UC SANTA BARBARA

Department of Electrical & Computer Engineering

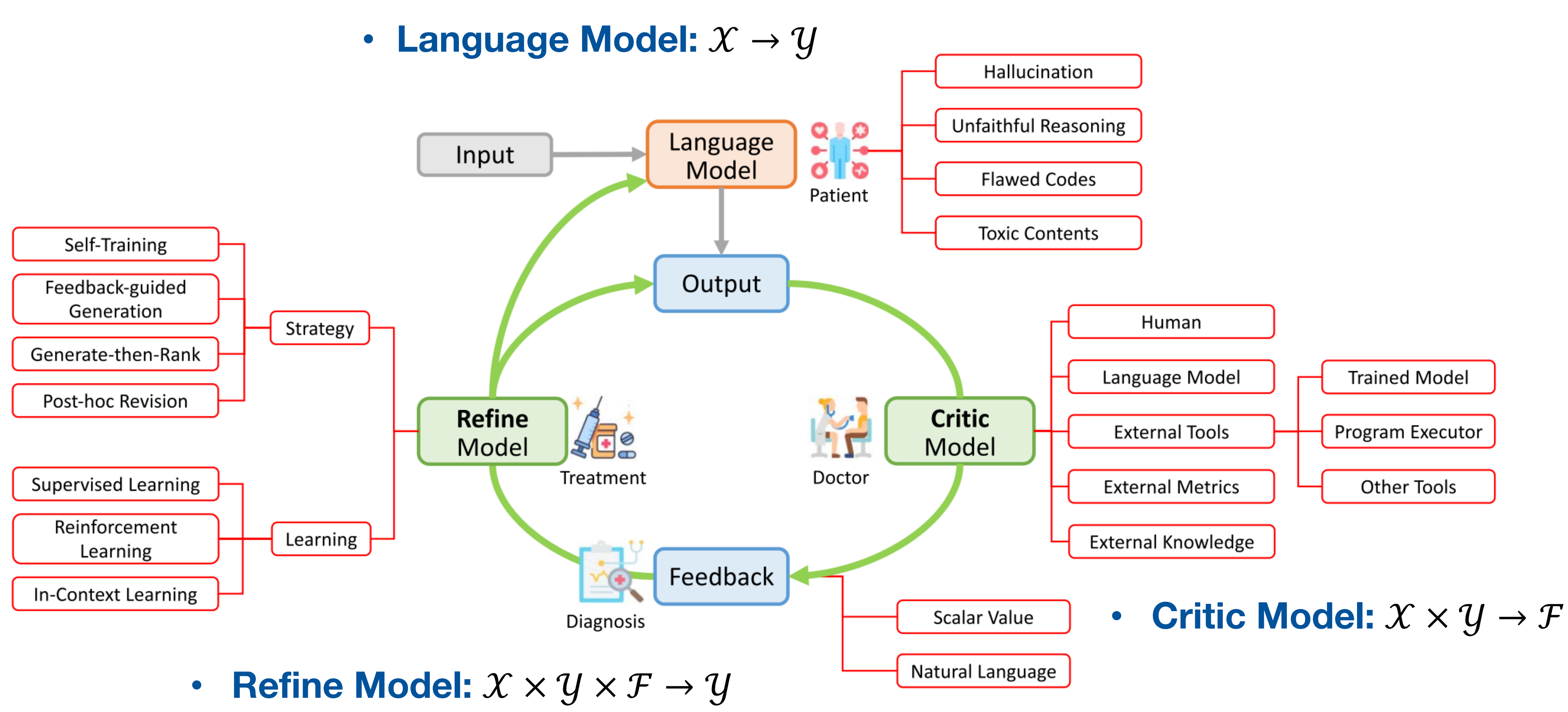
Paper

Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, William Yang Wang

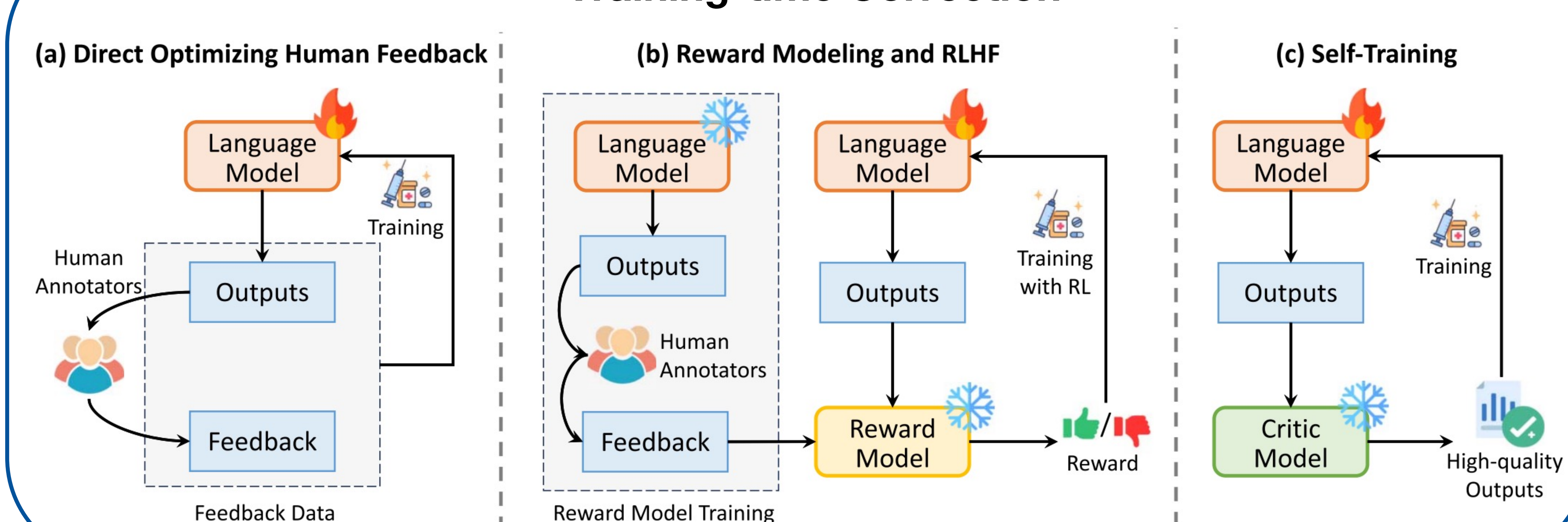
Landscape of Correcting LLMs with Automated Feedback

- **What gets corrected?**
  - ✓ hallucinations
  - ✓ reasoning Errors
  - ✓ biased / harmful content
- **Source of the feedback?**
  - ✓ self-feedback
  - ✓ external Feedback
- **Format of the feedback?**
  - ✓ scalar value
  - ✓ natural language
- **When to correct the model?**
  - ✓ training-time
  - ✓ generation-time
  - ✓ post-hoc
- **How to correct the model?**
  - ✓ on the output
  - ✓ on the parameters

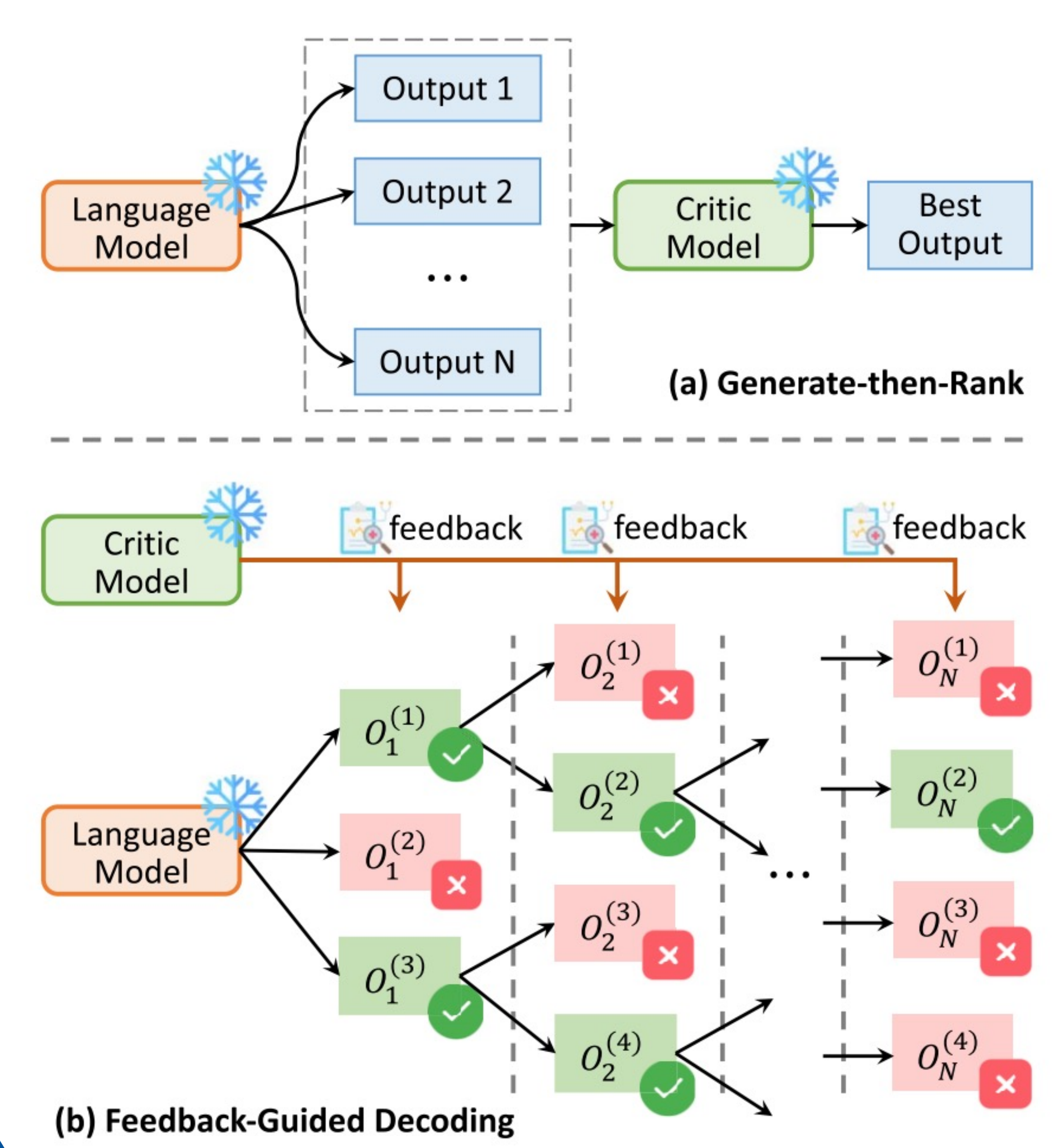


Typical Automated Correction Strategies

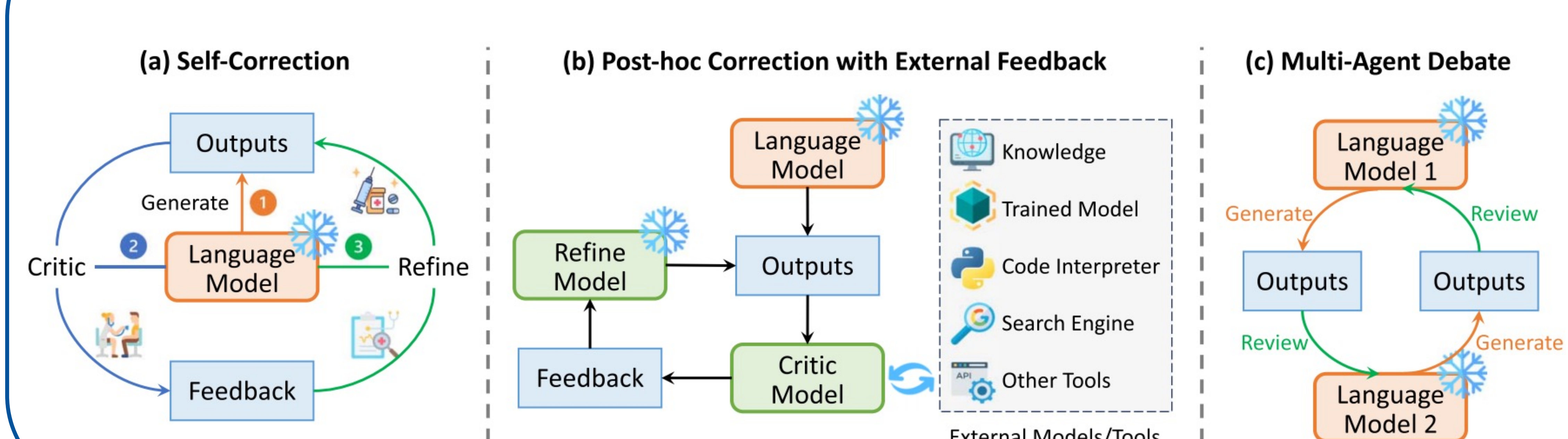
Training-time Correction



Generation-time Correction



Post-hoc Correction



Key Findings

- **Self-feedback** is bounded by LLM's own knowledge and capability
- **Leveraging external feedback** is encouraging, but high-quality external feedback is unavailable in many scenarios
- Training **high-quality feedback** model is the bottleneck

Future Directions

- Theoretical analysis of automated correction
- Benchmarking Automated Correction
- Continual Self-Improvement
- Self-Correction with Model Editing
- Multi-modal Self-Correction

@PanLiangming liangmingpan@ucsb.edu