THESIS REPORT #1

АГЕНТ СУУРИЛСАН ХИЙМЭЛ ОЮУН

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

2025 оны 10-р сарын 03

А г у у л г а

01	Оршил
02	Хиймэл оюун
03	Foundation Model
04	Rag&Agent
05	Microservice

00 О р ш и л

Б и з н е с үү д э д

- AI нь компаниудын заавал нэвтрүүлэх ёстой бай болж байна. (Harvard Review, 2023)
- 2024 онд S&P 500 Компаниуд AI нэвтрүүлэх нь өмнөх жилийхээс 3 дахин их ихэслээ.
- AI engineering нь хамгийн хурдацтай өсөх бизнес боллоо. (Github дээр stars-ийн тоогоор тэргүүлж байна.)

A I e n g i n e e r i n g

- Харилцагчийн дэмжлэгийн бот нь голчуу хувиар
- Тэгээд бизнесийн үйл ажиллагааны цувралыг автоматжуулах Агент нь эрэлтээр ихэсч байна.

Байгууллага нь маш олон агент бүтээж байна. Жич: 20ш

Энэ нь ихэвчлэн стандартгүй монолит агент байна

Олон мулти агент нь тестлэхэд, нэвтрүүлэхэд, өргөжүүлэхэд хэцүү байна

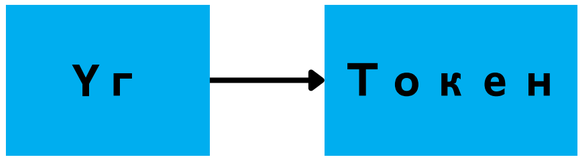
tightly-coupled

Иймээс бие даасан микросервис жсэн агент маш сайн өргөжихөөр байна

Event-Driven агент нь ирээдүйн бай болж байна

01 Хиймэл оюун

Language models



Токенжүүлэлт : Дунджаар токен нь үгийн ¾ урттай байна. Жич: 100токен=75үг

Language models-ийн үндсэн төрөл

1. Masked- Токен нөхөх — My favorite __ is blue (color)
2. Autoregressive- Дараагийн токен олох — My favorite color is ___ (blue>car)

Модел бэлтгэх үндсэн арга

1. Self-Supervision - Labeled датагаар бэлтгэгдэх (Өргөжихөд сайн, цаг, бас хугацааны хувьд үнэтэй)
2. Unsupervised - Ялгагдаагүй датагаао бэлтгэгдэх. (Алдаа ихтэй: Ценцургүй дата, буруу баримттай...)

Multimodal модел- Зураг, текст зэрэг зэрэгцээ ажилладаг модел

Foundation Model- Large language models, Largemultimodal model

01 Хиймэл оюун

AI Engineering vs. ML engineering

ML Engineering:	Data	→	Model	→ Product
AI Engineering:	Product	→	Data	→ Model

AI engineering

1. Бэлэн Foundation Model ашиглаад программаа хөгжүүлнэ. Модел хийх, сургах гэхээс илүүтэйгээр дасгах үйл хийнэ. (Adaptibility of LLM)
2. AI engineer-үүд илүү том хүч шаардсан модел дээр ажилладаг учир үр дүнтэй сургах, inference optimization хийх хэрэгтэй (Training:Finetuning, Sampling).
3. Open-ended гаралттай тул үнэлэхэд бэрхшээлтэй байдаг (Hallucination, inconsistency)

Г а р а л т :

RAG — Өөрийн дата дээр сургагдсан чатбот

AI Agent — Бизнесийн цуврал ажил хэргийг автоматжуулах агент

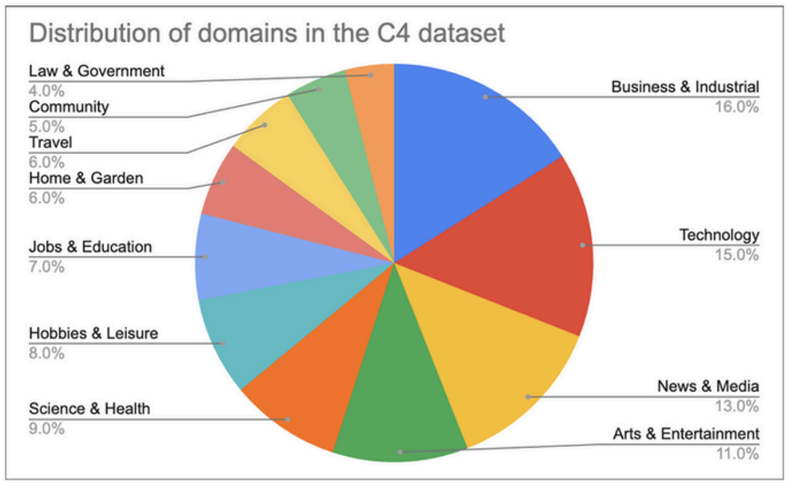
Category	Building with traditional ML	Building with foundation models
AI interface	Less important	Important
Prompt engineering	Not applicable	Important
Evaluation	Important	More important

02 Foundation Model -PreTraining

Сургалтын дата олох

Common Crawler ашиглах-

- 2022, 2023 онд нэг ашгийн бус байгууллага 2–3 тэрбум веб хуудас crawling хийсэн
- Бүх төрлийн дата байна. Байгууллагууд heuristic аргачлал ашиглаж муу чанарын датаг шүүдэг. (Худлаа мэдээлэл, хорт үзэл бодол) Жич: Reddit-ийн 5-аас илүү upvote-тэй датаг авах
- Сүүлийн жилүүдэд 7B модел нь 30B моделээс илүү үзүүлэлттэй байна



Domain-Specific Models

Тодорхой датасет ашиглана

drug discovery and cancer screening,
DNA, RNA data

-DeepMinds AlphaFold- trained on 3D structures of around 100,000 known proteins
-NVIDIAs BioNeM- bio molecular data for drug discovery.
- Googles Med-PaLM2 combined the power of an LLM with medical data to answer medical queries with higher accuracy.

02 Foundation Model -PreTraining

Multilingual Data

Language	Code	Pop.	CC size	
		(M)	(%)	Cat.
English	en	1,452	45.8786	H
Russian	ru	258	5.9692	H
German	de	134	5.8811	H
Chinese	zh	1,118	4.8747	H
Japanese	jp	125	4.7884	H
French	fr	274	4.7254	H
Spanish	es	548	4.4690	H
Italian	it	68	2.5712	H
Dutch	nl	30	2.0585	H
Polish	pl	45	1.6636	H
Portuguese	pt	257	1.1505	H
Vietnamese	vi	85	1.0299	H

Table 2-1. The most common languages in Common Crawl, a popular dataset for training LLMs. Source: Lai et al. (2023).

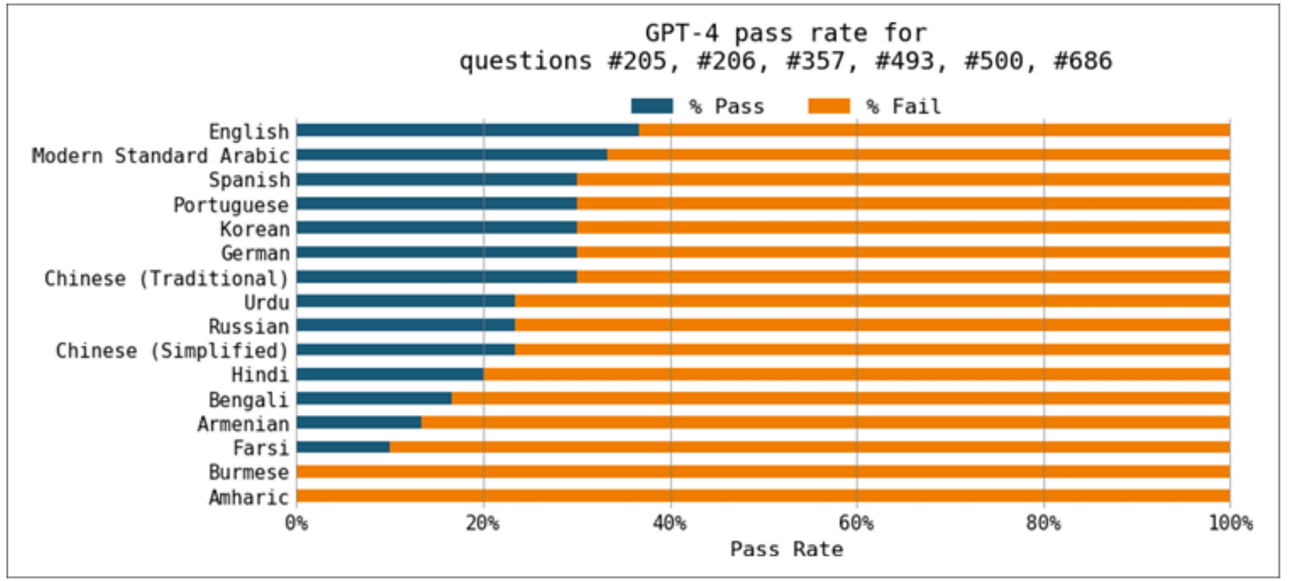


Figure 2-2. GPT-4 is much better at math in English than in other languages.

Уг хэлний бүтэц, соёл нь үр дүнд шууд нөлөөлж байна.

Шууд орчуулгаад хувиргах

- Олон хүмүүс хийж байна
- Гэхдээ орчуулах нь мэдээлэл алдагдуулж байна (Эзэн бие алга болох)
- Хятад хэлээр худал мэдээлэл гаргах нь илүү байна (Англи- 1/7, Хятад- 8/8)
- Дундаж токений урт- Англи 7, Энэтхэг 32, Burmese 72

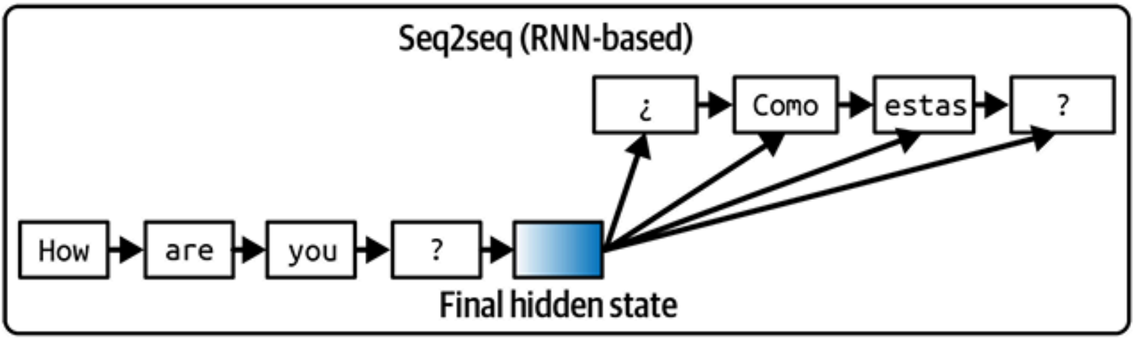
→

Зарим шийдэл

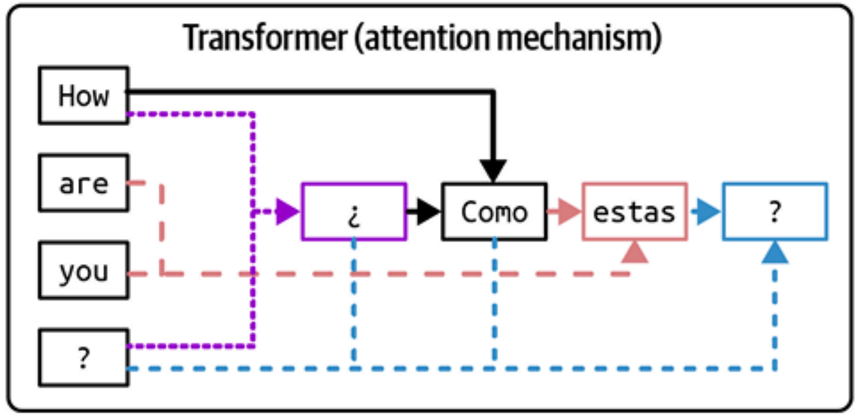
- Уг хэлд сургагдсан модель ашиглах
- Хятад (LLama-Chinse)
- Франц (Croissant-LLM)
- Виетнам (PhoGPT)

02 Foundation Model -PreTraining

Моделийн архитектур

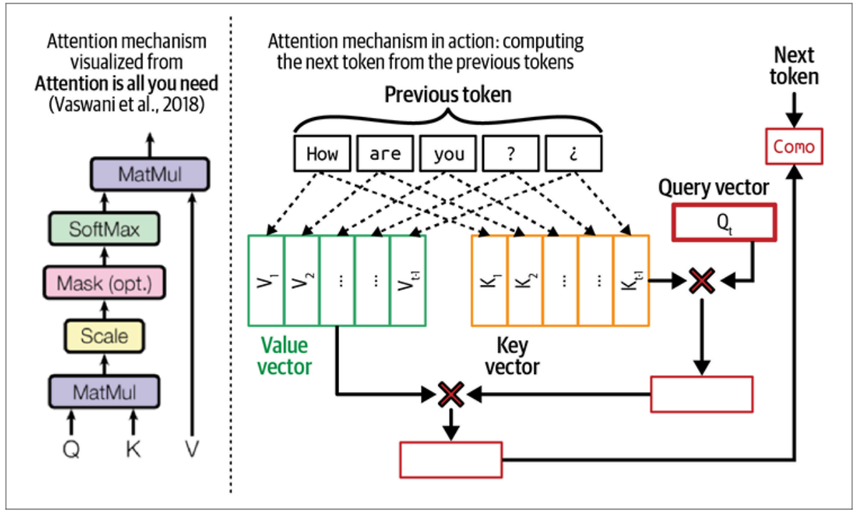


2016онд нэвтрүүлсэн (Google translate)
Параллель процест ганц engine-тэй
тул удаан хүлээгддэг



Өмнөх токеноо ашиглаж
үр дүнтэй хариу гаргах

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



02 Foundation Model -PostTraining

Post-Training нь хүнд тохирсон хариу гаргах ба 2 алхамтай:

1. **Supervised finetuning (SFT) -** Гүйцээлтээс илүү interactive байдлаар харилцан яриад сургах (Заавар бэлдэх, behaviour cloning хийх) - Demonstration Data-г гаргадаг хүний 90% нь эрэгтэй байна. Bias-тай байж болзошгүй. Иймээс зарим баг хүний гаргаснаас зааврааас илүү хиймэл оюуны зааврыг ашиглаж байна.
2. **Preference Finetuning** - Хамгийн чухал хэсгийн нэг. Дата ямар байдлаар шүүгдэх процесс нь энэ. Зорилго нь буруу датаг дамжуулахаас хамгаална, мөн хүнд тохирсон хариулт өгнө (Хөгжилтэй, Interactive, зөв). Хөгжүүлэлтэд RLHF Голчуу ашиглана. Хүний гаргах Онооны системээр сургагдах үйл явц юм.

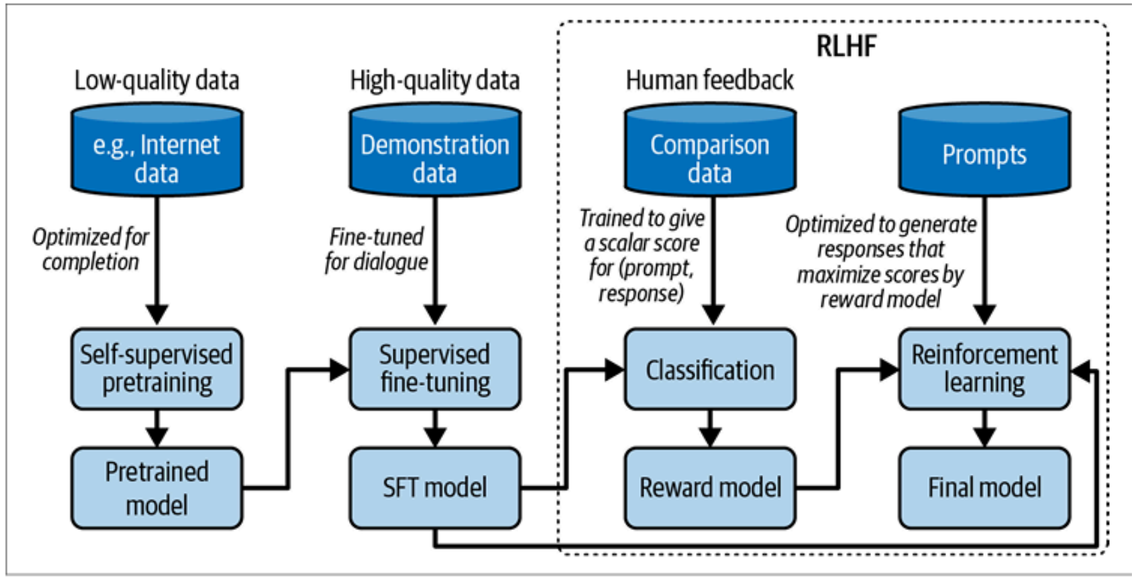
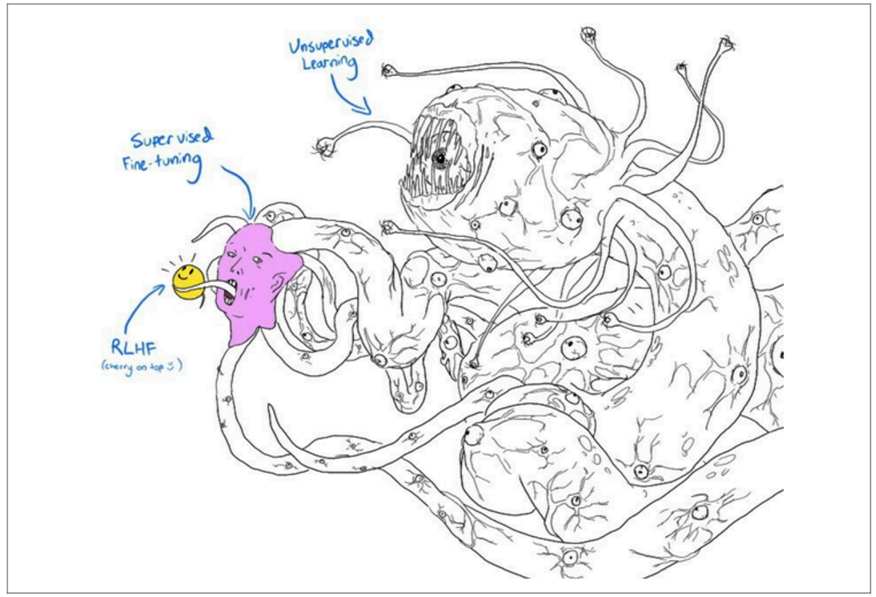


Figure 2-10. The overall training workflow with pre-training, SFT, and RLHF.



02 Foundation Model -PostTraining

Sampling - Магадлалт хариу боловсруулалт

Модел гаралтаа sampling гэх процессээр гаргадаг. Энэ нь магадлал дээр суурилагдсан байдаг. Нэгэн жигд хариу өгөхөөс сэргийлж Sampling нь олон стратегээр хариуг хөгжилтэй болгох, таамаглагдашгүй хариу гаргуулж болно.

Sampling Temperature

Түгээмэл токенг багасгаж, ховор токен гарах магадлал их болгодог арга.
Top-k: Тогтмол к өгч тооцооллын ажлыг багасгадаг.
Top-p: Гарч болох хариултыг динамикаар зааглах. (Тийм, үгүйгээс урт текст)

Бүтэцлэгдсэн гаралт

Энэ нь хариуны формат боловсруулахад хэрэгтэй. Дараах 2 тохиолдолд хэрэгтэй:

1. Хүний хэлээр API -тэй харьцах. Жич: RAG-аар хүн асуултаа асууж Tabular-аас SQL хэлээр хандаж дата авах
2. Хариултыг тодорхой форматны дагуу гаргах. Жич: REGEX ашиглах.

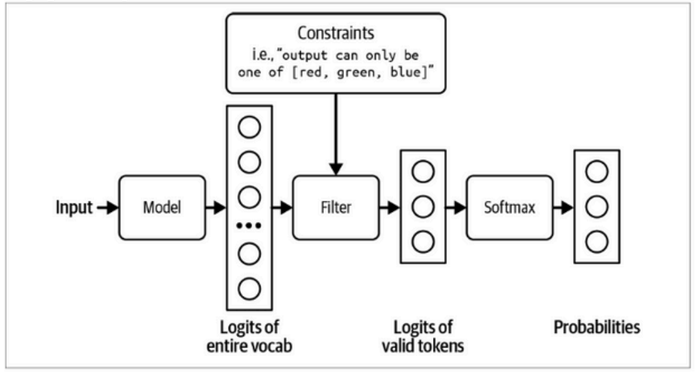


Figure 2-21. Filter out logits that don't meet the constraints in order to sample only

02 Foundation Model -PostTraining

Асуудал

Хиймэл оюун нь магадлалт суурилсан тул inconsistency, hallucination үүсдэг. Энэ нь хортой байна.

1. Inconsistency- Адилхан эсвэл ойролцоо асуулт нь маш өөр хариу гаргах (Моделийн чанартай шууд холбоотой)
2. Hallucination- Модел нь баримтад тулгуурлаагүй хариулт өгөх

Hallucination-ий 2 шалтгааны улмаас үүсдэг:

1. Snowball hallucination үзэгдэл- Анхны таамаглалаа буруу гаргангуут дараа нь энэ таамаглалаа буруугаар дэлгэрүүлдэг. Шийдэл: Хэрэглэгчийн асуултыг салгаж ReinforcementLearning хийх; Supervised Learning дээр анхаарах.
2. Лабел гаргагч хүн, бэлтгэгдсэн моделийн мэдлэг өөр байна. Модел нь мэдэхгүй, лабел гаргагч хүн нь мэдэх нөхцөл юм. Үүнийг яг тулгаж тааруулж сургах нь практикийн хувьд боломжгүй

03 Моделүнэлэх

- Модел нь хар хайрцаг шиг байна. Зөвхөн гарч байгаа гаралтаас шалгалт хийж болдог. Иймээс үнэлэх аргууд маш олон байдаг.
- Модел нь хүчтэй болсноор, эрсдэл дагаад ихэснэ. Иймээс үнэлэх нь чухал ба маш олон багууд үнэлж байна. Хиймэл оюунаар үнэлэх, хүнээр үнэлэх аргууд тус тус байна. Хүнээр үнэлэх нь маш чухал байдаг.
- Моделуудыг тус тусдаа оноогоор үнэлэж, сонголтоо хийх нь зөв аргачлал болж байна.

04 Prompt Engineering

Моделийн жинг өөрчлөлгүй зааврын дагуу хүссэн үр дүнгээ авах аргыг хэлнэ:

Best practice:

- Хоёрдмол утгагүй заавар ашиглах
- Моделд персона өгөх
- Жишээ өгөх- Хоёрдмол хариултыг багасгаж, хүссэн хариу авах
- Контекст өгөх- Дотоод мэдлэг нь hallucination үүсгэх эрсдэлтэй тул контекст өгч үүнийг сэргийлж болно
- Том даалгаврыг хэсэгчлэх

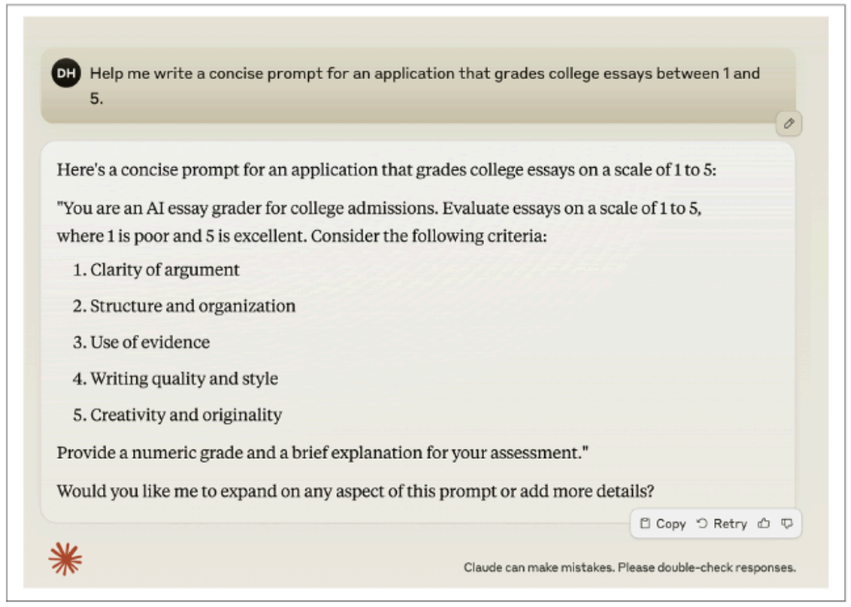


Figure 5-7. AI models can write prompts for you, as shown by this prompt generated by Claude 3.5 Sonnet.

Моделээр заавар боловсруулч болно.

05 RAG&Agents

RAG (Retrieval Augmented Generator) бүтэц.

- Retrieval- Документээс асуултын дагуу холбоотой хариултыг олж авах
- Generator- Гаргасан хариултаар PROMPT ENGINEERING хийж үр дүн гаргах.

Ихэнх хүмүүс их өгөгдөл нь RAG-ийг чадлыг хязгаарладаг гэж үздэг.

- Өгөгдөл устдаггүй, тогтмол нэмэгддэг
- Өгөгдөл нэмэгдэх тусам цаг, зардал ихэсч, хайлтын оновчлолыг багасгадаг.
- Иймээс зөв тактик ашигласнаар хүссэнээр өргөжүүлж болно.

Retriever-ийн чанар нь RAG-ийн үр дүнтэй шууд холбоотой. Дараах үндсэн функцтай

1. Indexing- Боловсруулахыг хүссэн датагаа индекслэх
2. Querying- Документээс холбоотой датаг хайх

05 RAG&Agents

Retrieval Algorithm

Term-Based Retrieval

- Түлхүүр үгээр хайлт хийх
- Google Search зэрэгт ашиглагдсаар ирсэн
- Elasticsearch, BM25 гэсэн 2 түгээмэл аргатай.
- Elasticsearch нь inverted index ашигладаг ба үгийн давтамж, n-gram холбоосоор хайдаг

Embedding-Based Retrieval

- Семантик утгаар хайлт хийх
- Орчин үеийн RAG, өргөжүүлэлтэд үр дүнтэй.
- Үр дүн нь Embedding моделийн чанартай шууд холбогддог
- Өгөгдлийг вектор баазд хийх амархан ч вектор хайлт хийх нь бэрхшээлтэй байдаг

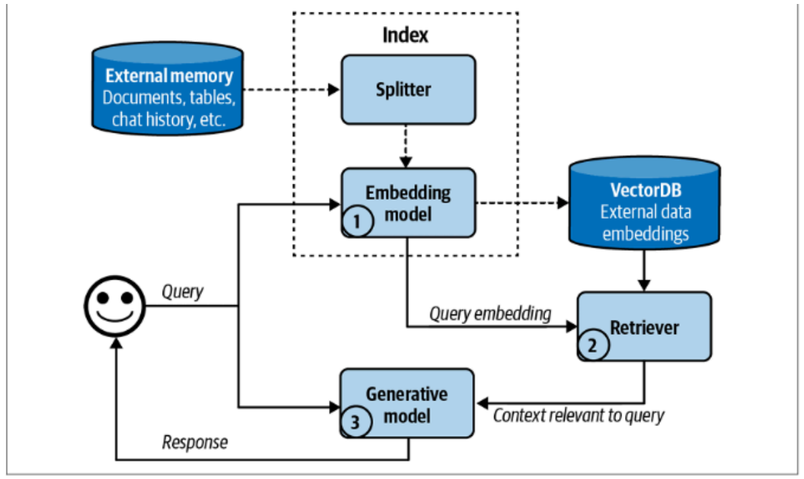


Figure 6-3. A high-level view of how an embedding-based, or semantic, retriever works.

05 RAG&Agents

Вектор хайлт

Жирийн вектор хайлт:
Ерөнхийдөө хамгийн ойр байгаа хөрш хайх аргачлал (k NN)

- Документийг хэсэгчилж вектор баазд оруулах
- Бүх векторд оноо өгөх
- Хамгийн өндөр оноотой к векторуудыг буцаах

Энэ нь тооцоолоход удаан урт текстийг дундуур таслах магадлалтай тул үр дүн муутай байдаг.

Ахисан вектор хайлт:
Векторыг өгөгдлийн бүтцийн олон төрөлт хувиргаад үр дүнтэйгээр хайна. Хэш, граф, кластард задлах, мод бүтэц рүү хөрвүүлэх.
Алдартай вектор хайлтын сангууд

1. FAISS (Facebook AI Similarity Search)
2. Google’s ScaNN
3. Spotify's Annoy
4. HNSwlib

05 RAG&Agents

Agents

Agent нь орчин, хэрэглүүрүүдтэй байж идэвхтэй ажилладаг. Үндсэн модел нь чанартай байх ёстой.

Хэрэглүүр:

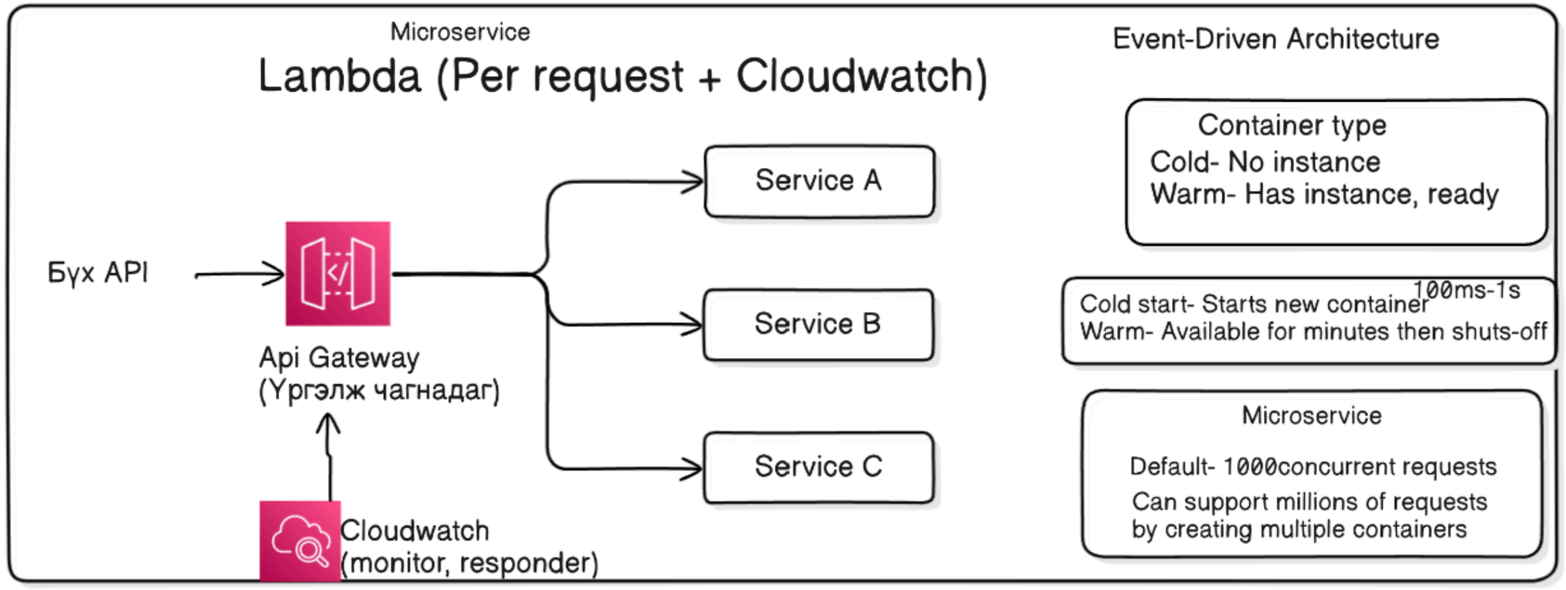
- Мэдлэгийн аугмент: Текст авах, Зураг авах, SQL хэрэгжүүлэгч, болон бусад мэдээлэл авах API байж болно.
- Чадлын сунгамж: Заримдаа модел нь тоо хуваах үйлдэл зэрэг хийхэд буруу хариу гаргадаг. Иймээс тооны машины API гаргаад холбох зэрэг зүйл хийнэ.
- Үйлдэл гаргагч: Дата нэмэх, устгах, өөрчлөх зэрэг хийнэ. Мөн цаашлаад банкны гүйлгээ хийх, хэрэглэгчдэд бүтээгдэхүүн санал болгох зэрэг хийнэ

06 Microservice

Том багууд Монолитийг ашиглахад дараах асуудал тулгарч байсан.

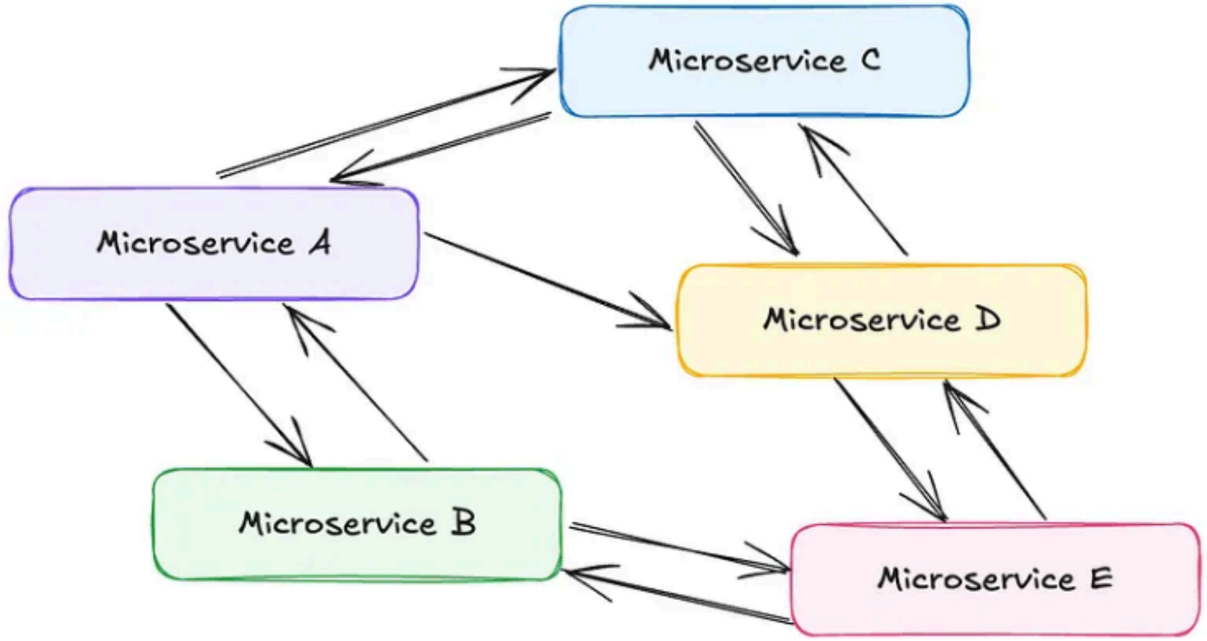
- Бүх бизнес логик tight-logic-ээр холбогдсон тул жижиг өөрчлөлт хийхэд маш удаан бүтэн build-ээ ачааллуулж асдаг байсан.
- Олон харилцагч дааж чадахгүй болсон.
- Программын хэмжээ ихэссэн

Иймээс том компаниуд Microservice хэрэгжүүлж эхэлсэн. Хамгийн алдартай нь AWS-ийн ламбда сервис



06 Microservice

Микросервисийг өргөжүүлэхэд дараах асуудалтай тулгарсан. Сервис хооронд NxM дууддаг тул хүлээлт үүсч, олон холболт даахаа больсон.



Microservices with Rigid Dependencies

06 Microservice

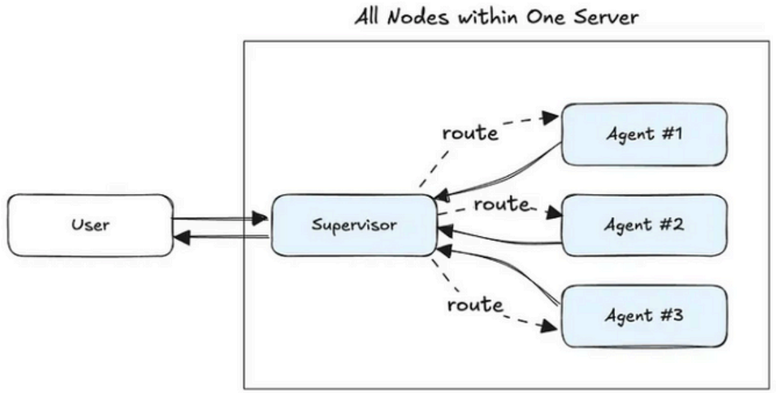
Иймээс EDA Буюу Event broker гаргаж олон хүсэлтийн холболтыг асуудалгүй явуулдаг болов.

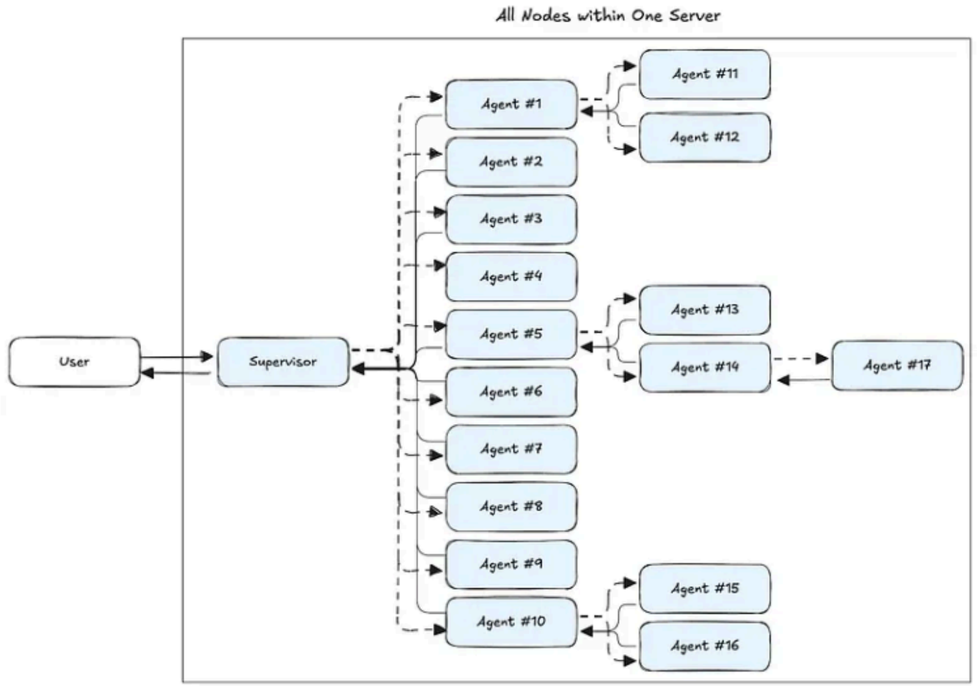
Үр дүнтэй ажиллахын төлөө программ олон хэсэгт задлагдах, өргөжих, event-driven суурьтай байх чанартай болов.

06 Microservice

Монолитик AI Agent

Баг AI-ийг нэвтрүүлэхдээ маш олон AI agent гаргаж болно. Энэ нь нэг серверийн орчинд tight-coupling-ий шалтгааны улмаас эвдрэл үүсэх өндөр эрсдэлтэй байдаг.





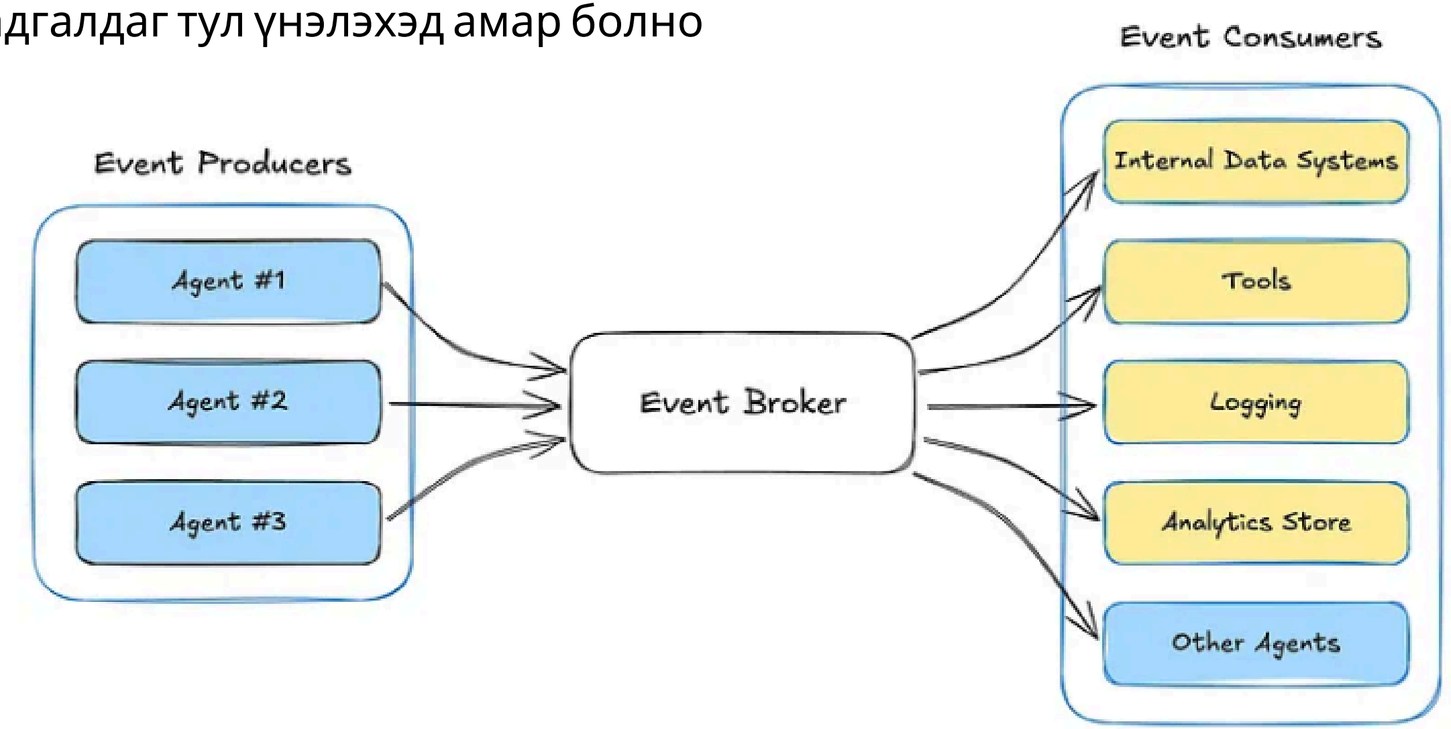
Agent Dependency Graph Quickly Growing Out of Control

06 Microservice

Монолитик AI Agent

Иймээс microservice-ээр салгах нь илүү найдвартай ажиллана

- Өргөжих чадвар - Тус тусдаа ажиллана
- Уян хатан байх чанар- Шинэ агент бусад процессд саад болохгүйгээр ажиллана. Нэг нь Унасан ч бусад нь унахгүй
- Динамик Пайплайн- динамикаар процессийн дарааллаа өөрчилж болно
- EDA нь лог хадгалдаг тул үнэлэхэд амар болно



АНХААРАЛ ХАНДУУЛСАНД БАЯРЛАЛАА