feedback

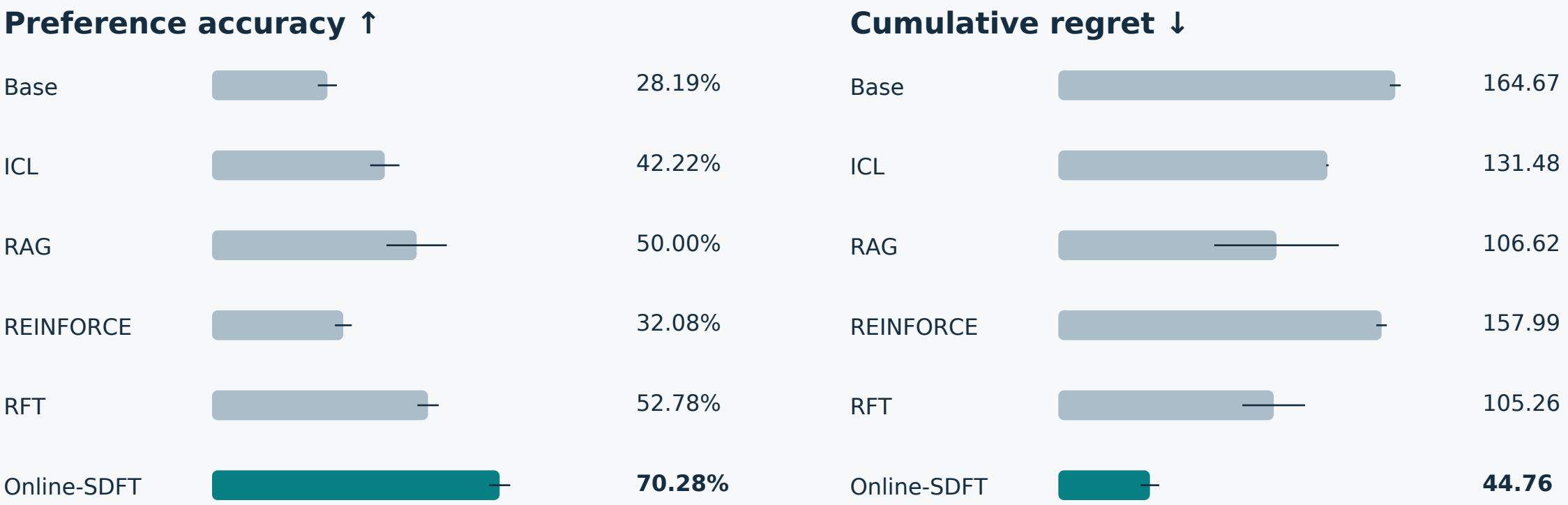
720 decisions: three paired synthetic streams, 240 decisions each; weekday, on-call, and off-hours regimes.

LiquidAI/LFM2.5-230M: frozen base with rank-4 LoRA; 172,032 trainable parameters.

Evaluate before learning: hidden sampled preferences and route utilities are evaluation-only.

Benchmark run: Apple MPS. Accuracy, regret, and observable reward measure different quantities.

# Online-SDFT leads on these synthetic streams




Means and nominal 95% confidence intervals across only three paired streams.

# Soft targets retain more usable evidence

**+17.50**

percentage points in accuracy vs. RFT

**60.50**

less cumulative regret vs. RFT

RFT accepted 75 of 311 teacher candidates (24.12%). Soft targets can preserve graded information that a verified hard label discards.

Comparison to the strongest baseline by mean accuracy and mean regret.

# Replay drives the largest ablation gain

| Variant | Accuracy | Cumulative regret |
|---|---|---|
| **Full Online-SDFT** | **70.28% ± 2.60** | **44.76 ± 4.57** |
| No exploration | 69.44% ± 4.72 | 62.00 ± 10.96 |
| No replay | 39.58% ± 1.70 | 134.90 ± 11.34 |
| Neither trick | 42.78% ± 3.41 | 130.42 ± 4.12 |

Removing replay costs 30.69 accuracy points. Removing exploration adds 17.24 regret with little change in accuracy.

Paired 2 × 2 ablation; same model, optimizer, hindsight target, and event streams.

# Android proof of concept: learning on one phone

Local LFM2.5-230M decisions; ONNX Runtime Training updates a rank-4 LoRA adapter.

Dismiss a notification until it stays quiet; ask for it back, and the next one is delivered. “Show now” rises to 94%.

The demo runs without network access; replay and checkpoints remain in app-private storage.

Engineering prototype. Host provisioning is still required; FP32 targets high-memory ARM64 devices.

# Promising evidence, with a narrow scope

Reliability-conditioned soft hindsight targets work well in this controlled, action-dependent setting.

Only three synthetic streams; the selected configuration lacks held-out-stream confirmation.

Android latency, memory, battery, thermal behavior, background reliability, and end-to-end privacy remain unmeasured.

Next: larger held-out comparisons and profiling on physical edge devices.

# Resources and selected references

## Project website & code

lin826.github.io/SLM-Online-SDFT/
github.com/lin826/SLM-Online-SDFT

Shenfeld et al. Self-Distillation Enables Continual Learning. 2026.
arxiv.org/abs/2601.19897

Qin et al. Enabling On-Device Large Language Model Personalization with Self-Supervised Data Selection and Synthesis. DAC 2024.

Figures and reported measurements reproduced from the source website.