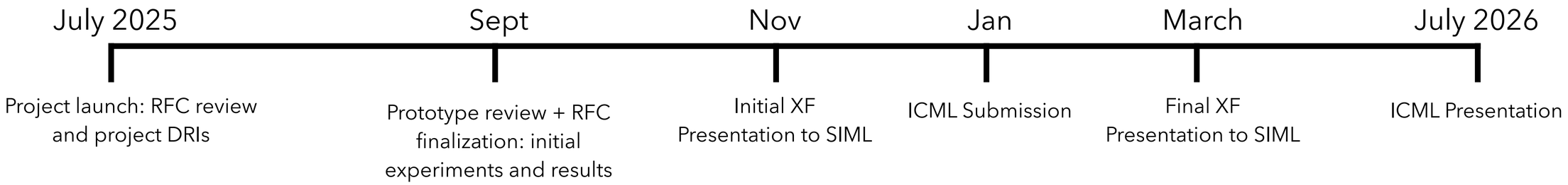
University Research Collaborations - CMU

Directed funding research projects -$100k funding for:
Carnegie Mellon University, Research in Requirements-Driven AI App Synthesis and Testing
DRIs - Eric Liang Yang, Andrew Walkingshaw

Summary

- Research team: HCI-focused researchers working on how humans interact with AI systems
- Research focus: constraint-driven LLM application development (prompt->eval)
- Expected output: co-development of framework/experiments. Intern->FTE opportunities
- Investment: 100k investment. Biweekly meetings (0.5hr, two Apple attendees)



Principal investigator



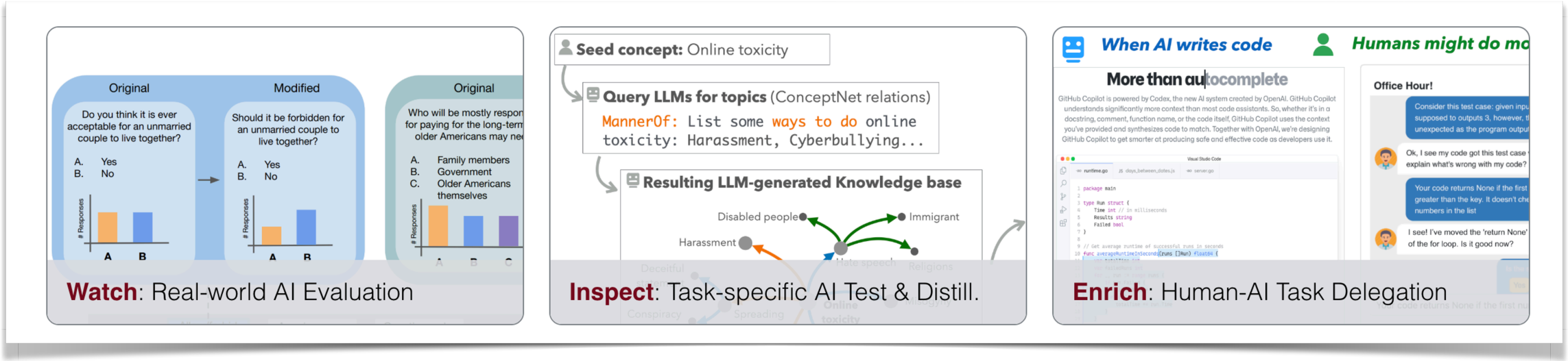
Prof. Sherry Wu

Assistant Professor (tenure track) in Human-Computer Interaction Institute, School of CS

- Area chair for top AI/HCI/NLP conferences.
- Works recognized by media coverage from The Washington Post, MIT Tech Review, CNET, etc.

CMU, Research in LLM App Dev | Research Lab Summary

WInE (Watch, Inspect, and Enrich) lab at Carnegie Mellon University, Human-Computer Interaction Institute.
Focuses on design, build, and evaluation of practical AI systems that can best support humans.



Core topics:

1. **Practical AI Evaluation**: We evaluate AI models beyond accuracy metrics, focusing on task-specific needs. This involves frameworks that augment domain knowledge and end-to-end benchmarks that assess AI's ability to adapt across varied fields.
2. **Tailoring AI Behaviors**: We investigate methods for refining AI behaviors to meet user requirements. This includes explicating users' implicit preferences and translating user inputs into actionable AI tasks.
3. **Human-AI Collaboration**: We explore how AI and humans can optimally collaborate, by improving task delegation and incorporating AI into education and real-world applications.

CMU, Research in LLM App Dev | Research Goals

Requirement-Driven LLM prototyping & deployment

Prompting is the best prototyping technique we have, but large models are slow and inefficient. We would like to have small, efficient task-specific models. How can we transfer insights from prompting to post-training?

We propose a requirement-centered framework where requirements unify task definition—guiding prompt synthesis during prototyping and enabling high-quality synthetic data generation for training smaller models.

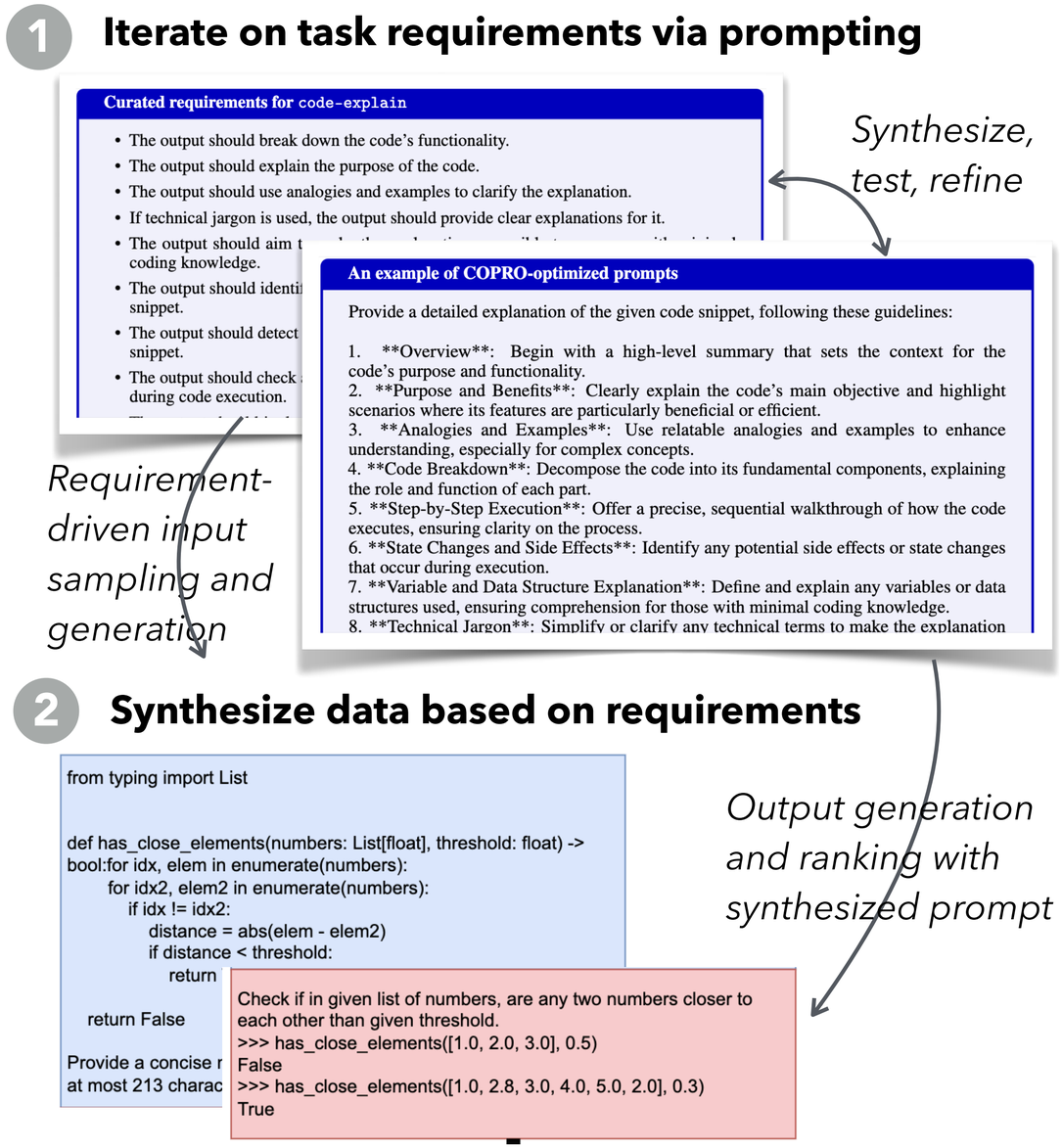
The steps are:

Iterative Requirement-Driven Prompting

- Developers refine use-case requirements by evaluating auto-synthesized prompts with requirement-specific checks (LLM- or rule-based).
- Tools to accelerate prompt engineering, like requirement validators to flag vague/conflicting constraints and requirement-aware prompt synthesizers

Synthetic data generation and model training aligned to requirements

- Map task instructions to requirement clusters using instruction-dataset pairs
- Generate multiple outputs per prompt and sample outputs that best satisfy requirements.



CMU, Research in LLM App Dev | Relevance, logistics, process

Relevance

- Enables us to scale our Apple Intelligence product process by:
  - Identifying which use cases require adapters vs just prompting
  - Accelerating development of synthetic data pipelines (for PEFT, training, evaluation)
    - Important for distillation into small (eg on-device) models
  - Giving us insight into better prototyping tool (informs work on *Belle*)

Timeline

Duration of engagement: 1 year

- H2/2025: Tools for requirement extraction and prompt synthesis
- H1/2026: Dataset generation + small model tuning
- H2/2026: Evaluation, benchmarking, open-source code release

Target publication venues

Bi-weekly research meeting with the PI and the student
Quarterly project status review
Paper submission to ICML/ICLR/NeurIPS 2026

Budget

$100k: 1 full-time PhD student (95k), with additional research cost e.g., compute, human-subject experiments, etc. (5k).

DRIs

University DRI: Prof. Sherry Wu
Apple DRI: Eric Liang Yang, Andrew Walkingshaw