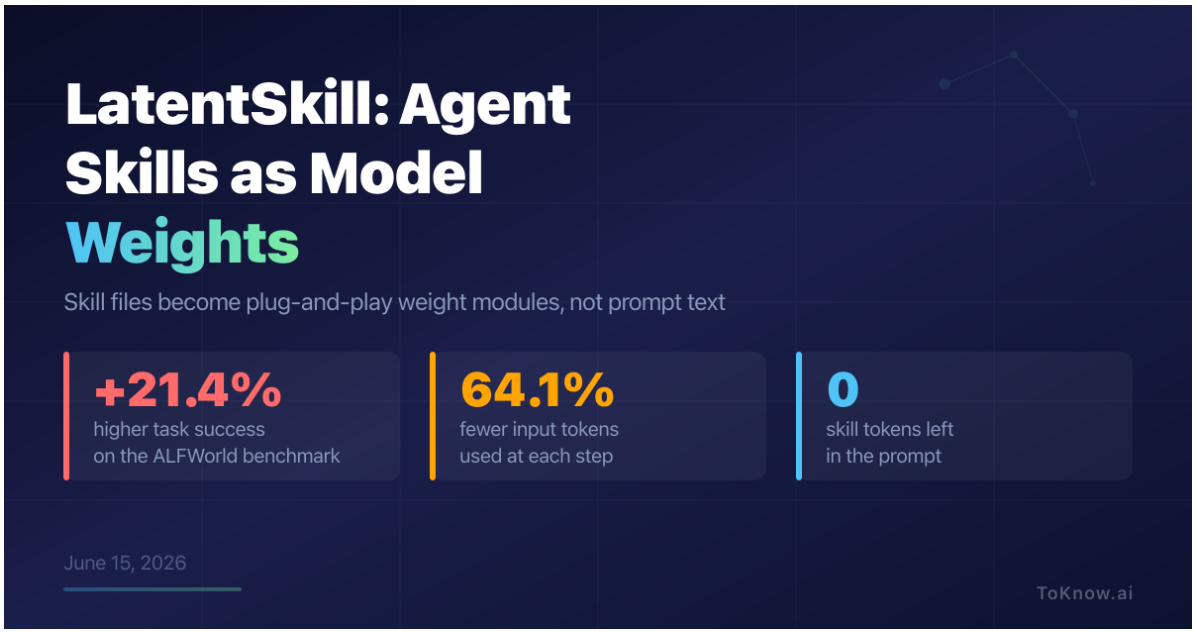
LatentSkill: Storing Agent Skills as Model Weights, Not Prompt Text

Kabui, Charles

2026-06-15

Read at ToKnow.ai



Researchers from Shanghai Jiao Tong University and OPPO Research Institute released LatentSkill, a method that moves an AI agent's skills out of the prompt and into the model's weights. Agent skills are reusable text instructions, like tool-use patterns and recovery steps, that tools like Claude Code paste into the prompt. But that text gets re-inserted every step, burning input tokens and sitting in the prompt as readable plaintext. LatentSkill trains a

skill compiler that reads a skill once and generates a small LoRA adapter, a bundle of extra weights mounted on a frozen model. On ALFWorld, a household-task benchmark, it raised success by 21.4% on familiar tasks and 13.4% on new ones with 64.1% fewer input tokens, and on a search-and-answer benchmark it lifted accuracy by 3.0% using 72.2% less skill-token overhead.

Re-reading the same skill on every step gets expensive at scale. A swappable weight module means no skill tokens in the prompt while keeping the plug-and-play feel: load it, swap it, or dial its strength up or down. Plain fine-tuning also pushes skills into a model, but bakes them in permanently and makes them hard to remove or combine. These adapters stay separate, and the skill no longer sits next to untrusted input, which cuts the prompt-injection surface.

It points to a third path between bloated prompt skills and frozen fine-tuning: skills as weight modules you compile once and combine with simple arithmetic. The same procedures now shared as text in agent skill marketplaces could ship as weights instead.

Sources:

- LatentSkill: From In-Context Textual Skills to In-Weight Latent Skills (arXiv)
- LatentSkill full paper (arXiv HTML)
- LatentSkill on Hugging Face Papers
- LoRA: Low-Rank Adaptation of Large Language Models (arXiv)
- ALFWorld: Aligning Text and Embodied Environments (arXiv)
