Yu WU, Yi YANG, 2026. A platform-aware survey of execution, verification, and risk of graphical user interface agents. *ENGINEERING Information Technology & Electronic Engineering,* 27(7):260103. https://doi.org/10.1631/ENG.ITEE.2026.0103

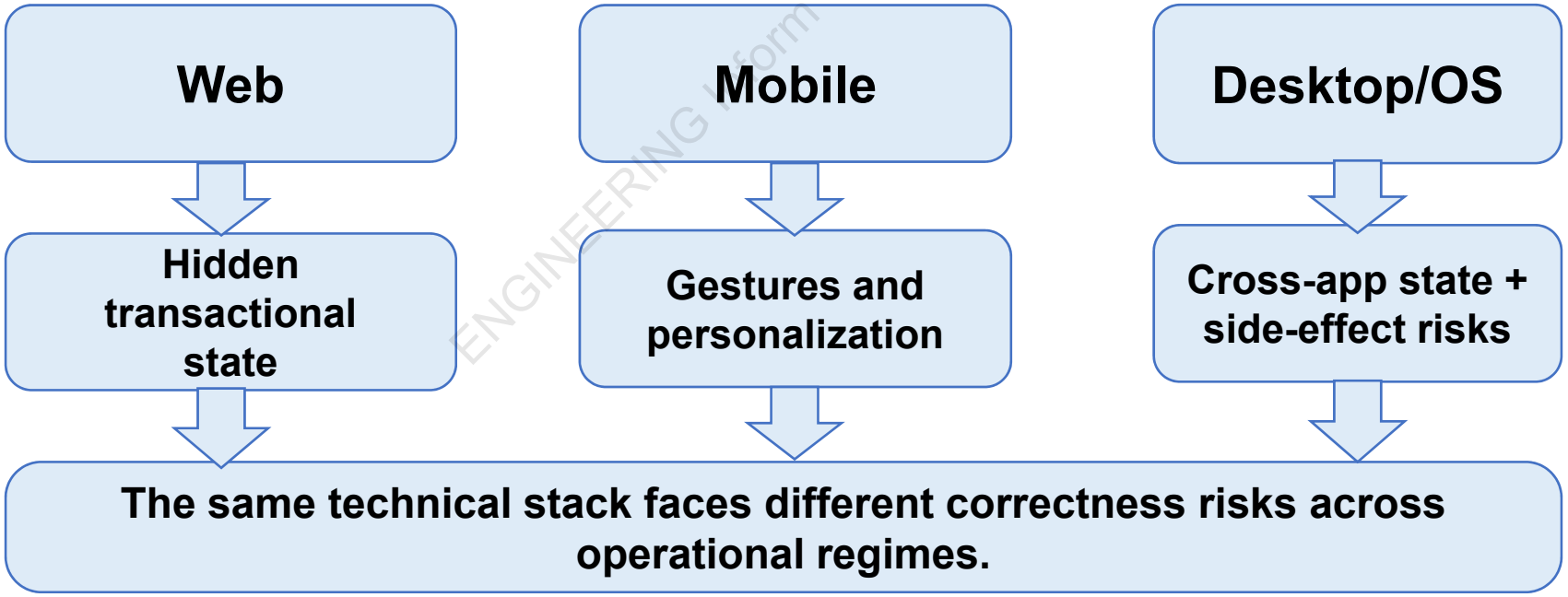
A platform-aware survey of execution, verification, and risk of graphical user interface agents

Key words: Graphical user interface (GUI) agents; Operational regimes; Trustworthy evaluation; Postcondition verification

Yi YANG
E-mail: yangyics@zju.edu.cn
ORCID: https://orcid.org/0000-0002-0512-880X

Motivation

•GUI agents increasingly automate tasks across websites, mobile apps, and desktop software.
•The visible interface exposes only part of the state that determines task success.
•A next-action-prediction framing obscures distinct failure modes, verification needs, and side-effect risks.



Core idea

- Shared stack: observation, grounding, planning, memory, execution, and verification.
- Distinct regimes: observability, hidden-state dependence, recovery needs, and side-effect risks.

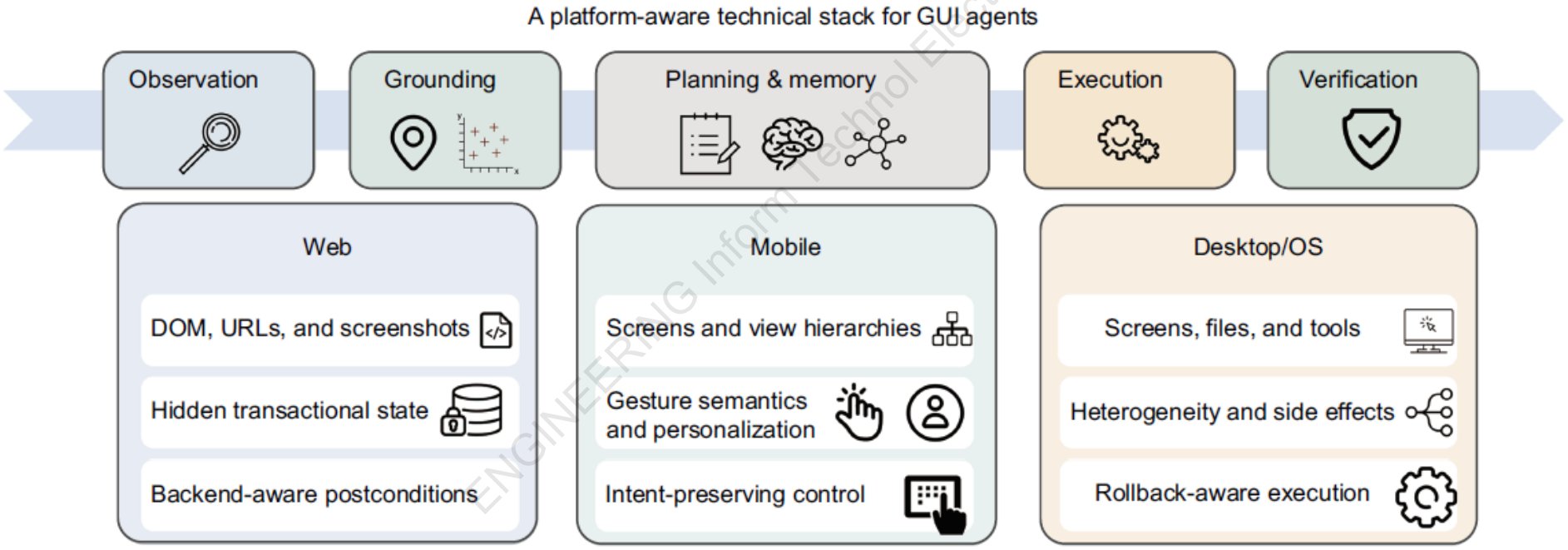


Fig. 3 A platform-aware technical stack for GUI agents. The high-level control loop is shared across platforms, but the main sources of uncertainty and the needed recovery strategies differ by operational regime

Core claim: GUI agents share a technical stack, but not a single correctness problem.

Review method

The survey reviews representative works through a shared technical stack and treats platform differences as an analytical variable.

Dimension	Review design
Scope	Large-model GUI agents across web, mobile, and desktop/OS
Coverage	2023 onward; search updated May 10, 2026; selected precursors included
Source	Google Scholar, arXiv, ACM Digital Library, IEEE Xplore, ACL Anthology, and OpenReview
Inclusion	Benchmarks, datasets, models, frameworks, evaluation protocols, and public technical/product reports
Analytical lens	Observability, action semantics, hidden-state dependence, verification, execution cost, and side-effect risk

Benchmarks and resources

Resources are grouped by what they measure: static supervision, interactive environments, and outcome-oriented evaluation.

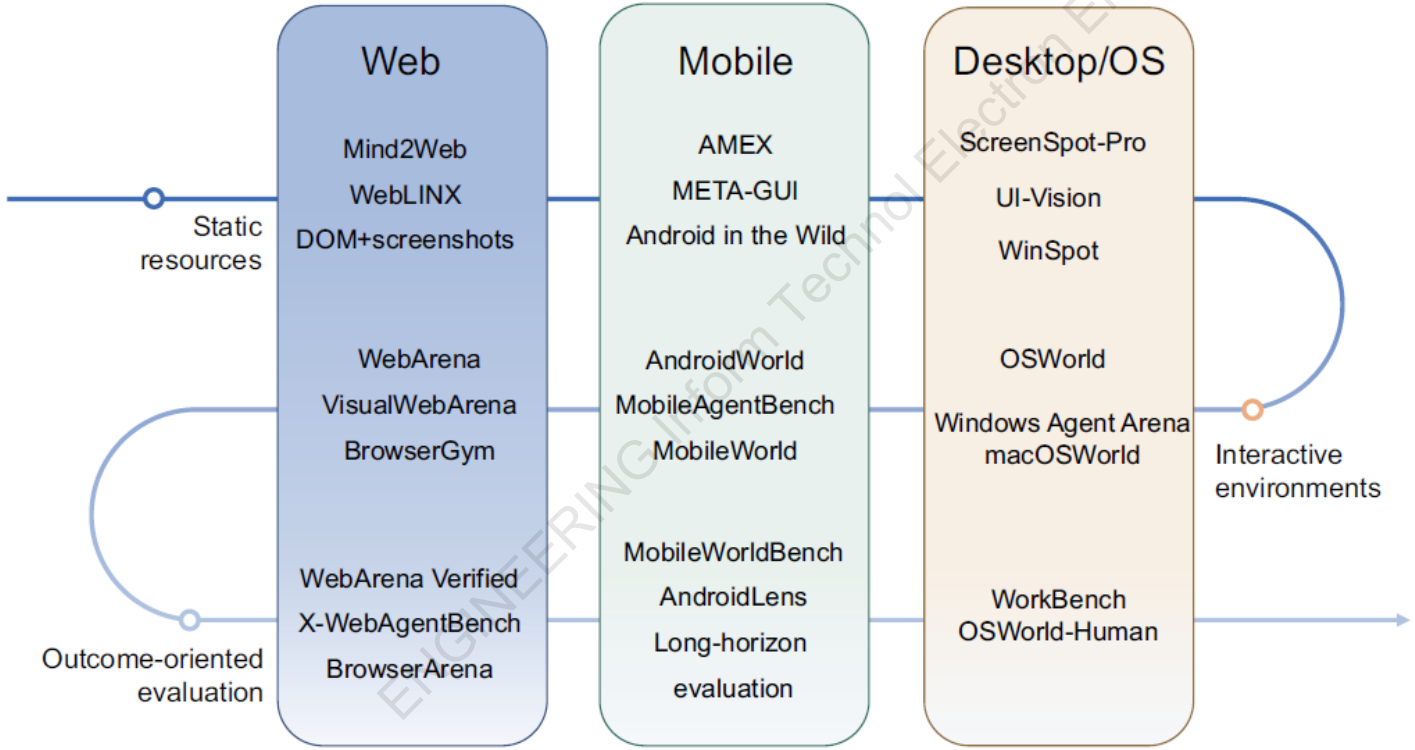


Fig. 2 Benchmark and resource map across the three operational regimes. The organization separates static supervision, interactive environments, and outcome-oriented evaluation instead of listing benchmarks as a flat list

The field is moving from next-action prediction toward executable tasks and verified outcomes.

Model and training design space

The survey compares interface representations, control strategies, and training signals, while emphasizing their integration.

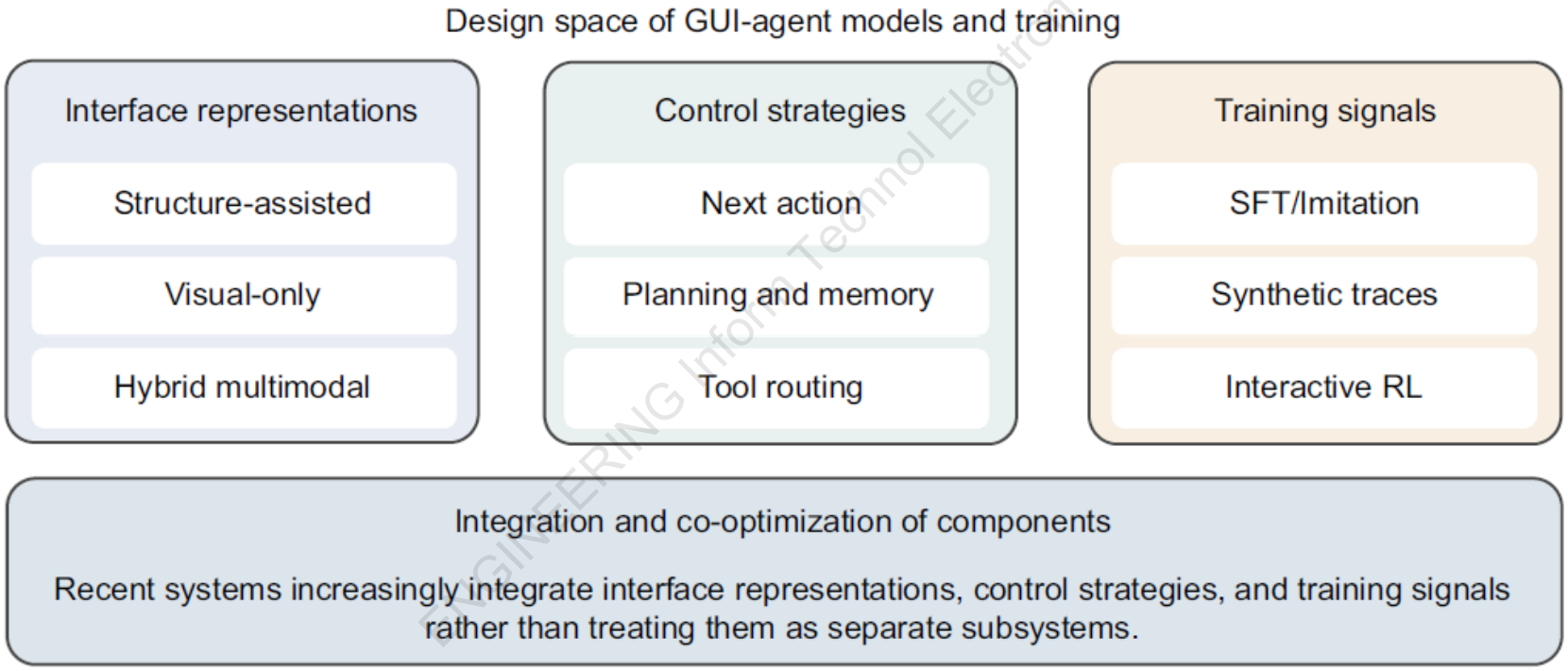
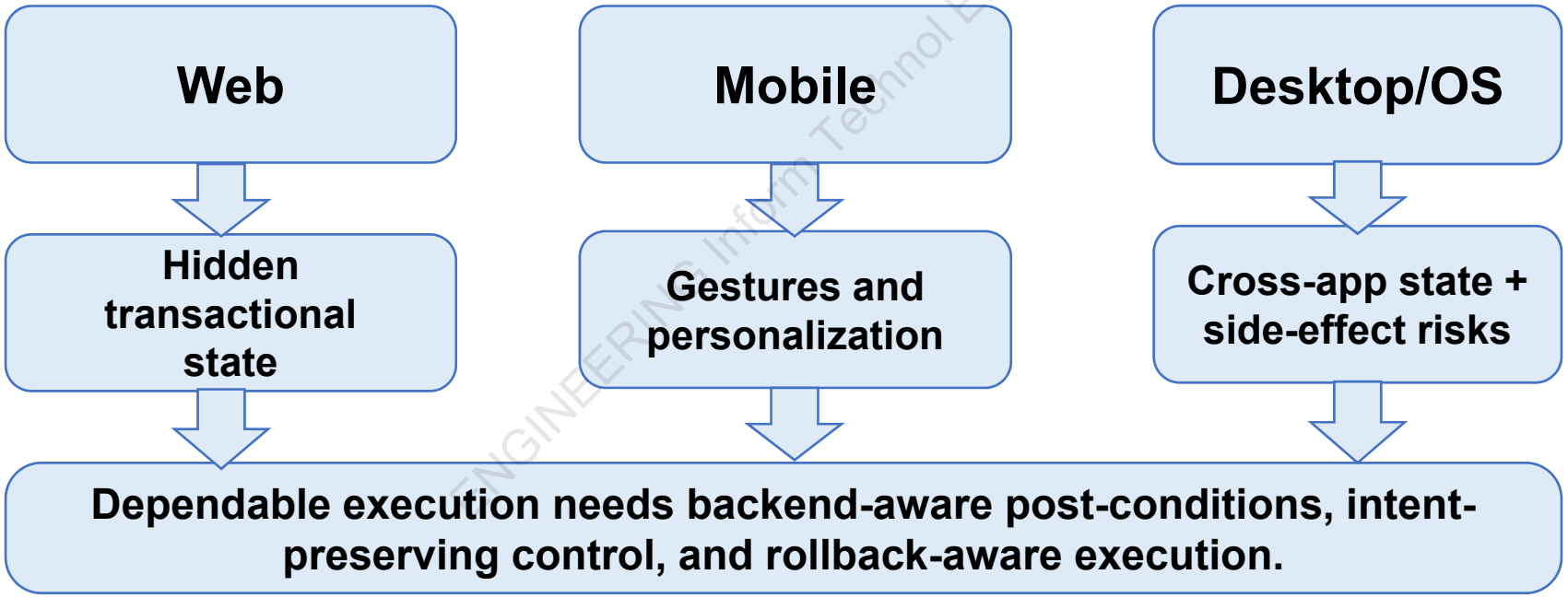


Fig. 4 Design space of GUI-agent models and training processes. The diagram groups recent systems by interface representations, control strategies, and training signals, and highlights the integration and co-optimization of these components

Analytical lens: interface representations × control strategies × training signals.

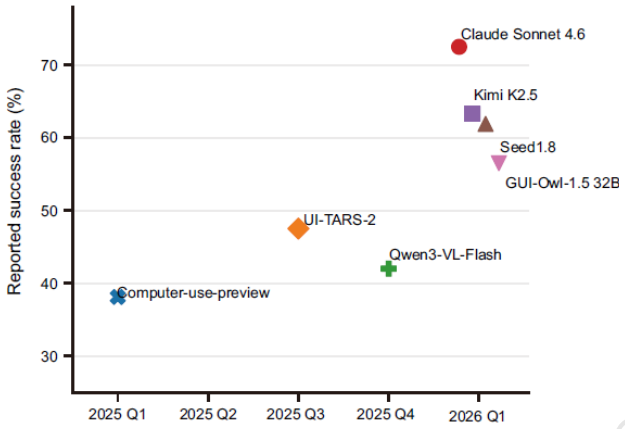
Major finding

Uncertainty concentrates in different system states, helping explain why benchmark performance transfers less effectively across operational regimes than headline scores suggest.

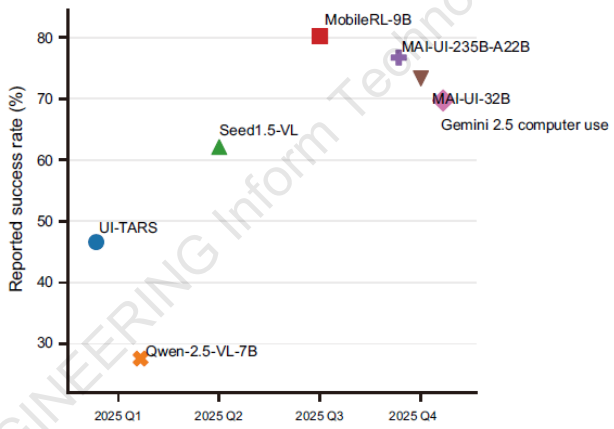


Benchmark scores are system-level outcomes

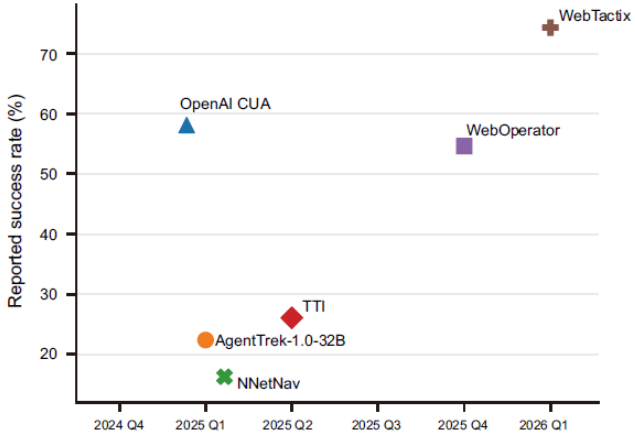
- Public reports show rapid gains across the OSWorld, AndroidWorld, and WebArena benchmark families.
- These snapshots are not a unified leaderboard: observation contracts, tool access, retry budgets, and verifier designs differ.



OSWorld family



AndroidWorld family



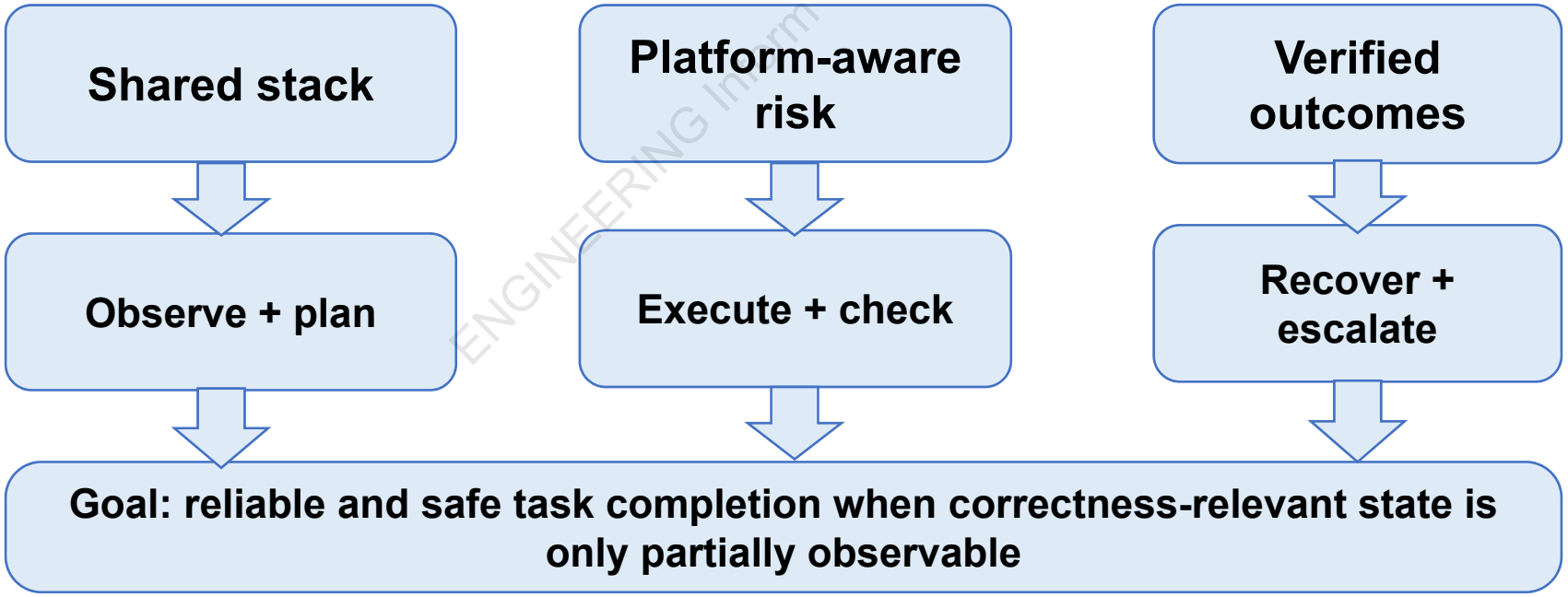
WebArena family

Future outlook

Research direction	From bottleneck to evidence of progress
State-aware execution	Hidden state and false local success → belief tracking, postcondition checks, runtime monitoring, recovery → outcome evidence, calibrated uncertainty, lower false-success rate
Adaptive sensing	Costly or incomplete observations → selective zooming, structure requests, tool queries, evidence thresholds → fewer observations, lower latency, more stable success rate
Hybrid automation	Fragile GUI-only execution and side effects → GUI–tool arbitration, APIs/scripts, sandboxing, rollback, confirmation → fewer irreversible errors, clearer audit trails, safer recovery
Human oversight	Risk, uncertainty, and irreversibility → staged autonomy, escalation, confirmation, evidence display → predictable and controllable use

Conclusions

- GUI agents share one technical stack but not one uniform correctness problem.
- Benchmark scores are system-level outcomes shaped by observation privilege, tool access, evaluator fidelity, execution fidelity, and recovery cost.
- Dependable computer use requires state-aware verification, adaptive sensing, GUI–tool arbitration, rollback, and human oversight.



Yu WU serves as Vice Dean, Professor, and Doctoral Supervisor at the School of Artificial Intelligence, Wuhan University. He is an expert on the expert panel for the implementation plan of National Key R&D Programs and a recipient of the National High-Level Young Talent Honor. He presides over multiple national-level research projects, including key sub-projects under the National Major Project for New-Generation Artificial Intelligence (AI). His research mainly focuses on multimodal large models. He has published more than 80 papers in top-tier academic conferences, with over 8000 citations on Google Scholar and over 1000 citations for a single paper. He has long served as Area Chair for top international AI conferences such as NeurIPS, CVPR, and ICML, and acted as General Chair and core organizer of CVPR 2023. His accolades include the 2020 Google PhD Fellowship (only 10 global recipients), the 2024 AAAI Academic Star Award, and championship awards in seven international academic competitions hosted by CVPR and ICCV.

Yi YANG is an IEEE Fellow and IAPR Fellow, Qiushi Chair Professor of Zhejiang University, and a Distinguished Expert of the National Special Recruitment Program. He currently holds multiple administrative positions: Director of the Institute of Artificial Intelligence, Zhejiang University; Director of the Microsoft–MOE Key Laboratory of Visual Perception; Deputy Director of the Provincial–Ministerial Co-Innovation Center for AI; Chairman of the Digital Entertainment and Intelligent Generation Committee of China Society for Image and Graphics (CSIG). He also leads numerous major national research assignments, including the national major project under the 2030 “New-Generation AI” Initiative of the Ministry of Science and Technology and key special projects of the National Natural Science Foundation of China. He has long dedicated himself to fundamental theories of AI and interdisciplinary applications, achieving a series of breakthrough outcomes across multiple research areas. His publications have accumulated more than 100 000 Google Scholar citations with an H-index of 158, and he has been listed on Clarivate Analytics’ Highly Cited Researchers global list for eight consecutive years. His research achievements have won wide recognition worldwide, earning him dozens of prestigious honors: National Excellent Doctoral Dissertation Award (Ministry of Education), First Prize of Zhejiang Natural Science Award (2015), Gold Medal for Disruptive Innovation from Australian Computer Society (2016), Google Research Scholar Award (2016), Australian Lifetime Achievement Award in Research (2019), Amazon Machine Learning Research Award (2020), AAAI Most Influential Paper Award (2021), Exclusive Best Paper Award of ACM MM (2023), First Prize of CSIG Natural Science Award (2023), First Prize of Zhejiang Science and Technology Progress Award (2024), the First Zu Chongzhi Award for Frontier Innovation in AI (2024), First Prize of Wu Wenjun AI Science and Technology Award in Natural Science (2025), alongside over 30 world championships in international research competitions.