Bian Que: An Agentic Framework with Flexible Skill Arrangement for Online System Operations

Bochao Liu*, Zhipeng Qian*, Yang Zhao*, Xinyuan Jiang*, Zihan Liang, Yufei Ma, Junpeng Zhuang, Ben Chen†,

Shuo Yang, Hongen Wan, Yao Wu, Chenyi Lei, Xiao Liang

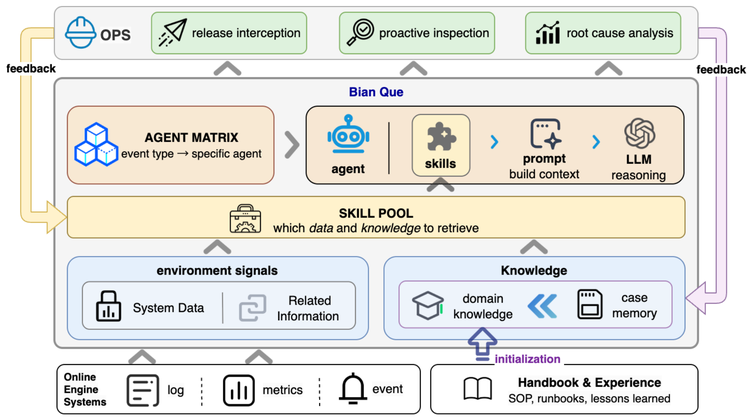
Kuaishou Technology * Equal contribution † Corresponding author: benchen4395@gmail.com

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Online engine operations span dozens of daily releases. The key bottleneck is selecting the right data and operational knowledge for each event.

Architecture



An event selects an Agent and matched Skills. Skills assemble logs, metrics, changes and domain knowledge for diagnosis.

Three operational patterns

Release interception

Check release events and identify unsafe changes.

Proactive inspection

Analyze multiple events to uncover latent system risks.

Alert root cause analysis

Correlate evidence to explain an alert and guide response.

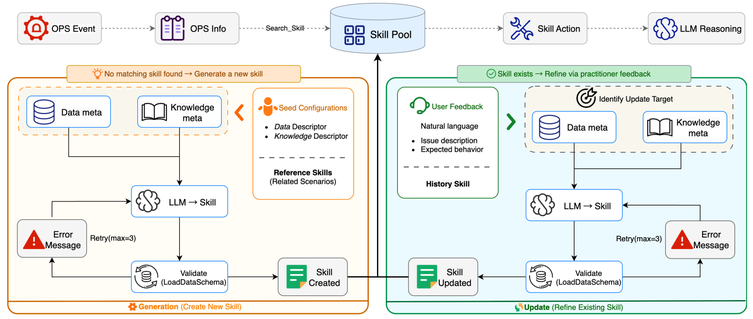
Core idea: separate scenario logic in Agents from business-module context in reusable Skills.

Flexible Skill Arrangement

Agents generate and revise Skills. Engineers refine them through natural-language feedback without code edits.

Skill field	Role
LoadDataSchema	Data calls and knowledge queries
Prompt	Business-module reasoning steps
Meta	Name, version and matching tags

Skill lifecycle



An unseen key triggers generation. Schema execution checks source access and required fields, with up to three retries.

One correction, two pathways

Knowledge pathway

Store corrected cases immediately. Distill reusable experience into long-term knowledge and consolidate daily.

Skill pathway

Revise retrieval schemas or reasoning prompts based on the diagnosed failure and practitioner instructions.

Prompt-only edits require subsequent SRE review.
Schema validation does not verify diagnostic correctness.

Production impact

Six-month deployment on an e-commerce search engine

75%
fewer human-facing alerts

80%
root cause analysis accuracy

>50%
reduction in mean time to resolution

95%
diagnoses returned within 5 minutes

Descriptive before–after observations. P0 alerts are never suppressed. Persistent or unhandled cases escalate to SREs.

Offline evaluation

104 real events: 44 alerts and 60 inspections

Method	pass@1	pass@5
Fixed Skills	65.4%	83.7%
No knowledge retrieval	71.2%	86.5%
Bian Que	**78.8%**	**94.2%**

Original single run. pass@k counts a case as solved if at least one of k autonomous samples succeeds. Three-run means: 77.9 ± 2.5% at pass@1 and 94.6 ± 0.6% at pass@5.

99.0%
cumulative accuracy with SRE correction

103/104 cases after up to five autonomous generations and three correction rounds on residual failures.
This result includes human assistance.

Code: github.com/benchen4395/BianQue_Assistant