

From Model to Harness, Close the Loop: Building Reliable AI Agents on Alibaba Cloud

2026. 09. 07. 11:05

MKIK Prompt Aréna

ENG · 40 perc

**Xiao Yang**Technical Service Expert · Alibaba Cloud
Central Europe

Xiao Yang, az Alibaba Cloud Közép-Európa műszaki szolgáltatási szakértője elmagyarázza a megbízható AI-ügynökök valós alkalmazásokban történő létrehozásának követelményeit, hangsúlyozva a robusztus rendszerhüvely szükségességét.

MI VOLT TÉMA

Xiao Yang hangsúlyozza, hogy a modell önmagában nem elegendő a magas kockázatú feladatokhoz, mivel a felhasználóknak képesnek kell lenniük ellenőrizni, hogy a rendszer hogyan jut el egy adott következtetéshez. A prezentáció az ügynökök cselekvéseit kockázati szintek szerint kategorizálja, megjegyezve, hogy míg egy dokumentum összefoglalása alacsony kockázatú, egy fizetés jóváhagyása jelentős pénzügyi következményekkel jár, amelyek emberi felügyeletet igényelnek. Például egy ügynököt fel lehet hatalmazni egy e-mail megfogalmazására felülvizsgálatra, de nem szabad engedélyezni a fizetés végrehajtását ellenőrzési lépés nélkül.

A funkcionális AI-ügynök lényege a meghajtóban rejlik, amely eszközöket, kontextust, ellenőrzéseket és teszteket tartalmaz. Xiao Yang bemutatja, hogy egy keresőeszköz hozzáadásával az ügynök hozzáférhet az adatbázis információihoz, miközben a világos kontextus biztosítja, hogy a modell megértse a konkrét részleteket, például a pénztípusokat. A prezentáció azt a következtetést vonja le, hogy az explicit szabályok és a szigorú tesztelés kötelező ahhoz, hogy az ügynök pontos ítéleteket hozzon, és az ember felé terelje a bizonytalan kérdéseket. Például beállítható egy szabály, amely automatikusan elutasít minden olyan számlát, amely a megrendelési összeget több mint 2%-kal meghaladja.

FONTOS TÉNYEK

- A modell csak egy komponense a rendszernek, és egy kötéletre van szüksége az adatok összekapcsolásához és a szabályok érvényesítéséhez.
- Az ügynököknek úgy kell megtervezni, hogy a bizonytalanság vagy hiányzó információ esetén mindig emeljék a feladatot egy embernek.
- Az ügynök feladatának kockázati szintje meghatározza a szükséges emberi beavatkozás és ellenőrzés mértékét.

ÜZLETI HATÁS

Ezen építészeti elvek megvalósítása lehetővé teszi a vállalkozások számára, hogy AI-ügynököket alkalmazzanak összetett munkafolyamatokhoz, például számlázás és beszerzési megrendelés ellenőrzés. Eszközök és szabályok integrálásával a vállalatok automatizálhatják a rutinellenőrzéseket, miközben fenntartják a pénzügyi biztonságot és az adatok pontosságát. Ez a megközelítés biztosítja, hogy a rendszer elég megbízható maradjon a termelési környezetekben, ahol a hibák közvetlen pénzügyi veszteséghez vezethetnek.

KÖVETKEZTETÉS

Xiao Yang azzal zárja, hogy felszólítja a közönséget, lépjen túl a egyszerű modellpromóciókon, és összpontosítson átfogó rendszerek építésére. Ajánlás, hogy eszközöket, egyértelmű kontextust és szigorú szabályalapú ellenőrzéseket építsenek be, hogy egy alapvető modellt funkcionális, megbízható AI-ügynökké alakítsanak.

From Model to Harness, Close the Loop: Building Reliable AI Agents on Alibaba Cloud

2026. 09. 07. 11:05

Prompt Arena

ENG · 40 perc

**Xiao Yang**Technical Service Expert · Alibaba Cloud
Central Europe

Xiao Yang, a Technical Service Expert from Alibaba Cloud Central Europe, establishes a framework for building reliable AI agents by emphasizing the necessity of a robust system harness beyond the base model.

WHAT WAS DISCUSSED

Xiao Yang explains that a model alone is insufficient for real-world applications because it lacks the ability to perform actions or verify data accurately. The presentation categorizes agent capabilities by risk levels, noting that while summarizing a document is low-risk, approving a payment involves significant financial risk and requires human oversight. For example, an agent might successfully summarize a text, but it requires a different level of verification before it can be allowed to send an email on behalf of a user.

The core of a functional AI agent lies in the "harness," which consists of tools, context, checks, and tests. Xiao Yang demonstrates how adding a lookup tool allows an agent to access a database, while providing clear context ensures the model understands specific details like currency types. The system must also incorporate hard rules, such as a 2% variance limit on invoices, and an escalation path to a human for uncertain queries. These elements combined transform a simple model into a reliable system capable of making governed judgments.

KEY TAKEAWAYS

- A model is only one component of a system and requires a harness of tools and rules to function in production.
- Risk assessment determines what actions an agent can perform autonomously versus what requires human intervention.
- Escalation protocols are essential for ensuring that the AI agent asks a person for help when it encounters uncertainty.

BUSINESS IMPACT

Implementing these structured harnesses allows businesses to automate complex workflows, such as invoice processing, while maintaining strict financial controls. By integrating database tools and specific business rules, companies can reduce manual labor and minimize errors in multi-currency transactions. The use of a 2% threshold for invoice approval serves as a concrete example of how automated checks can protect a company's bottom line while still allowing for minor discrepancies.

CONCLUSION

Xiao Yang concludes by urging the audience to move beyond simple model prompts and focus on building comprehensive systems. The recommendation is to systematically add tools, context, and rules to ensure the agent can perform specific tasks reliably. The presentation ends with a call to prioritize testing and human-in-the-loop escalation to ensure safety in real-world deployments.