THESIS REPORT

МИКРОСЕРВИС АРХИТЕКТУРТ СУУРИЛСАН ХИЙМЭЛ ОЮУН АГЕНТУУД

МУИС, МТЭС, МКУТ, Мэдээллийн технологи хөтөлбөр, 4-р түвшний оюутан Б.Раднаабазар

2025 оны 12-р сарын 09

Агуулга

01 Нэр томьёоны тайлбар

Мэдээлэл	Хүний мэдрэхүйн дамжиж хүний ой эрхтэн, ухаан, баримт, эд зүйл,бичиг зайд ямар нэг хадгалж байгаа дижитал орон зайд ямар нэгэн байдлаар үлдэнэ
Мэдээлэл	Тодорхой зорилгод чиглэсэж боловсруулсан, утга санаа илэрхийлсэн өгөгдөл
Өгөгдөл	Ямар нэг баримт (факт), статистик эсвэл мэдээллийн элемент бөгөөд тоон шинжийг агуулна
Эдийн засаг	Бараа, бүтээгдэхүүн (үйлчилгээ) үйлдвэрлэх, худалдах, худалдан авах хэлбэрээр орлого олж, ашиг олох үйл ажиллагаа.
Цахим бизнес	Интернэт, ICT-ийн тусламжтай онлайн орчинд явуулж буй бизнесийн үйл ажиллагаа;
Хиймэл оюун	Компьютерийн системийг хүний сэтгэн бодох, шийдвэр гаргах, суралцах чадвартай болгох технологийн салбар.
Суурь загвар (Foundation Model)	Олон төрлийн өгөгдөл дээр урьдчилан сургасан, олон даалгаварт ашиглагдах чадвартай хиймэл оюуны том загвар.
Том хэлний загвар (Large Language Model, LLM)	Хүний хэлний бүтэц, утгыг ойлгож, бичвэр үүсгэх чадвартай хиймэл оюуны загвар. Transformer архитектур дээр суурилдаг.

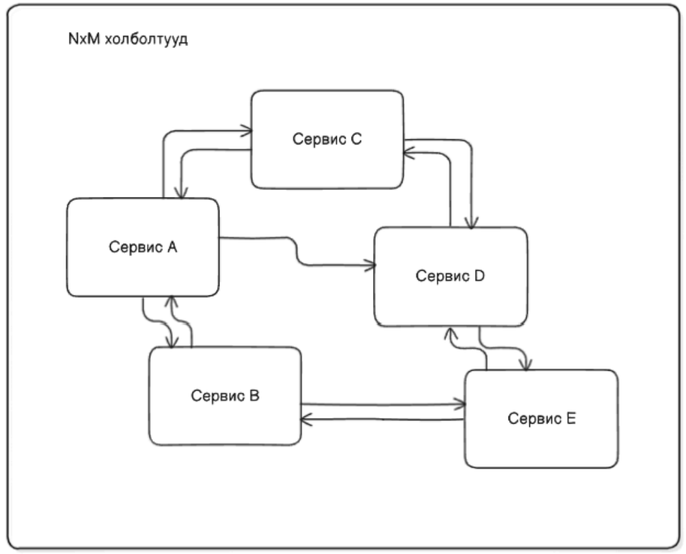
01 Нэр томьёоны тайлбар

Хиймэл оюуны инженерчлэл (AI Engineering)	Бэлтгэгдсэн суурь моделийг ашиглан бодит хэрэглээний аппликейшн, систем хөгжүүлэх хөгжүүлэлтийн арга барил.
Prompt инженерчлэл (Prompt Engineering)	Хиймэл оюуны моделд өгч буй зааварчилгаа, асуултыг оновчтой бичих замаар хүссэн үр дүн гаргуулах арга.
Хиймэл оюун агент (AI Agent)	Өөрийн орчныг мэдрэх, нөхцөл байдлыг ойлгож, төлөвлөгөө гарган, хэрэгсэл ашиглан үйлдэл хийдэг ухаалаг систем.
ReAct буюу Төлөвлөгч агент	Төлөвлөгч агент нь системийн төв тархи болж, хэрэглэгчийн хүсэлтийг ангилж, зохих агент руу чиглүүлэх үүрэгтэй. Энэ нь ReAct (Reasoning and Acting) загварын гол бүрэлдэхүүн юм.
RAG — Хайлтаар нэмэгдүүлсэн үүсгэлт	Хиймэл оюуны модел гадаад өгөгдлийн сангаас холбогдох өгөгдлийг хайж, хариулт гаргахдаа ашигладаг арга.
Үзэгдэлд суурилсан архитектур (EDA)	Үзэгдлийг хадгалж, боловсруулдаг программ хангамжийн зохиомжийн загвар
Микросервис архитектур	Том системийг жижиг, бие даасан, тусдаа ажиллах сервисүүдэд хувааж хөгжүүлдэг программ хангамжийн архитектур.
Монолит архитектур	Бүх функц, логик, өгөгдлийн сан нь нэг програмд нэгтгэгдсэн уламжлалт программ хангамжийн архитектур.

01 Оршил

- Хиймэл оюун нь компаниудын заавал нэвтрүүлэх ёстой бай болж байна. (Harvard Review, 2023)
- Хиймэл оюун агентууд цахим бизнесд шинэ дэвшилтэт боломж нээж, агент суурилсан хөгжүүлэлтийн чиг хандлага давамгайлж байна. (Huyen, 2024)

Уламжлалт олон агент систем



Асуудал

N×M холболт = Нягт хамаарал = Нэг агент унавал бүгд дамжин унана

Huyen, C. (2024). AI Engineering. O’Reilly Media.

Falconer, Sean. ”The Future of AI Agents is Event-Driven”.

01 Зорилго

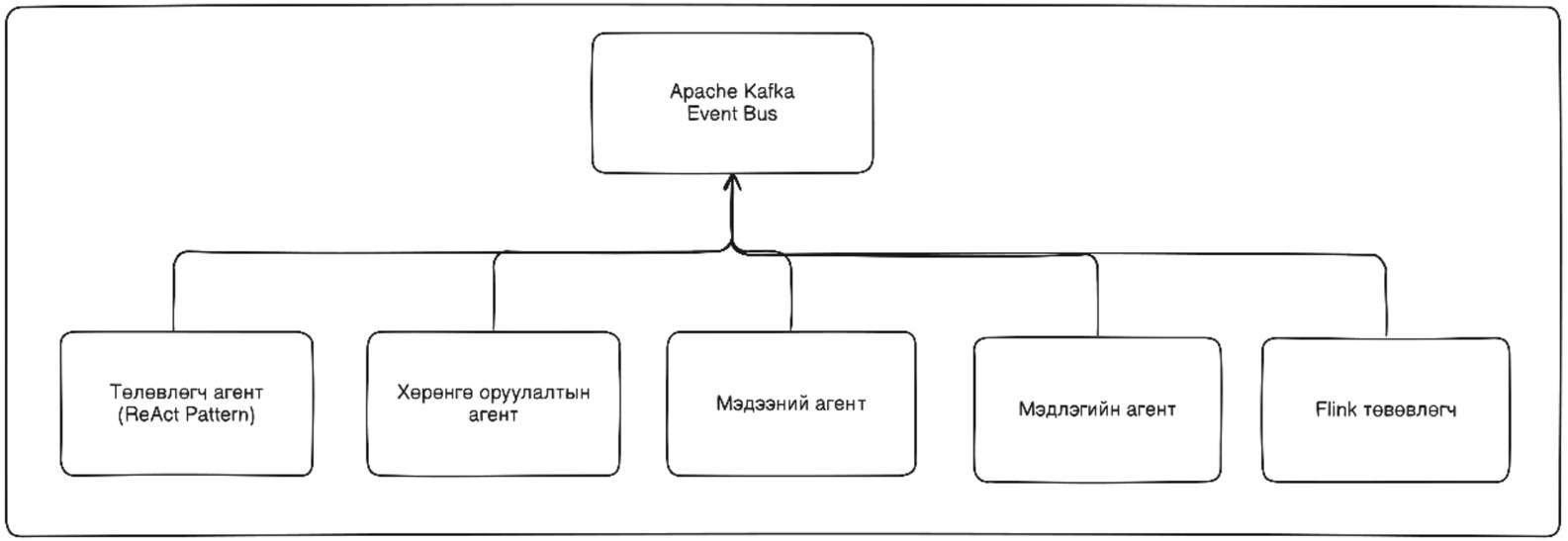
Тархмал системийн онцлогийг ашиглан хиймэл оюун агентуудыг бие даасан, хэвтээ чиглэлд хялбар өргөжих боломжтой, найдвартай системийн зохиомжийг гаргаж, бататгах

Зорилт

1. Хиймэл оюун агентуудыг тархмал микросервис болгон хөгжүүлэх
2. Үзэгдэлд суурилсан архитектураар (EDA) уян хатан системийн зохиомж гаргах
3. Монголын хөрөнгийн биржийн өгөгдөлд тулгуурлан демо систем бүтээх

02 EDA суурилсан олон агент зохиомж

Уламжлалт N×M холболтын асуудлыг шийдэхийн тулд үзэгдэлд суурилсан архитектур (EDA) ашигласан



N+M холболт = Тархмал = Бие даан ажиллана

Falconer, Sean. ” The Future of AI Agents is Event-Driven” .

Давуу талууд

- Асинхрон харилцаа - зэрэгцээ боловсруулалт
- Алдаа тусгаарлалт - нэг унахад бусад хэвээр
- Хэвтээ өргөжих - агент нэмэхэд хялбар
- Event лог- дахин тоглуулах боломжтой, аудит, тест

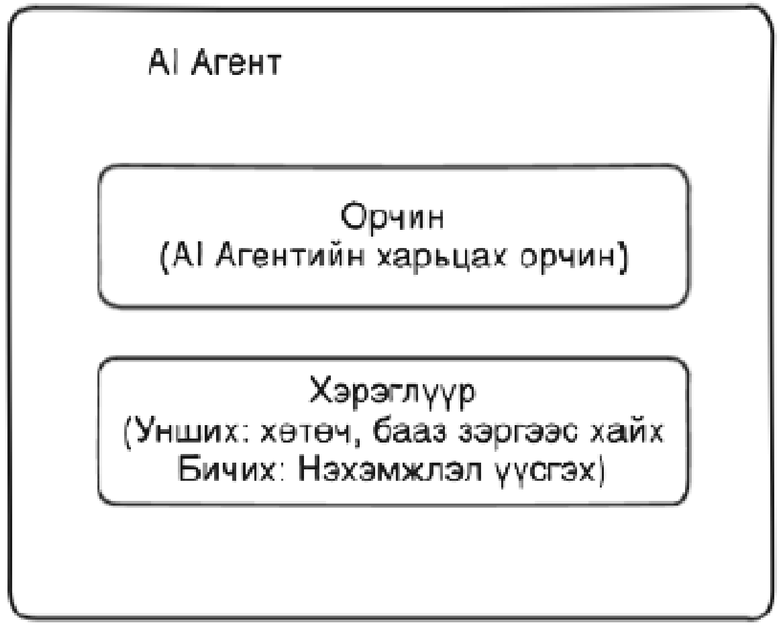
Олон агентийн шаардлага

Бизнесийн доторх үйл ажиллагаа ялгаатай байдаг тул агентууд адилаар ялгаатай байх хэрэгтэй байдаг.

02 Хиймэл оюун агент

Агент гэж юу вэ?

Өөрийн орчныг мэдрэх, түүн дээр үйлдэл хийх чадвартай систем



Huyen, C. (2024). AI Engineering. O’Reilly Media.

Үргэлжлэл

Хиймэл оюунд гадна орчинтой хандах хэрэглүүр өгөх нь маш олон боломжийг нээж өгдөг. .

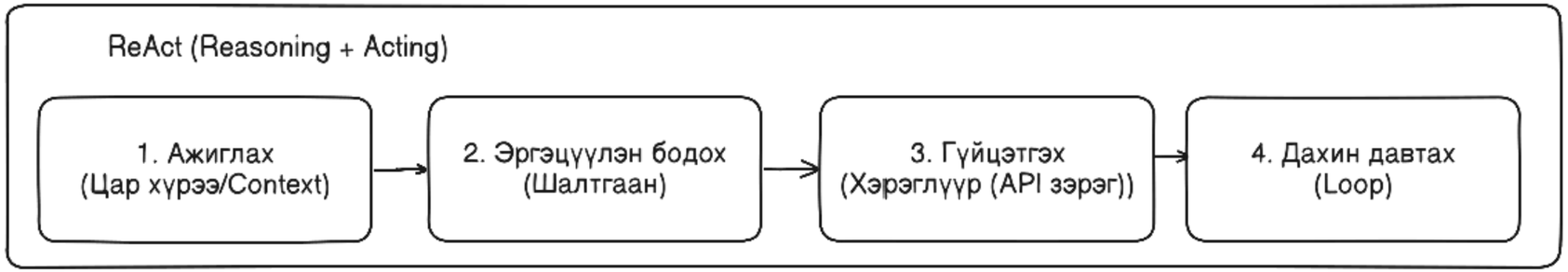
Жишээ хэрэглээ

- Хэрэглэгчийн туслах: Тусламжийн тикет боловсруулах, чиглүүлэх
- Хувийн туслах: уулзалт автоматаар товлох
- RPA: Нэхэмжлэл боловсруулах

02 ReAct буюу төлөвлөгч агент

Төлөвлөгч нь эхлээд асуултыг ойлгож (Reason), дараа нь зөв хэрэглүүр дуудаж (Act) процессоо гүйцэтгэдэг.

Олон агентуудыг уяж ажилладгаараа онцлогтой



Хэрэгжүүлэлт (Демо)

1. Хэрэглэгчийн хүсэлтийг ангилах (Нийт 6 ангилал)
2. Энгийн/Төвөгтэй эсэхийг тодорхойлох
3. Зөв агент руу чиглүүлэх (Динамик чиглүүлэгч)
4. Процессийг гүйцэтгэх

Huyen, C. (2024). AI Engineering. O'Reilly Media.

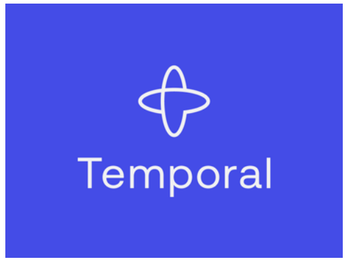
IBM (2025). What is ReAct agent?

02 Ижил төстэй системийн судалгаа

Хиймэл оюун агент сэдэвт ижил төстэй системийн судалгаа

Inngest

- Event-д суурилсан serverless workflow
- LLM дэмждэг
- ReAct төлөвлөгч байдаггүй
- Хаалттай эх, жижиг багт зориулагдсан



Temporal

- Төлөв хадгалалт сайтай
- Зөвхөн workflow-оор ажилладаг (өргөтгөл хязгаартай)
- Хаалттай эх
- Төлөвлөгч агент (ReAct)

Энэ судалгааны ажлын зохиомж

- AI-agent-first, тархмал микросервис
- Нээлттэй эх (Kafka, Flink), event log-д суурилсан
- Төлөвлөгч агент (ReAct)

ДЕМО

Монголын Хөрөнгийн Биржийн бодит өгөгдөлд тулгуурлан олон агент бүхий системийг хөгжүүлж, хэрэглэгчдэд монгол хэлээр хөрөнгө оруулалтын зөвлөмж өгөх

Технологийн стек: Docker, Apache kafka, PyFlink, NextJS v16, ExpressJs, Typescript

Агентууд:

1. Төлөвлөгч агент - Хүсэлт ангилаад, чиглүүлэх
2. Хөрөнгө оруулалтын агент - МХБ дата дээр шинжилгээ хийх
3. Мэдээний агент - өдөр тутам мэдээ явуулах, бүртгэх үед мэдээ явуулах
4. Мэдлэгийн агент - RAG систем ба дотор нь мэдээнүүд ба агентийн үүрэг байна
5. PyFlink төлөвлөгч - Нарийн даалгаврыг төлөвлөж гүйцэтгэнэ

04 Демо үр дүн

- 5 AI агент Kafka-аар харилцаж ажиллаж байна
- N×M → N+M холболтын нарийн төвөгтэй байдлыг бууруулсан
- ReAct pattern ашиглан ухаалаг чиглүүлэлт хийсэн
- Монгол хэлээр хувийн зөвлөгөө өгч байна

Гүйцэтгэл

- Кафка дамжуулалт: 5-10мс (10K+ мессеж/s)
- API Gateway: 200-500мс
- AI шинжилгээ: ~10-20 секунд
- Агент Uptime - 99%
- Нийт санах ой ашиглалт - ~123mb

ДҮГНЭЛТ

Энэхүү судалгааны ажлын хүрээнд хиймэл оюун агентуудыг микросервис архитектурт нэгтгэн хэрэгжүүлэх зохиомжийг боловсруулж, үр дүнг амжилттай ажиллаж буй демо системээр бататгалаа.

Хиймэл оюуны хоцрогдол, өртөг зэрэг хязгаарлалтууд байгаа бөгөөд цаашид агентуудын санах ойн удирдлага, олон агентын хамтын ажиллагаа зэрэг чиглэлээр судалгааг үргэлжлүүлэх боломжтой.

АНХААРАЛ ХАНДУУЛСАНД БАЯРЛАЛАА

Ном зүй

- Huyen, Chip. AI Engineering. O'Reilly Media, 2024.
- Vaswani, A., et al. "Attention Is All You Need". Advances in Neural Information Processing
- Falconer, Sean. "AI Agents are Microservices with Brains". March 2025.
- Falconer, Sean. "The Future of AI Agents is Event-Driven". BigDataWire, March 2025.