E-COMMERCE

FAQ Surge Assistant for E-commerce Sellers

Shopping Festival Scenario · Executive Brief

Prepared by

Rex Chang

GenAI & Cloud Solution Architect / Program Lead



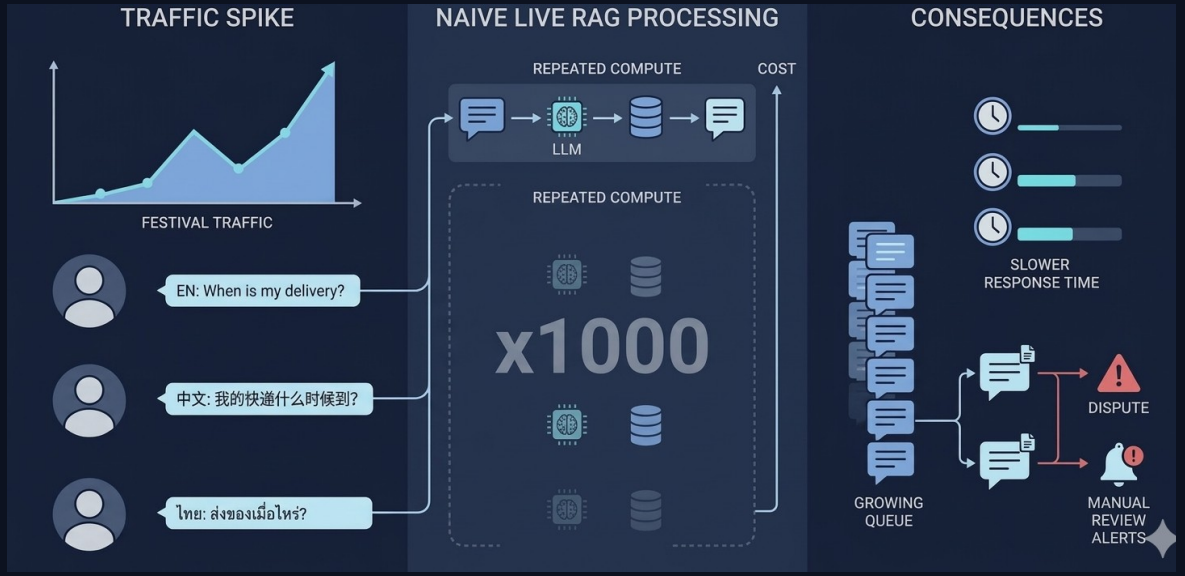
Peak shopping traffic breaks traditional support economics

THE PAINS

- Peak shopping events can sharply increase multilingual support demand.
- Most questions repeat, but live answering still pays full cost each time.

⚙ MAIN ISSUE

Treating every repeat as a new request can make peak support slow and expensive.



A controlled pipeline replaces ad-hoc answering

ROUTE

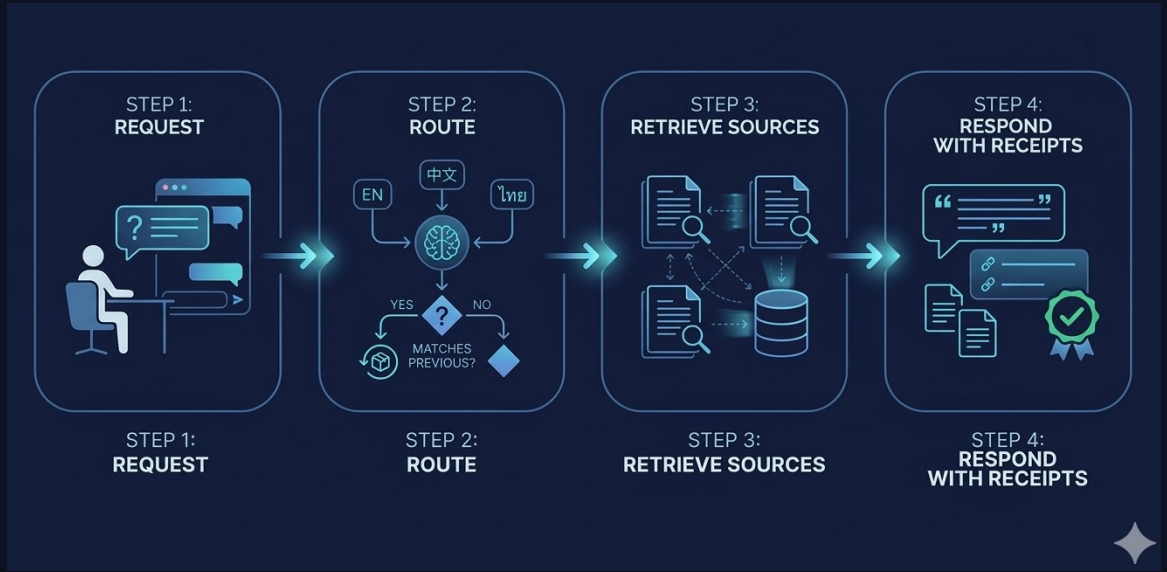
Use the selected language's support content.

SOURCE

Retrieve supporting information.

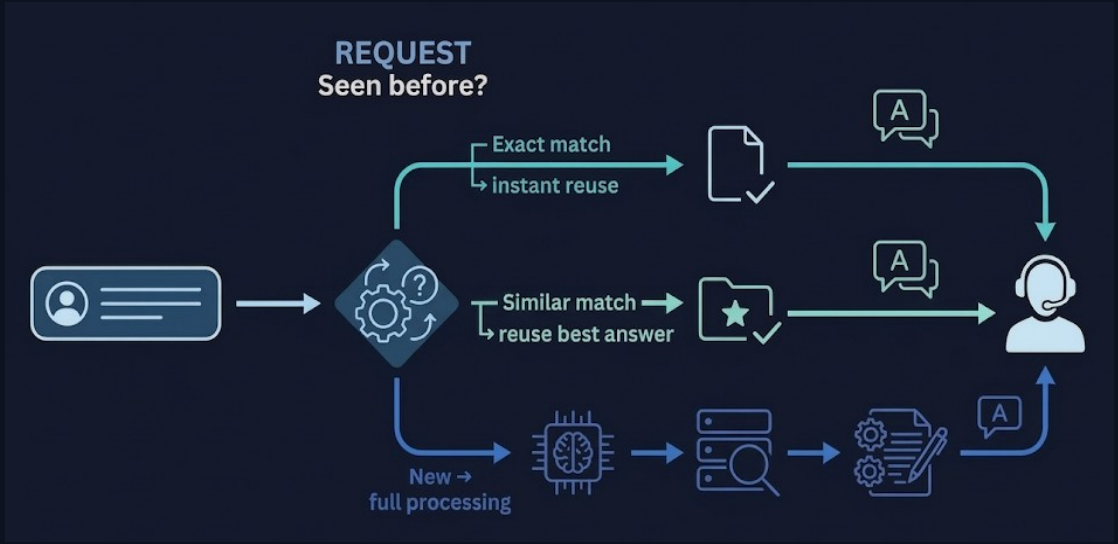
TRACE

Return source links and route details.



Make the answer path inspectable.

The system avoids redoing work for repeat questions



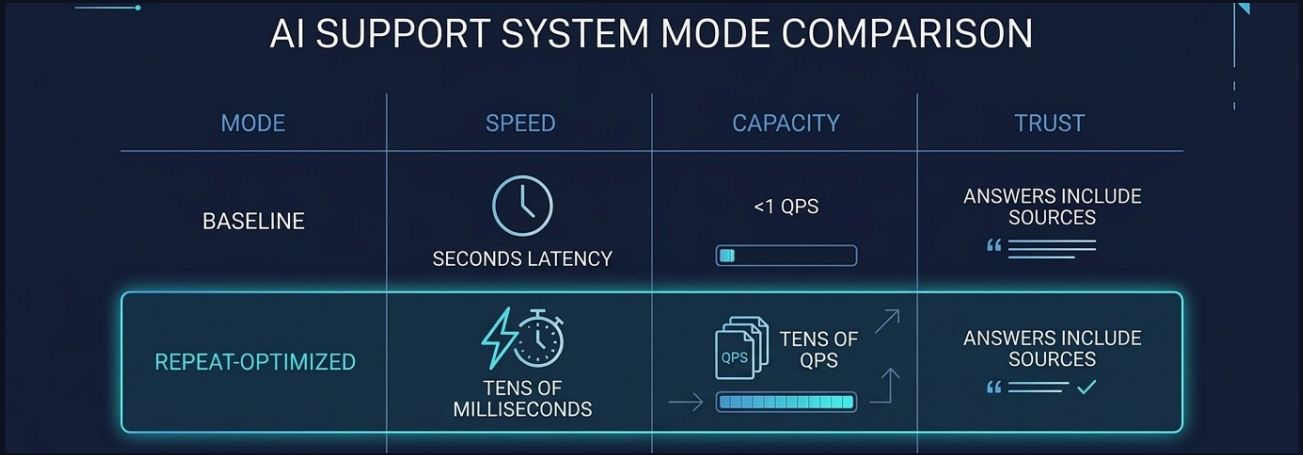
During peak events, the same questions appear thousands of times.
The system reuses prior answers by two layers:

1 **Exact question repeat (L1)**
→ Reuse immediately

2 **Similar question (L2)**
→ Reuse closest matched answer

Reuse matching answers to avoid repeated generation.

100×+ Faster Replies to Repeat Questions



BASELINE

seconds; <1 QPS

REPEAT PATH

p95 12–14 ms; 34–43 QPS

REVIEWABILITY

Source links and route details

Primary-path benchmark after warm-up; 20 requests per setup. Source links support review.

Thank You

Rex Chang

GenAI & Cloud Solution Architect / Program Lead

PORTFOLIO	LINKEDIN	EMAIL
rexchang.dev	linkedin.com/in/rex-chang-tw	rex@rexchang.dev

GenAI Systems Portfolio