

# Biz user를 위한 AI agent 활용방법

## 목 차

0. 이 문서 download방법
1. AI agent란 무엇인가?
2. LLM(Large Language Model)
3. LLM만을 이용한 질의 처리
4. Generation과 Understanding
5. RAG(Retrieval Augmented Generation)
6. RAG와 embedding model
7. AI agent 관련 문서 및 사용 web 주소
8. RAG적용 사례
  - a. 간단한 문서의 RAG 적용
  - b. embed model에 따른 query결과
  - c. 영어와 한글 질의에 대한 차이
  - d. LLM 답변이 항상 맞는 것은 아님.
  - e. RAG에서 memory mode이용
9. RAG 활용시 주요 고려사항 및 추가 활용 사례

- a. 시사성이 높은 문서를 예측에 이용할 경우 주의 요망
- b. drill down 예1 – 서울시 환경정책
- c. drill down 예2 – 서울시 자율주행 관련계획
- d. drill down시에 고려할 사항
- e. drill down 예3 – 이재명정부 123대 국정과제
  - 기초연구 생태계 조성
- f. RAG와 similarity score

#### 10. AI agent에서 mcp tool의 이용

- a. mcp tool 사용시 고려사항
- b. mcp tool 사용 예 – ppt file내용 검색,  
query의 표현이 살짝 다름
- c. LLM only와 mcp tool을 agent가 판단해서 실행

#### 11. AI agent에서 web tool의 사용

- a. web tool 시용예1 – 최근 뉴스1
- b. web tool 시용예2 – 최근 뉴스2
- c. web tool 시용예3 – 참신성이 거의 없는 query

#### 12. AI agent의 특성을 결정 짓는 agent code

#### 13. agent 앱의 query처리 우선 순위

#### 14. 까다로운 웹검색 예

#### 15. RAG의 실질적 query적용 예

#### 16. web search와 연동되는 drill down예

17. 반사속고 모드(reflection mode)
18. API key적용 방법
19. 무료 test application
20. 끝맺음

## 0. 이 문서 download 방법

<https://www.ko-agent.com:8080> 을 여시면 다음 페이지가 보입니다.

The screenshot shows the AI Agent website interface. At the top, there is a logo for 'AI Agent' with the tagline '똑똑한 개인비서' (Smart Personal Assistant). Below this, there are three bullet points: 'AI agent는 똑똑한 개인비서입니다.' (AI agent is a smart personal assistant.), '개인비서 지도, 훈련은 개발자가.' (Personal assistant guide, training is for developers.), and '여러분은 개인비서를 잘 활용하시면 됩니다.' (You can use the personal assistant well.).

The main section is titled '문서 및 File Download' (Documents and File Download). It is divided into three numbered sections:

- 1) Biz User를 위한 AI agent**: Contains a link for 'Biz User를 위한 AI agent.pdf' (PDF).
- 2) 무료 Test를 위한 실제사례 - 성경(가톨릭) AI Agent App**: Contains links for '성경(가톨릭) AI agent App.pdf' (PDF), 'bible.zip' (ZIP), and 'bible.7z' (7Z).
- 3) 참고문서**: Contains a link for '구글 API key.pdf' (PDF).

At the bottom, there is a section for '문의하실 곳' (Where to ask questions), providing the website 'ko-agent.com' and the email address 'koagentcom@gmail.com'.

(이 Home Page의 모양과 내용은 향후 자주 변동될 것입니다.)

여기서 위에 보이는 Biz User를 위한 AI agent.pdf button을 click하시면 됩니다.

## 1. AI agent란 무엇인가?

AI agent는 똑똑한 내 비서로 여기시면 됩니다.

이 문서에서는 가능하면 기술적 용어나 복잡한 개념 설명은 피하고 주요특성 파악을 위한 경우를 제외하고는 Biz user가 AI agent를 쉽게 활용하는 것에 초점을 맞추어 설명하겠습니다. 근래에 AI의 일상생활, 직장에서의 업무에 대한 AI의 침투에 따른 변화 등으로 많은 분들이 우려와 걱정을 하실 수 있다고 봅니다. 그러나 AI의 여러 분야에 대한 영향은 현재 거스를 수 없는 대세이나 AI를 쉽게 충분히 이해하고 활용하면 미래를 겁낼 필요가 없다고 판단합니다. 또한 앞으로 수십년 뒤의 세상에 대한 예측은 본인의 능력범위를 벗어나는 것이므로 이는 현재로서는 논쟁을 피하길 바라고, 본인이 겪은 경험에 비추어, 아직도 AI분야는 더 적극적으로 발전이 필요하고, 더 많은 분들이 AI 발전에 따른 좋은 영향을 받게 되기를 희망합니다. 솔직히 말하자면, 주변에 AI의 발전을 걱정스럽게 여기시고, 걱정부터 하시는 분도 계시지만, 제 소신은 인간의 지혜는 아직까지는 AI로 대체될 수 없다고 판단합니다. 다만, AI의 활용으로 인간의 지혜성장/발전에 긍정적인 효과가 클 것으로 판단합니다.

이러한 배경하에 이 문서는 과학자/기술자가 아닌 일반 사용자, biz user가 쉽게 AI agent를 이해하고 충분히 특성을 파악하고 경험하는데 도움을 드리고자 합니다. 이에 따라서 제가 실제로 적용한 사례를 보여드리고 (youtube 주소를 안내해 드리겠습니다.), 또한 지적재산권

침해가 발생하지 않는 공개된 문서를 사용하는 특정한 실제 사례는 직접 test해 보실 수 있도록 Windows에서 실행시킬 수 있는 .exe file 을 한정된 기간에는 구동시키실 수 있도록 배포할 예정이니 직접 DOWNLOAD 받으셔서 압축을 풀고 실제 test를 해 보시면 더욱 잘 이해하시는 데 도움이 될 것으로 보입니다. 그런데 이때는 API(application programming interface) key를 직접 받으실 필요가 있습니다. 이 API key는 제 판단으로는 대부분의 사용자가 무료로 충분히 test하실 수 있을 것으로 보입니다. 다만, 제가 youtube통해서 보여 드릴 때에는 이러한 실제 사례보다는 여러 기능이 같이 있는 application도 보여 드릴 예정인데, 이 이유는 이러한 여러 기능은 기술지원이 별도로 필요한 부분이라서 현재로서는 무료로 제공하지 못함을 뵈으신 아량으로 양해해주시기 바랍니다.

그러면 AI agent는 무엇일까요? 쉬운 예로 여러분이 이미 활용하고 계신 ChatGPT/Gemini/Claude/Grok등 chat bot은 AI agent의 매우 잘 알려진 예로 보실 수 있습니다. 그런데 이러한 것은 OpenAI, Google, Anthropic, xAI등의 대형기업이 수많은 사용자를 대상으로 그 기업의 여러 목적에 맞추어 여러 사용자들에게 유료/무료의 서비스를 제공하는 것으로 일반 기업, 개인 사용자들의 여러 요구를 일일이 맞추기 보다는 수많은 사용자들의 요구사항을 나름 예상하여 개발하고 서비스를 제공하는 것입니다. 이 문서에서 설명 드리고자 하는 것은 이와는 다르게 기업, 개인이 보유하는 문서에 맞추어 분석하고, 분석된 내용을 토대로 사용자의 질의에 맞춘 답변을 제공하거나, 미리 사전에 준비된 여러 도구를 사용하여 질의에 답변하거나 하게 됩니다. 여기서 도구는 매우 다양한 도구가 될 수 있으며, 이것은 기업, 개인의 요구

에 따라 다르게 됩니다. 그리고, 이러한 도구중에서 특별히 웹검색 도구도 같이 사용할 수 있는 바, 이 웹검색 도구는 분석된 내용에 대한 답변이나, 추가적 도구를 사용한 답변이 부족하다고 판단할 때에 추가적인 도구로 사용할 수 있게 되는 것입니다. 여기서 AI agent에는 LLM(Large Language Model: 거대언어모델)이 중심적인 역할을 하고, 이에 따라서 이러한 AI agent사용에는 필수적으로 OpenAI나 Google, Anthropic등의 여러 회사에서 제공하는 API key를 얻어서 연결시켜야 합니다.

## 2. LLM(Large Language Model)

앞서 LLM이 AI agent의 중심적 역할을 한다고 말씀 드렸는데, 그것은 현재 LLM이 상당히 발달이 된 상태이고, LLM의 여러 우수한 능력을 잘 활용하는데 우리의 목적이 있습니다. 아마도 이미 잘 아실 것이라 믿는데, LLM의 개발에는 천문학적인 비용이 듭니다. 또한 여러 회사에서 경쟁적으로 성능 좋은 model을 엄청난 비용을 들여서 만들고 있는 상태이니 우리는 우리 형편에 맞게 좋은 model을 골라서 사용하면 됩니다. 물론 여러 다양한 model이 있고, 성능, 속도, 사용비용이 제각각 이라서, 제가 이 문서를 작성 완료할 시점까지 최강의 model을 잘 선택해서 보여 드릴 것을 장담하기는 어렵습니다만, 제 경험에 비추어 현실적으로 보여 드리는 model들이 어느 정도 기간내에는 꽤 실용적인 대안이 될 것으로 믿습니다. 이러한 model의 사용에는 API의 사용이 따르게 되고, 이러한 기술적인 문제는 저 같은 개발자가 해결해야 할 과제이며, 여러 biz user는 이러한 API를 사용한 LLM model을 AI agent로 활용하여 똑똑한 비서로 여기시고 적극적으로 활용하시면 되는 겁니다.

다만, 미리 말씀드릴 것은 제가 보여 드리는 것은 여러 실용적인 model들에 대한 대안을 보여 드리는 것이고 실제의 사용은 여러분이 선택해서 사용하시게 됩니다. 그런데 어떠한 model이 가장 좋으냐는 어떠한 법칙으로 정해져서 결정되는 것이 아니고, 매우 다양한 조건에 따라서 여러 가지 model을 선택해서 사용 해보면서 경험적으로 판단해야 할 경우가 상당히 많습니다. 저는 이 문서가 그러한 경험을 효과적으로 하시도록 도움이 되기를 바라는 마음으로 작성하겠습니다.

일반적으로는 LLM model을 훈련시키는 data의 양과 내부 변수의 양으로 표현합니다. 다만 이러한 정량적인 표현이 솔직히 말씀 드리자면 저같이 주어진 model을 사용하는 입장에서는 큰 의미가 없고, 질의에 대해서 얼마나 정확히 잘 답변하느냐, 또 얼마나 빨리 답변하느냐, 비용이 얼마나 드느냐가 더 큰 의미를 갖습니다. 이러한 관점으로 말씀 드리자면, 제 경험으로는 대체로 더 복잡하고 큰 model이 좀 더 정확히 답변하고, 다만 좀 더 느리게 답변하고 비용은 상당히 수십 배 이상으로 올라갈 가능성이 높게 됩니다. 그런데, 많은 경우 질의에 대한 답변이 작은 model로도 빠르고 충실하게 얻을 수 있으므로 어떠한 model을 사용할지에 대해서는, 다시 한번 말씀 드리지만, 경험적으로 판단해서 정하는 것이 필요합니다. 질의와 현재 상황과 결합된 정보를 알아야 더 나은 model을 선택할 수 있는데, 이 것은 너무나 많은 경우가 있게 되므로 경험적인 판단이 필요한 것입니다.

그리고 덧붙여서 말씀 드리자면, LLM은 계속 새로운 model이 나오는 상황이므로 제가 보여드리는 예에서도 시기적으로 계속 변동되어 실제 사용하시는 시점에는 조금 다르게 보일 가능성이 있습니다. 다만, 현재까지의 경험으로는 LLM model을 제공하는 회사에 따라서 성능이나 특성이 크게 다르기 보다는 같은 회사에서 제공하는 model의 종류에 따라서 성능이나 특성이 다르게 됩니다.

또 하나 중요하게 말씀드릴 것은 사진이나 그림 같은 image나 소리, 동영상 정보를 이용하거나 분석 또는 생성하는 것은 본 문서의 범위

를 벗어나고 순수하게 문자 정보(text 정보)를 분석하거나 생성하는 것으로 제한합니다. LLM의 발전이 근래에는 이러한 것들까지 포함시켜서 Multimodal LLM(MLLM), 또는 LMM(Large Multimodal Model)으로 불리는 것들도 이미 여러가지 나와 있지만, LLM agent의 핵심은 아직도 text를 다루는 LLM이 대세로 판단되며 이에 대한 충분한 이해 및 경험 습득은 향후의 MLLM이용에도 도움이 될 것으로 믿습니다. 이에 따라서 이 문서에서는 LLM을 이용한 AI gent로 국한 하겠습니다.

### 3. LLM만을 이용한 질의 처리

LLM자체가 상당한 data를 기반으로 만들어진 것이므로 웬만한 질의에 충분히 잘 답변할 수 있습니다. 그러나, 이러한 사용은 ChatGPT, Gemini, Claude, Grok등의 chat bot사용에 비해 차별화 점을 찾기가 어렵고, 오히려 여러가지로 기능 보완을 해서 API key값을 이용해서 구현을 하더라도 일반 chat bot의 성능을 뛰어 넘기가 매우 어렵습니다.

그러므로 이 문서에서는 LLM만을 사용하는 경우는 LLM과 tool을 같이 사용하여 질의에 대한 답변을 구할 때에 굳이 tool을 사용하지 않고 LLM만으로 답변을 할 수 있는 경우에 한하여 LLM만을 사용하게 될 것입니다. 즉, 질의 시에 이미 LLM만이 아닌 tool을 같이 사용하여 답변을 얻을 것을 예상하는 경우에 한하여 LLM만을 사용하게 될 겁니다. 이러한 예는 youtube에 올리는 동영상을 통하여 더 쉽게 이해하시게 될 겁니다.

#### 4. Generation과 Understanding

아마도 생성형 AI(generative AI)라는 용어는 이미 많은 분들이 들어 보셨을 겁니다. 이 용어에 대해서 많은 분들이 이미 잘 알고 계실 수도 있지만, 이것은 LLM이 인간이 문장으로 잘 알아볼 수 있고, 그림이나 사진으로 볼 수 있고, 음성으로 들을 수 있게 문장을 생성하거나 image를 생성하거나, 음성을 생성하거나, 동영상을 생성해서 보여주는 AI를 말합니다. 그리고 이렇게 LLM이 생성하는 것을 generation이라고 기술적 용어를 사용하는데 너무 깊게 생각하실 필요 없이 그렇구나 라고 여기시면 됩니다. 이 문서에서 설명드리고자 하는 AI agent도 생성형 AI를 사용한다고 볼 수 있지요.

그러면 LLM이 이러한 능력을 갖추기 위해서 먼저 무엇이 필요할까요? 여러분이 이미 학창시절에 많은 학습과 훈련에 의해서 경험하신 것이지만, 우리가 다른 사람에게 보일 만한 글을 쓰기 위해서는 먼저 많은 학습을 하고 글을 쓰기 위한 내용을 내 자신이 잘 알고 있어야만 좋은 글을 쓸 수 있습니다. LLM도 마찬가지로 이러한 것들을 하기 위하여 많은 data로 훈련을 받아야 합니다. 앞서서도 말씀 드렸지만, 고성능 LLM model을 만들기 위한 훈련비용에는 천문학적 비용이 들어갑니다. 우리는 이렇게 만들어진 LLM model을 잘 골라서 사용하면 되는 겁니다.

그런데, LLM이 이렇게 훈련을 받고 나서 여러분의 질의에 대해서 답변을 하려면 질의의 내용이 무엇인가를 파악을 해야겠지요. 이렇게 인간이 LLM에게 입력한 것을 LLM이 파악하려면 generation의 반대 의

미인 understanding이 필요한 것입니다. LLM이 understanding단계를 거친 후에 내부적인 여러 과정을 거쳐서 여러분께 질의에 대한 답변을 드리게 되는 겁니다. 그런데, 실제의 AI agent사용에는 generation, understanding 이라는 용어는 굳이 아실 필요 없이 편하게 AI agent가 '똑똑한 내 비서다'라고 여기시고 활용하시면 됩니다.

## 5. RAG(Retrieval Augmented Generation)

LLM이 상당한 능력을 갖추어서 수많은 질의에 질 높은 답변을 할 수 있는 것은 사실입니다만, LLM만으로는 답변이 어려운 질의도 많이 존재합니다. 물론 AI agent가 아무리 뛰어나도 모든 질의에 제대로 답변할 수는 없고, 제가 보는 관점에서는 앞으로 수십 년이 지나도 세상의 모든 질문에 완벽한 답변을 할 AI이건, agent이건, computer이건 세상에 존재하지 못할 것이라 믿습니다.

다만, 모든 질문이 아닌, 우리가 도움을 청해서 agent가 우리에게 유용한 답변을 할 것을 기대하는 질의에 대해, 이 문서가 어떠한 종류나 형식의 질의가 도움이 될 지를 더 잘 알게 되고, 이에 따라서 LLM이건 agent건 우리에게 도움이 되는 답변을 얻게 되는데 유용하게 되기를 기대합니다.

그런데, 우선 LLM만으로는 답변이 제대로 안 될 수 있는 이유에 대해서 먼저 아는 것이 도움이 될 것으로 여겨 제 나름대로 파악한 이유를 들어 설명 드리도록 하지요.

가) 우선 LLM이 훈련에 사용하는 것은 상당한 양의 문서나 data이기는 하지만, 한계가 있고, 또한 새로운 문서나 data가 자꾸 생겨날 수 밖에 없기 때문입니다.

나) LLM을 훈련시킬 때에는 아직도 수개월 이상의 훈련이 필요하므로 최신의 data와는 시간상 격차가 생길 수 밖에 없습니다.

다) 기업이나 정부기관이나, 병원이나, 기타 여러 단체, 개인 등이 보유한 data를 보안상, 또는 기술적, 비용적 한계로 외부로 노출시키지 못 하는 data의 양도 엄청납니다. 이러한 data는 LLM훈련에 사용될 수 없고, 이러한 data를 사용하려면 특별한 방법으로 대처 해야 합니다.

여기서 다)의 문제 해결을 위하여는 여러 방법이 존재하는데, 큰 기업이나 정부 기관에서는 미리 훈련된 model에 추가적으로 필요한 data를 추가하여 재 훈련시키는 방법을 사용할 수도 있습니다. 그러나 이러한 방식은 꽤 많은 비용을 감수해야 합니다.

이러한 방식 이외에 다른 방식으로 RAG(Retrieval Augmented Generation)를 적용하면 재 훈련시키는 비용보다 훨씬 저렴한 비용으로 보유 문서, data를 이용하여 질의에 대해서 답변을 얻을 수가 있습니다. RAG는 아직 우리말로 보편적으로 사용하는 용어가 없어서 저는 앞으로 RAG로 부르면서 사용하겠습니다.

## 6. RAG와 embedding model

RAG를 사용하기 위해서는 필수적으로 보유 문서를 분석을 먼저 해야 합니다. 이러한 문서 분석에 반드시 사용해야 하는 전제 도구로 embedding model(또는 간단히 embed model로도 부름.)이 있습니다. 기술적인 설명은 최대한 자제하려고 하지만, 주요한 특성이므로 아시는 것이 도움이 되리라 생각하고 간단히 설명 드리겠습니다.

우선, 일반적으로 가장 많이 사용하는 것으로 OpenAI의 small, large 2가지 embedding model을 선택하실 수 있습니다. embedding model의 특성을 나타내는 요소로 dimension이 있는데 small은 dimension이 1536이고 large는 이의 2배인 3072입니다. 이것은 OpenAI의 API key가 있어야 사용이 가능하고, 무엇보다, 제가 보기에 기업 입장에서는 분석하고자 하는 문서를 OpenAI 서버로 보내야 하므로 기업 비밀 보호상 어려움이 발생할 수가 있습니다.

제가 youtube video로 보여 드리고 설명하는 내용 중에는 OpenAI 서버로 보낼 필요 없이 자체 보유 서버를 이용하여 사용할 수 있는 fast, better, NOA-large도 같이 있습니다. fast는 매우 반응이 빠르는데, 어떠한 질의에 대응하기 위해서는 분석된 내용이 불충분 할 수 있어서 이보다는 좀 더 충실히 분석할 수 있는 better가 있고, 많은 test를 해보니, 일부 질의에는 OpenAI large나 NOA-large를 사용해야 제대로 답변을 얻는 경우도 있었습니다. 그리고 제가 이 문서를 작성할 초기에는 고려를 못하다가 나중에 사용하는 embed model로 gemini-embed model이 있습니다. 꽤 많은 test결과를 통해서 확인한 바로는 OpenAI

large로 제대로 된 답변을 얻지 못하는 경우에 NOA-large나 gemini-embed가 해결책이 되는 경우가 제법 있었습니다. 다만, gemini-embed도 google server를 이용하므로 기업 비밀 보호상 어려움이 발생할 수가 있습니다.

그런데, 무슨 LLM이 어떠한 경우에 적합한 지를 판단하는 것도 많은 경험으로 판단해야 한다고 말씀 드렸는데, embedding model선택도 마찬가지입니다. 더욱이, embedding model과 LLM model pair에 대한 선택도 좀 더 복잡한 문제가 되는데, 이것도 많은 경험에 의해서 판단되어야 한다고 저는 생각하며, 다만, 이 문서와 제 youtube video, 공개되는 test용 실제 사례 application등이 이러한 판단에 도움이 되기를 간절히 바라고 있습니다.

## 7. AI agent 관련 문서 및 사용 web 주소

AI agent관련 문서는 다음 2가지가 있습니다.

- 1) Biz User를 위한 AI agent.pdf
- 2) 무료 test를 위한 실제사례

지금 보여 드리는 문서는 1)이고 이것이 AI agent관련하여 가장 중요한 main 문서입니다. 2)는 여러분이 직접 test하실 수 있도록 비교적 쉽고 상세하게 설명 드리는 사용방법입니다. 이것은 실제 사례에 따라 여러가지 종류의 문서 제목으로 보이게 될 겁니다. 여기서는 download 받으시는 압축파일을 푸시고 어떻게 사용하실지 설명드리게 될 겁니다. 굳이 여러분이 익숙하지 않으시는 program, software등을 설치하실 필요는 없으나 Gemini, tavily, brave search API key를 얻기 위해서는 Gemini나 ChatGPT 안내를 받으시면 쉽게 얻으실 수 있고, 무료로 신용카드 정보 입력 없이도 받으실 수 있습니다. 그런데 API key를 받으시는 것은 Gemini, ChatGPT를 월간/연간 유료로 구독하시는 것과는 다른 것임에 유의하시기 바랍니다.

무료 test 를 위한 실제사례 사용방법에 대해서는 여러 개념을 충분히 이해하시고 사용하시면 더 크게 도움이 될 것으로 판단하여 이 문서의 거의 끝부분에 다시 설명을 드리도록 하겠습니다.

## 8. RAG적용 사례

RAG를 적용한 사례는 여러 가지가 있는데 여기서는 중요한 예를 youtube site를 이용하여 설명 드리겠습니다. RAG로 출발하여 웹검색 까지 이용하게 되는 심화질의로 계속 이어지는 예는 향후에 mcp tool 사용에 이어 web search tool까지 설명 드린 후에 더 보여 드리겠습니다.

### a. 간단한 문서의 RAG 적용

이것은 수료증.pdf라는 1 페이지 짜리 문서를 서버에 올리고 문서 분석 후에 '이 문서의 목적은?'라는 단순한 질의에 간단히 답하는 내용입니다. 문서 내용과 질의 내용이 모두 단순하므로 embed model은 무엇이래도 질의에 대한 답변을 잘 하게 됩니다. youtube에서는 가장 가벼운 embed model을 적용했습니다.

여기서는 AI agent의 실제 예를 보여드리게 되는데 이것은 앞으로 설명드릴 test application이나 동영상에서나 다음 주소를 web browser로 여시면 보시게 됩니다.

<http://localhost:8000>

여기서 보여드리는 AI agent의 사용 예와 test application의 경우 화면에 보이는 모습이 크게 다를 수도 있습니다. 이것은 AI agent를 우리

의 목적에 맞추어 모습을 크게 다르게 바꿀 수 있음을 보여드리는 겁니다.

query 및 동영상 주소는 아래와 같습니다.

query: 이 문서의 목적은?

youtube address: <https://youtu.be/kJKyTXSAFyo>

동영상 제목: 1RAG기본

그리고 동영상에 언급하는 공개된 source에 관해서는 수정된 내용으로 말씀을 드립니다. 일반 사용자가 공개된 source를 이용하여 필요시 python등을 설치하고 test하는 방법보다는 일반 사용자는 더 test하기 쉬운 더 완성된 형태의 application으로 제공 받아서 test하시는 것이 방법상으로도 더 쉽고 준비할 사항도 더 적어서 test할 기회가 더 늘어 날수 있다는 판단과 함께 partner들을 고려하니 source공개는 함부로 임의대로 해서는 안되겠다는 뒤 늦은 깨달음도 같이 작용했음을 알려 드립니다. 이에 대해서는 깊이 이해해 주시고 앞으로 보이는 동영상에도 공개된 source라는 표현에는 이러한 점이 적용됨을 깊이 헤아려 주시기 바랍니다.

#### b. embed model에 따른 query결과

이제는 수백 페이지 이상의 좀 더 복잡한 문서를 이용해서 RAG를 test해보겠습니다. 이 문서는 기술적인 문서이므로 안의 복잡한 기술적 내용보다는 간단히 저자를 파악하는 query를 해 보겠습니다. 간단한 embed model은 답변을 제대로 못하고 성능이 좋은 embed model

을 사용해야 답변이 제대로 얻어 집니다. NOA-large 나 OA-large를 사용해야 author, 즉 저자에 대한 정확한 답변을 얻게 됩니다. 다만, OA-large는 비용이 좀 들고 시간이 더 많이 걸리며 무엇보다도 보안 문제를 해결해야 하는 문제를 안고 있습니다. NOA-large는 on-premise server의 성능에 크게 처리 시간이 좌우받게 됩니다. query 및 동영상 주소는 아래와 같습니다.

query: 이 문서의 저자는?

Youtube address: <https://youtu.be/s1dzBShuKPI>

동영상 제목: 2 embed model의 영향

### c. 영어와 한글 질의에 대한 차이

각각 수백 페이지 이상의 4개의 pdf문서를 서버에 올린 후에 better embed model로 문서분석을 한 후에 다음 질문을 하면

Q1: 문서 1의 저자는?

답변을 제대로 얻지 못합니다. 그런데 같은 내용으로 영어로 질문을 하면

Q2: who are authors of document 1?

이에 대해서는 정확한 답변을 얻게 됩니다. 이 것은 아직은 우리가 사용하는 LLM model이 영어로는 훈련이 잘 되어서 답변을 잘 할 확률이 높게 됨을 보여줍니다. 이러한 내용을 잘 알고 대처해서 가능하면 query할 때에 영어로 하는 것이 답변을 잘 얻게 될 가능성이 높아집니다. 물론 Language를 한국어로 선택하면 답변을 한국어로 얻게 되기도 하나, 이것도 경우에 따라서는 영어로 답변을 얻기도 합니다. query가 영어인 경우에는 agent의 상태에 따라서 정해지므로 항상 한

국어로 답변을 얻는 것이 아님을 기억하시기 바랍니다. query와 동영상 주소는 아래와 같습니다.

Q1: 문서 1의 저자는

Q2: who are authors of document 1?

Youtube address: <https://youtu.be/lens5EPu7qo>

동영상 제목: 3 restart + several files upload + english korean query result

참고로, NOA-large나 OA-large를 사용하시면 한국어나 영어나 관계없이 저자에 대한 질의에 잘 답변합니다.

그리고 이 동영상에서는 새로 시작(restart)단추와 여러 개의 문서를 서버에 올리고 이중 일부 또는 전체를 제외시켜서 문서분석을 하는 방법에 대해서도 설명합니다. 문서분석을 하면 이전에 한 문서분석결과(vector store)는 사라지고 새로운 결과로 대체됩니다. 이러한 점을 기억하시고 문서분석을 하시기 바랍니다. 기업이나 기관에서 대량의 문서를 분석하고 이 결과를 저장하여 차후에 사용하려면 별도의 database사용이 필요합니다. 이러한 것은 본 문서의 범위를 벗어나는 것입니다.

d. LLM 답변이 항상 맞는 것은 아님.

이번에는 다른 질의를 해 봅니다.

Q1: 문서 1의 발행일은?

이 질의에 대해서도 한국어로 질의를 하면 better embed model로 분

석한 결과를 사용하면 답변을 제대로 얻지 못 합니다. 영어로 다음과 같이 질의하면,

Q2: what is the publication date of document 1?

이에 대해서는 맞는 답변을 합니다. 문서 2 ~ 4에 대해서도 같은 질의를 하면 모두 답변을 합니다. 그런데 문서 4에 대한 답변은 오류를 보입니다. 만일 좀 더 성능이 좋은 embed model인 NOA-large를 사용하면 맞는 답변을 하게 됩니다. 이 예에서 보이듯이 LLM 답변이 항상 맞는 것은 아닙니다.

그러므로 항상 LLM이 오류를 가질 수 있음을 알고 사용해야 합니다. 그렇지만 이러한 오류가능성에도 불구하고 아예 사용하지 않는 것보다는 이에 대비하면서 사용하는 것이 우리에게 큰 도움이 되는 것입니다. agent도 부족한 점이 있을 수 있지만 이러한 점을 염두에 두고 잘 활용하면 똑똑한 내 비서로 역할을 톡톡히 하게 될 것입니다. query와 동영상 주소는 아래와 같습니다.

Q1: 문서 1의 발행일은?

Q2: what is the publication date of document 1?

Youtube address: <https://youtu.be/BxlwgyqBWn0>

동영상 제목: 4 llm answer is not always correct

#### e. RAG에서 memory mode이용

RAG에서 이전의 질문들에 대한 맥락을 이용해서 질의를 할 수 있는데 이럴 경우 문서 1의 발행일은? 이라는 질의를 앞에서 했다면 문서 2는? 이라는 질의는 당연히 문서 2의 발행일은? 이라는 질의로 여기

고 답변을 해야 할 것입니다. 그러나 이러한 답변을 하기 위해서는 반드시 Memory Mode toggle switch를 ON으로 해야 합니다. 이 것은 memory mode를 OFF로 하여 질의하는 것이 더 질의에 대한 답변을 정확히 하게 되는 경우도 많기 때문입니다. memory mode를 사용하기 위해서는 반드시 이전 질의에 대한 기억을 필요로 하고, 이것을 무한정 기억할 수는 없으니 현재는 10개의 과거 질의를 기억하는 것으로 정하여 처리하고 있습니다. 그런데 memory mode는 편한 점도 있지만 때로는 혼란을 줄 수도 있으니 주의하여 사용하시기 바랍니다.

동영상 주소는 아래와 같습니다.

Youtube address: <https://youtu.be/az1ssm8WmLo>

동영상 제목: 5 RAG에서 memory mode이용

## 9. RAG 활용시 주요 고려사항 및 추가 활용 사례

### a. 시사성이 높은 문서를 예측에 이용할 경우 주의 요망

RAG는 보유문서를 분석해서 질의에 대한 답변을 얻는 것이므로 만일 보유문서가 시사성이 높은 것이라면 질의 내용이 시기적 관련성이 높은 예측 등에 대한 사용에는 주의를 해야 합니다. 예를 들어 다음 문서를 사용해 봅니다.

2025년 8월 KDI에서 발행한 KDI 경제전망 수정본(제42권 제3호)을 이용해 보겠습니다. 이렇게 이 문서는 비교적 최근에 발행이 된 것입니다. (AI agent관련 본 문서 작성시기 2025년 10월 기준)

이에 따르면 환율은 변동이 거의 없을 것으로 전망하고 국제유가는 두바이유 기준으로 2025년 71달러, 2026년 67달러로 예상합니다.

그런데 실제 2025년 10월 현재 환율은 2025년 8월부터 지금까지는 상승세이고, 인터넷 검색결과 두바이유는 2025년 10월 21일 현재 61.82달러로 나타납니다. 이처럼 시사성이 높은 문서를 RAG를 이용하여 예측하는 것 보다는 웹검색이 더 정확한 정보를 제공할 수 있습니다. 이것은 RAG는 기존 문서에 있는 내용으로 정보를 제공하는 것이고 이에 따라서 기존 문서가 오류가 있으면 이것을 그대로 전달하기 때문이며, 향후의 예측이나 전망을 어떤 기관이 더 정확하게 할 수 있느냐에 대한 판단은 현재로서는 agent가 제공할 수 있는 것이 아니고 biz user가 독립적으로 잘 찾아야만 하는 과제입니다.

동영상주소는 아래와 같습니다.

Youtube address: <https://youtu.be/exAJB4OEako>

동영상 제목: 6 RAG이용해서 예상할 경우의 주의 사항

## b. drill down 예1 – 서울시 환경정책

원유 시추시에 점점 더 깊게 파고들어 더 유용한 원유를 얻게 되듯이 agent에서도 어떠한 주제에 대해 질의를 하고 이에 대한 답변 중에서 더 관심이 있는 것을 파고들어서 심화질의를 하는 것을 drill down으로 부를 수 있습니다.

서울시가 2023년도에 발행한 2040 도시환경기본계획이라는 pdf문서를 이용해서 drill down예를 보이는 바, 처음 질의로

Q1: 서울시의 환경 정책에 대해 자세히 알려주세요.

라고 한 것에 대해서 얻은 답변 중에

\* 풍부한 자연환경을 시민이 편리하게 이용할 수 있도록 자연자원에 대한 접근성을 강화하고 녹색 네트워크를 구축합니다.

내용을 출력창에서 mouse로 선택하고 '선택영역을 입력창으로 복사'라는 button을 이용하면 그대로 입력창에 복사가 되므로 앞에서의 출력 내용을 이용하여 심화질의를 할 때 매우 편리하게 이용할 수 있습니다. 이렇게 복사한 내용을 이용하여 더 심화질의를 합니다. 즉,

Q2: \* 풍부한 자연환경을 시민이 편리하게 이용할 수 있도록 자연자원에 대한 접근성을 강화하고 녹색 네트워크를 구축합니다.에 대해서 더 자세히 알려주세요

라고 질의를 합니다.

그런데 drill down을 이용하실 때에는 앞에서 한 query와 연속적으로 같은 맥락을 유지한 상태로 query를 하게 되므로 memory mode toggle button을 ON상태로 두시고 query를 하시는 것이 훨씬 더 효과적입니다. 왜냐하면 이것이 OFF상태이면 query처리는 앞서 처리한 query결과와는 관계없이 지금의 query와 vector store만을 이용하여 답변이 얻어지기 때문입니다.

이번에는 보강 모드를 이용해서 추가 지시사항으로

초등학생이 알 수 있도록 쉽게 알려주세요

라는 지시사항도 덧붙여서 질의를 합니다. 그러면 초등학생이 알기 쉬운 내용으로 답변을 하게 됩니다. 추가 지시사항에 입력한 것은 여러분이 수정을 하지 않으시는 한에는 계속 그대로 유지됩니다. 그러므로 위에서 사용한 지시사항을 연속적으로 적용을 원하실 때에 유용하게 이용하실 수 있습니다. 물론 보강 모드에서 다시 일반 모드로 전환하시면 이것은 적용되지 않습니다. 그러므로 일시적으로 이러한 지시사항을 적용하지 않으시려면 이렇게 일반 모드로 전환하시면 됩니다. 다시 보강 모드로 복귀하면 앞에서 입력한 추가 지시사항은 그대로 유지된 채로 전환이 됩니다.

앞에서 얻은 답변내용을 이용하여 다시 다음 심화질의를 합니다.

Q3: 그리고 '녹색 네트워크'는 서울에 있는 여러 초록색 공간들(산, 강, 공원 등)을 서로서로 연결하는 거예요. 마치 여러 개의 작은 놀이터들을 긴 미끄럼틀이나 다리로 이어서 하나의 아주 큰 놀이터를 만드는 것과 비슷하죠.에 대해서 더 자세히 알려주세요

계속적으로 앞에서 얻은 내용을 이용하여 다시 다음 심화질의의를 합니다.

Q4: 서울시는 이런 좋은 자연을 친구들이 더 쉽고 편하게 찾아가서 즐길 수 있도록 만들 거예요. 마치 여러 놀이터를 서로 연결해서 더 크고 재미있는 놀이공간을 만든 예를 보여주세요.

이에 대해서 나름 좋은 설명을 얻게 됩니다.

이렇게 drill down을 이용하여 기존 보유문서에 대한 깊이 있는 분석 및 이해에 도움이 되는 query결과를 얻을 수 있습니다.

이에 대한 동영상 주소는 아래와 같습니다.

Youtube address: <https://youtu.be/qZyyVWgxBHE>

동영상 제목: 7 drill down 1 서울시 환경정책

drill down은 웹검색과 같이 사용하면 더욱 유용한 agent로 활용할 수 있는데, 이것은 향후에 웹검색 설명을 한 후에 다시 설명 드리겠습니다.

### c. drill down 예2 – 서울시 자율주행 관련계획

앞서 사용한 같은 문서를 이용해서 이번에는 서울시의 자율주행관련 계획에 대해서 알아보고 drill down을 진행합니다. 그런데 이 질의 내용에 대해서는 embed model이 성능 좋은 것이라야만 좋은 답변을 얻을 수 있습니다. 즉, NOA-large나 OA-large를 사용해야만 query에

대한 답변을 얻을 수 있습니다. 이렇게 같은 문서라 하더라도 주제에 따라서 다른 embed model사용이 필요할 수 있습니다. 처음 query는 다음을 사용합니다.

Q1: 서울시 자율주행 관련 계획을 자세히 알려주세요.

이로부터 얻은 출력으로부터 계속적으로 다음 심화 질의들을 얻습니다.

Q2: - \* 기존의 교통수단과 자율주행자동차, UAM, 마이크로모빌리티 등 새롭게 등장하는 다양한 교통수단을 통합하여 관제하고 운영·관리할 수 있는 기반을 조성합니다.에 대하여 더 자세히 설명해 주세요

Q3: \* \*\*광역형 허브:\*\* 서울 및 수도권외의 주요 거점과 GTX, UAM 등 광역교통망을 연계하는 복합환승센터를 구축하고, 공공·상업시설 등 도시 서비스를 제공하는 스마트 물류 플랫폼을 구축합니다.에 대하여 더 자세히 설명해 주세요

Q4: \* \*\*3 차원 물류네트워크 통합:\*\* 도심 내 급증하는 물류 수요에 대비하여 '공중+지상+지하'를 모두 활용하는 '3 차원 물류네트워크'를 구축합니다. 이는 공중 활용을 위한 드론 배송 인프라 조성, 지상부의 물류 로봇 배송, 지하철을 활용한 지하 배송체계 구축 등을 포함합니다.에 대하여 더 자세히 설명해 주세요

Q5: \* \*\*스마트 물류 플랫폼 연계:\*\* 광역형 모빌리티 허브 구축 시 공공·상업시설 등 도시 서비스를 제공하는 스마트 물류 플랫폼을 구축하여 3 차원 물류네트워크와 연계합니다.에 대하여 더 자세히 설명해 주세요

Q6: \* 플랫폼은 빅데이터 분석을 통해 물류 흐름을 최적화하고, 교통량 예측, 서비스 수요 분석 등을 수행하여 효율적인 운영을 지원합니다. 이를 통해 실시간으로 변화하는 도시 환경에 유연하게 대응할 수 있습니다.에 대하여 더 자세히 설명해 주세요

Q2 ~ Q6 모두 앞의 답변을 이용하여 얻은 질의입니다. 그런데, 앞의

답변 내용을 보면서 다음 query를 얻을 때 앞에서 query한 것을 모두 잘 기억하면서 query를 만들지 않는 한에는 다시 중복된 query를 만들 가능성도 있습니다. 즉, query에 대한 답변이 이미 질의한 내용과 겹치는 내용일 수 있으므로 앞의 query가 어떤 것인지 기억하면서 현재의 query내용을 선택해야지 그렇지 않으면 또다시 유사한 query를 만들 수 있다는 말씀입니다. 이러한 현상의 원인은 LLM이 query에 대한 답변을 할 때에 중복된 내용을 자주 사용하면서 내용이 풍부하게 보이도록 하는 현실로부터 기인한 것으로 추정이 됩니다. 그런데, drill down을 하기 위해서는 query에 대한 답변이 충실해야지 너무 간단하면 다음 심화질의를 만들기 매우 어렵게 됩니다. 이것은 다음에 다시 설명드리겠습니다.

이렇게 다음 query를 만들면서 agent가 만드는 drill down에 의한 query에 대한 답변을 보면서 web search를 사용하지 않고 RAG만에 의해서 drill down을 할 경우에 결국은 기존 문서의 내용을 이용해서 LLM이 답변을 하게 되므로 drill down 횟수가 많이 질수록 답변이 앞에서 보인 내용과 중복된 내용이 많아질 수 있습니다. 이것은 RAG나 drill down의 효과와 한계를 보이는 좋은 예가 되었습니다. RAG는 기존 문서의 넓이와 깊이를 벗어 날수가 없습니다. web search와 연동이 안 되는 RAG만을 이용하는 drill down도 이러한 RAG의 특성을 그대로 갖습니다. 다만 drill down은 분석된 문서에 대해서는 확실히 심도 있게 이해하는 데에는 큰 도움이 됩니다. 그런데, drill down이 효과를 보이려면 분석에 사용되는 문서가 drill down에서 사용되는 query에 대하여 내용이 충실해야 함은 물론입니다.

RAG만에 의한 drill down을 극복하는 방법은 web search를 이용하는 것입니다. 물론 주제에 따라서 web search로도 답을 얻지 못할 경우도 있습니다만, 기존 문서만을 이용하는 RAG query보다는 우리가 알고자 하는 주제에 대하여 답을 얻을 가능성이 훨씬 높아지게 됩니다. web search와의 연동은 web search tool에 대하여 설명 드린 후에 다시 보여 드리겠습니다. 동영상 주소는 아래와 같습니다.

Youtube address: <https://youtu.be/5K3EUwrqngk>

동영상 제목: 8 drill down 2 - 서울시 자율주행 관련계획

#### d. drill down시에 고려할 사항

drill down에서는 가능하면 많은 내용의 출력 내용을 이용하여 이 중에서 관심있는 내용으로 좁혀서 심화질의 하는 것이 효과적인 접근 방법으로 판단 됩니다. 지금 RAG에서는 가능하면 서버에 있는 문서 분석 정보를 이용하여 query에 대한 답변을 충실하게 하도록 되어 있으므로 별도의 조치없이 query를 하시면 가능하면 많은 내용의 출력이 얻어집니다. 이렇게 얻은 출력내용을 아래와 같이 보입니다.

#### 일반 출력:

서울시의 환경 정책은 자원 순환, 온실가스 저감, 과학적 관리체계 구축 등 다양한 측면에서 추진되고 있습니다.

먼저, 재활용과 자원 순환을 극대화하기 위해 재활용 최대화, 재사용, 새활용(업사이클링)을 위한 신기술 개발을 지원하고 있습니다. 이를 통해 재활용 산업의 성장과 시장 안정화를 위한 기반을 마련하고 있습니다. 또한, 사업장과 생활폐기물의 원천 감량과 온실가스 저감을 통합적으로 관리하는 과학적 관리체계를 구축하고 있습니다.

이와 더불어, 저층 주거지역 내 노후 환경 개선을 위한 계획적 관리도 추진하고 있습니다. 양호한 저층주거지는 지역 특성을 유지하면서 노후 주거환경이 개선되도록 이면도로 정비, 주차장 및 공원·녹지 확충 등을 추진합니다. 또한, 무허가 건축물 밀집지역 등에는 주거환경 개선과 저지대 수해, 산사태 등 자연재해, 각종 재난, 범죄에 대한 선제적 대응 방안도 마련하고 있습니다.

서울시는 공공기여를 활용한 기반시설 설치를 통해 지역격차 해소와 공공성 높은 시설 조성에도 힘쓰고 있습니다. 공공임대주택 등 공적자원 확보, 시민 접근성 및 효율성 강화, 재정 기반 마련 등이 이에 포함됩니다.

또한, 보행환경 개선과 도로관리, 방범·치안 사각지대 해소를 위한 물리적 환경정비와 비물리적 관리체계 구축 등도 환경 정책의 일환으로 추진되고 있습니다. 이를 통해 어린이, 여성, 1인 가구, 저소득층 등 안전취약계층이 안심하고 생활할 수 있는 범죄예방형 도시환경을 조성하고 있습니다.

이처럼 서울시는 자원순환, 온실가스 저감, 주거환경 개선, 공공기여를 통한 기반시설 확충, 안전한 도시환경 조성 등 다양한 환경 정책을 종합적으로 추진하고 있습니다.

그리고, 추천 드리는 하나의 방법은 만일 출력결과가 처음에 약간 부족한 것으로 판단이 되면 다시 동일한 query를 하면 좀더 풍부한 결과를 얻을 수도 있습니다. 이것은 LLM의 동작 특성상 자주 보이는 양상이므로 참고하시기 바랍니다. 그런데, 출력 내용을 너무 간단하게 하면 여기서 유용한 정보를 찾기는 매우 어렵게 됩니다. 즉, 앞에서 접근하는 방법과 매우 대조적으로 출력내용을 간단하게 얻기 위한 query를 하면 이것은 좋은 심화질의 찾기가 매우 어려워집니다. 보강모드에 있는 추가 지시사항에 다음처럼 입력을 하면 출력창에 보이는 내용이 매우 간단해 집니다.

추가 지시사항: 핵심내용위주로 간단하게 답하세요.

앞에서 보이는 일반 출력시에 사용한 동일한 embed model, LLM model, query를 사용하고 추가 지시사항을 적용한 결과를 아래에 보입니다.

### 간단한 출력

서울시 환경 정책은 재활용 최대화와 재사용, 새활용을 위한 신기술 개발 지원에 중점을 두고 있습니다. 또한 재활용 산업의 성장과 시장 안정화를 위한 기반을 마련하며, 사업장과 생활폐기물의 원천 감량과 온실가스 저감을 통합 관리하는 과학적 관리체계를 구축하는 것을 목표로 합니다. 이를 통해 지속 가능한 자원순환과 환경 보호를 강화하고 있습니다.

여기서 좋은 심화질을 하기는 매우 어려워 집니다. 그러므로, 이러한 사용은 drill down에서 반드시 피해야 할 방법으로 판단이 됩니다.

이에 관한 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/VFZN6eeBTrs>

동영상 제목: 9 drill down 3 drill down시 고려사항

e. drill down 예3 – 이재명정부 123대 국정과제 – 기초연구 생태계 조성

이재명정부 123대 국정과제는 2025년 9월에 발행된 문서로 최근의 문서이며(2025년 10월 현재 기준) 앞서 보여드린 전망, 예측에 대한 문서보다는 향후 짧은 기간내에 내용이 크게 바뀌지 않을 것으로 예상되는 문서라서 AI agent로 분석하기에 좋은 예로 볼 수 있습니다. 약 200페이지 정도 되는데 여기서도 과학기술 분야인 기초연구 생태계 조성과 과학기술 인재강국 실현이라는 주제에 대해서 drill down을 진행합니다.

여기서는 수 차례에 걸친 drill down을 통하여 점점 더 깊은 관심분야로 agent가 대답하는 것을 볼 수 있습니다. 다만, 마지막 추진 일정에 대한 대답은 이 문서의 내용이 과학기술정보통신부기 작성한 내용으로 추정되는 바, 이전부터 진행하던 과제의 내용으로 답변합니다. RAG는 기존문서 분석 내용으로 답변하는 것이므로 처음 기대와는 다르지만 내용상의 오류는 없습니다.

Youtube address: [https://youtu.be/T\\_UkvzS1AE8](https://youtu.be/T_UkvzS1AE8)

동영상 제목: 10 drill down 4 이재명정부 123대 국정과제 기초연구 생태계조성

이 예에서도 drill down이 보유문서에 대한 심도 있는 분석과 이해에 도움이 될 수 있음을 확인하게 됩니다.

#### e. RAG와 similarity score

이 문서의 6 RAG와 embedding model에서 embed model을 설명했는데 문서분석은 서버에 올려진 문서를 embed model을 이용하여 vector store라는 것을 만드는 것입니다. 그러면 이 vector store와 질의 내용이 얼마나 관련성이 있는지 이를 계산해서 relevance score라는 값을 얻고 이 중에서 높은, 즉 값이 큰 것부터 나열해서 스코어 한계값(threshold)이상의 것들만 출력창에 보이게 됩니다. 이 score는 similarity score라고도 불립니다. 이 값은 0 부터 1 사이의 값이고 이에 따라서 한계값을 정하는 slide도 0 부터 1사이의 값을 정하게 됩니다. d. 에서 사용했던 문서와 질의를 사용해서 스코어 값을 보이는 동영상은 다음을 참조하세요. 여기서 스코어 값을 보려면 스코어 보이기 toggle switch를 ON으로 해야 합니다. 그리고 출력창에 질의에 대한 답변을 제대로 보기 위해서는 반드시 한계값을 넘는 스코어가 최소 1개 이상있어야 합니다.

Youtube address: <https://youtu.be/JIVBTLlq9W4>

동영상 제목: 11 RAG와 similarity score

## 10. AI agent에서 mcp tool의 이용

이제까지는 RAG만을 이용하는 예에 대해서 보였는데 사실상 agent가 더 강력한 결과를 보이고 특히 drill down에서 더 유용한 결과를 얻으려면 웹검색도 같이 연동되면 훨씬 강력하게 됩니다. 웹검색도 하나의 agent도구로 볼 수 있는데 웹검색 이전에 우선 agent의 강력한 도구인 mcp tool들을 살펴보겠습니다. 다만, mcp tool들을 여기서 설명하는 것은 현재까지 준비한 tool들만 이용하는 것으로 각 기업이나 기관에서는 여러가지 다양한 tool들을 이용할 수 있습니다. 그런데 이러한 다양한 tool을 지원하기 위해서는 별도의 기술지원이 필요하므로 이를 원하시면 별도로 연락하시기 바랍니다. 여기서는 이러한 tool들에 대한 기본적인 예만 보여 드리는 겁니다. 그리고 공개되는 test application에서는 mcp tool을 사용하는 사례가 없습니다. mcp tool 사용에는 tool 종류에 따라 사전에 준비되어야 할 사항이 꽤 필요할 수 있어서 그러한 환경을 제공해 드릴 수 없기 때문입니다. 이점 넓은 양해 부탁드립니다. 동영상으로는 mcp tool 사용의 가능성을 엿 보시는 것을 목표로 하시기 바랍니다. 다만, mcp tool이 아닌 웹검색 tool은 tavily나 brave API key만 얻으시면 이용하실 수 있을 겁니다.

사실 기업이나 기관내부의 자료는 database등에 상당히 저장되어 있고 이들을 살펴보거나 알아보려면 특정한 application에 접근해서 미리 정해진 방법으로나 가능하던 것이 종래의 일반적인 예인데, 이제 LLM model을 이용한 AI agent로는 미리 준비가 되어야 하지는 하지만 자연어 처리로 이러한 특정한 application사용방법을 대체하는 것은 매우 멋진 일이 된 것이지요. AI agent로 모두 대체하는 것은 아니

고 AI agent와 기존 application이 서로 장단점이 있으므로 서로 협동하여 사용자에게 편리함을 제공하는 것이 매우 유익한 일이 될 것입니다.

#### a. mcp tool 사용시 고려사항

mcp tool을 사용할 때에 너무 많은 수의 tool사용을 시도하면 LLM이 혼란스러워 하여 제대로 맞는 tool을 선택하지 못하게 될 수 있습니다. 이러한 것을 고려하면 분야별, 주제별로 agent당 담당분야, 담당주제를 연동하여 사용하는 것이 바람직합니다. 이 것은 필요로 하는 mcp tool을 미리 구현하고 여러개의 mcp tool이 준비되면 이 때 분류해서 사용하는 것이 바람직 할 것입니다. 소수의 mcp tool만 사용하면 llm이 tool을 제대로 선택할 가능성을 높이게 될 것입니다. 다만, 소수라도 매우 비슷한 성격, 특성의 tool이라면 혼동될 가능성도 있습니다.

또한, mcp tool을 사용하실 때에는 LLM model이 성능이 좋은 것이 항상 좋은 결과를 보이지는 않습니다. 이것은 query에 따라서 잘 맞는 LLM model이 달라진다고 보시는 것이 좋습니다. 우리말로써 쉽게 말씀드리어서 query에 따라서 궁합이 잘 맞는 LLM model이 있다고 이해하시면 됩니다. 이것은 앞으로의 예에서 잘 보여 드리게 될 것입니다.

mcp tool을 사용할 때에는 굳이 embed model을 사용할 필요없이 llm model만 선택하면 됩니다. 그런데 llm model을 2025년 10월말 현재

는 10월초에 비하여 다른 model을 사용하게 됩니다. OpenAI model은 다음을 사용합니다.

gpt-5

gpt-5-mini

gpt-5-nano

gpt-4.1

gpt-4.1-mini

gpt-4.1-nano

gpt-4o

gpt-4o-mini

Google model은 다음을 사용합니다.

gemini-2.5-pro

gemini-2.5-flash

gemini-2.5-flash-lite

gemini-2.0-flash

gemini-2.0-flash-lite

(2026년 1월 25일부터는 gemini-3-pro-preview와 gemini-3-flash-preview가 추가로 사용되고 있으나 기존에 만든 동영상에는 보이 않습니다.)

openai model은 gpt-5가 가장 성능이 좋고 price도 가장 비싸게 됩니다. 물론 대체로 성능이 좋을수록 처리속도는 느리게 됩니다. 그리고, mini, nano가 붙은 것은 더 저렴하고 더 가볍게 처리하는 model로 간주하시면 됩니다. 처리속도는 대체로 더 빨라집니다. google model은

gemini-2.5가 더 최신 model이고 성능도 더 뛰어나게 됩니다. 특히 gemini-2.5-pro는 가격도 비싸고 성능도 좋다고 평합니다. 다만 처리 속도는 매우 느리게 됩니다. 그리고, flash, flash-lite가 붙으면 pro보다는 가격이 더 저렴하고 가벼운 model을 의미하며, 대신에 처리속도는 빨라지게 됩니다.

b. mcp tool 사용 예 - ppt file내용 검색, query의 표현이 살짝 다름

여기서 보여드리는 예는 AI agent앱 뿐 아니라 보조앱도 사용하여 더 잘 이해하시게 될 겁니다. 그리고 다양한 LLM model에 따라서 query에 대한 결과가 여러가지로 바뀌게 될 것입니다. 또한, 이제부터는 같은 query라 하더라도 출력창에 보이는 agent의 답변에 LLM model도 같이 보이도록 하여 LLM model에 따른 다른 답변을 명확히 알 수 있게 변경하였습니다. 이와 더불어, LLM model이 달라지는 것 뿐 아니라 더 답변을 잘 얻기 위해서는 query의 내용도 다르게 하는 것이 필요함도 여기서의 예를 통하여 아시게 될 겁니다.

그런데, 어떠한 query가 더 좋은 query인지는 매우 중요하면서도 쉽지 않은 내용인 바, 원칙적으로는 query의 표현이 구체적이고 LLM이 이해하기에 명확한 표현이 바람직하다는 원칙은 말씀드릴 수 있는데, 이외의 더 상세한 query의 표현 방법은 여러가지로 다양한 query를 사용해 보면서 경험을 쌓는 것이 필요하다는 말씀을 드릴 수 밖에 없습니다. 왜냐하면, query의 종류가 거의 무한정으로 바뀔 수 있으므로, 이것을 일일이 지금 분류하여 대응하는 것은 불가능하기 때문입니다.

mcp tool을 사용하기 위해서는 LLM+Tools toggle switch가 ON이어야 합니다. 이럴 경우 RAG에서처럼 굳이 분석할 file들을 서버에 올리고 문서분석을 하지 않아도 입력창에 입력해서 질의를 할 수 있습니다.

우선 openai 가벼운 LLM model을 사용하고 점차 성능 좋은 model을 사용한 후에 google model은 나중에 사용하겠습니다.

mcp tool의 간단한 소개 및 LLM model, 3개의 표현이 살짝 다른 한국어 query에 대해서 설명한 동영상은 다음에 보입니다.

Youtube address: <https://youtu.be/Nip-sK-ilLA>

동영상 제목: 12 mcp tool - start

다음 query를 이용한 mcp tool 사용예는 다음 youtube에 보입니다.

Q1: AI agent라는 문구가 file 내용에 들어있는 ppt file 목록을 보여주세요.

Q2: file 내용에 AI agent가 들어 있는 ppt file 목록을 보여주세요.

Q3: file 내용에 AI agent라는 문구가 들어 있는 ppt file 목록을 보여주세요.

Youtube address: <https://youtu.be/WYFQJlsv9bc>

동영상 제목: 13 mcp tool 예1

여기서 사용되는 search\_ppt는 사실은 제가 다른 기존의 application을 이용해서 database에 power point자료를 올린 것을 검색해서 제공하도록 만든 mcp tool입니다. 이렇게 mcp tool은 기존의 database등을 이용한 application과 연동이 되든지, 다른 형태의 application과의 연동등 다른 준비가 필요합니다.

이번에는 gpt-4.1-nano를 제외하고 gpt, gemini 모든 LLM model이 맞

는 답변을 했습니다. Q2에 대해서는 gpt-4.1-nano는 mcp tool 선택에 실패하고 web search를 시도했으나 이 역시 원하는 결과를 얻지 못했습니다. Q3에 대해서는 gpt-4.1-nano는 Q1에서 처리한 것과 같이 file 목록을 가져오는 단계는 성공했지만 output window에 그 목록을 보이는 단계는 실패한 겁니다.

여기서 보이듯이 AI agent에서 AI agent라는 문구로 표현을 살짝 바꾸니 LLM model이 좀더 명확한 표현으로 인식하여 mcp tool선택에는 성공하여 file 목록을 가져오는 단계는 성공했습니다. 그러나 gpt-4.1-nano model의 한계로 인하여 output window에 AI agent라는 문구를 갖는 ppt file list를 보이는 것은 Q1을 사용한 경우에서와 같이 실패하게 되는 것입니다.

그러면 다른 query로 같은 내용이지만 영어로 바꾼 것을 이용해 보겠습니다.

Q4: please show list of ppt files which has 'AI agent' in the content.

Q5: please show list of ppt files which has content with 'AI agent'

Youtube address: [https://youtu.be/jX\\_BoGV24PQ](https://youtu.be/jX_BoGV24PQ)

동영상 제목: 14 mcp tool 예2

Q4는 Q1과 Q3와 매우 유사한 결과를 보입니다. 특히 Q5는 gemini-2.0-flash도 gpt-4.1-nano와 같은 양상을 보입니다.

이렇게 LLM model과 query가 서로 (궁합이) 잘 맞는 pair가 있으니 이것을 잘 찾아서 query하는 것이 agent사용 효과를 높이게 될 것입니다. 그리고 동영상 예에서 보이듯이 어떤 경우는 비단 이러한 mcp

tool에 대한 답변뿐 아니라 종종 보이는 현상이기도 한데 LLM답변이 너무 과신하는 표현을 사용합니다. 이것은 chat bot을 사용할 때도 보이는 현상이기도 합니다만, 이것을 액면 그대로 받아들이면 나중에 실망이 클 수도 있으니 이러한 LLM의 특성을 미리 알고 사용하는 것이 필요합니다.

앞에서 보여드린 Q1 ~ Q5에 대한 결과는 다음 table처럼 요약해서 볼 수 있습니다.

※ X는 query실패, △는 불충분한 답변을 의미

	Q1	Q2	Q3	Q4	Q5
<b>gpt-4.1-nano</b>	△	X	△	△	△
<b>gpt-5-nano</b>	80.37	80.04	78.26	81.67	78.78
<b>gpt-4o-mini</b>	22.93	19.26	22.89	21.37	34.8
<b>gpt-4.1-mini</b>	22.39	16.35	17.72	17.97	20.43
<b>gpt-5-mini</b>	69.35	59.61	119.63	87.53	72.74
<b>gpt-4o</b>	13.98	14.36	15.71	14.63	15.9
<b>gpt-4.1</b>	17.22	15.61	16.84	14.28	17.53
<b>gpt-5</b>	88.28	76.2	141.06	116.18	148.93
<b>gemini-2.0-flash-lite</b>	12.47	12.84	12.07	12.71	12.17
<b>gemini-2.5-flash-lite</b>	13.38	11.3	11.47	14.96	13.5
<b>gemini-2.0-flash</b>	13.96	13.66	13.32	13.79	△
<b>gemini-2.5-flash</b>	25.95	28.6	25.16	25.86	27.46
<b>gemini-2.5-pro</b>	78.34	77.73	82.64	79.26	76.7

이 mcp tool 사용예에서는 영어 query가 한국어 query에 대해서 특별히 유리한 점을 보이지는 않습니다. 그리고 좀 특이한 점은 기존의 선

입관과는 다르게 gpt-4o, gpt-4.1에 대해서는 오히려 mini가 붙은 것이 더 속도가 약간이나마 느리게 됩니다. 물론 여기서 보이는 것은 너무 작은 양의 예라서 이것으로 특별히 결론을 내리기는 너무 성급할 수 있습니다.

그런데, 여기 보이는 특성의 차이는 뒤에서 더 설명드릴 내용인 AI agent code에 따른 영향일 수 있으므로 나중에 AI agent code가 수정, 개선되는 경우 조금 다른 양상을 보일 수도 있음을 고려하시기 바랍니다. 저는 지속적으로 AI agent의 성능을 개선하는 방향으로 AI agent의 code를 수정, 개선을 할 것이므로 이러한 점을 미리 말씀드립니다.

c. LLM only와 mcp tool을 agent가 판단해서 실행  
 이제는 다른 query를 사용하여 LLM only로 처리하든가 MCP tool을 사용하여 처리하는 예를 보여 드리겠습니다.

Q1: 3 더하기 6은?

Q2: 3 더하기 6은? 도구 사용.

이 query는 분명히 간단해서 LLM only로서 처리가 가능합니다. 다만 도구사용이라는 명시적 추가 지시사항으로 MCP tool중에서 add라는 tool을 사용하도록 명령을 내린 결과도 같이 보여 드립니다. 그런데 여기서 보시듯이 gpt-5는 다른 것과 달리 매우 간단히 답변을 하는 예도 보입니다. 이것은 성능이 좋은 LLM model이라서 답변이 상세하고 길게 보여질 것이라는 예상을 벗어나는 매우 좋은 예입니다. 여기서 보이듯이 LLM only가 당연히 MCP tool을 사용하는 것보다는 응답

시간이 빠르게 됩니다.

Q1, Q2에 대한 여러 LLM model에 대한 결과는 다음 동영상에서 보여 드리겠습니다.

Youtube address: <https://youtu.be/sGLJK4nhXN0>

동영상 제목: 15 mcp tool 예3

여기서 보이듯이 LLM only의 처리 속도는 mcp tool을 이용한 처리속도와는 비교가 안되게 훨씬 빠릅니다. 그러나 LLM only로 처리할 수 있는 것은 매우 제한적이므로 RAG, mcp tool, web tool등을 이용하여 AI agent의 응용범위를 넓히게 되는 것입니다.

## 11. AI agent 에서 web tool 의 사용

RAG나 LLM only, mcp tool로 query에 대한 원하는 답을 찾기 어려우면, agent의 또 다른 도구인 web search tool을 사용해 볼 수 있습니다. web search tool을 web search나 web tool로 줄여서 부르기도 합니다. web tool은 mcp tool과 달리 사전 준비할 내용이 상당히 많이 줄어들게 되나, 경우에 따라서는 mcp tool보다도 유용한 결과를 얻기도 합니다. 물론 internet 또는 web상의 정보가 당연히 query와 밀접한 관계가 있을 때에나 가능한 것이지만요.

web search engine은 여러 종류가 가능하지만, 우리는 tavily search, brave search, duckduckgo search만 사용할 겁니다. 실질적으로 이정도만 해도 대부분을 커버할 수 있다고 믿습니다.

그런데, 이 3가지 중에서 tavily(이제부터 search를 빼고 간단히 부르겠습니다.)가 가장 비용이 비싼 편이지만, 아주 많은 양의 웹검색을 하지 않는 한에는 매우 작은 비용으로도, 또는 무료로 꽤 긴 기간 동안 사용할 수 있습니다. 그런데, tavily 경우에는 pay as you go라는 유료 option을 같이 사용하는 것이 더 정확한 웹검색 결과를 얻는 것으로 경험상 알게 되었습니다. 그런데 이 유료 option은 실제 계속 비용을 지불하는 것이 아니고 아주 소액만 이리 예치를 해두면 pay as you go로 되고, 매달 free 제공 건수(tavily는 1000건, brave search는 2000건)부터 먼저 사용하고 이것을 넘기는 경우에만 pay as you go로 넘어가게 되니 실제로는 아주 많은 웹검색을 하지 않는 한에는 그대로 무료 사용이 되는 겁니다.

그리고, tavily나 brave나 API key를 필요로 하므로 이 key를 얻고 사용 설명에 따르는 대로 .env file이나 env.txt file을 application을 구동시키는 .exe file과 같은 directory에 같이 포함시켜서 사용하시기 바랍니다. duckduckgo search는 이러한 AP key를 필요로 하지는 않으나 상대적으로 검색 결과가 다른 것에 비해서 덜 안정적입니다. 오히려 AP key사용이 더 안정적인 검색을 제공하게 되는 것입니다. OpenAI, Google, Tavily, Brave 모두에 대한 API key값을 가지는 .env file이나 env.txt file의 내용은 아래와 같을 것입니다.

```
--- .env file 내용 ---
OPENAI_API_KEY = "openai api key값"
GOOGLE_API_KEY = "google api key값"
TAVILY_API_KEY = "tavily api key값"
BRAVE_SEARCH_API_KEY = "brave api key값"
```

chatgpt, gemini, claude, grok등의 chat bot 모두 이러한 API key를 얻는 상세한 방법을 잘 설명하고 있으니 참고하시기 바랍니다. 무료 test를 위한 test application에서는 많은 경우에 이러한 API key를 얻으시면 web tool은 그대로 사용하실 수 있도록 될 것이니, 여러가지로 web tool을 사용해 보시면 web tool사용에 대한 경험을 더 풍부히 가지실 수 있을 것으로 예상합니다.

#### a. web tool 사용예1 – 최근 뉴스1

web tool을 사용하시려면 강제 웹검색 toggle switch가 ON이 되거나 또는 이것이 OFF인 경우에도 query 명령에 웹검색을 명시적으로 요구하면 web tool을 이용하시게 됩니다.

앞에서 설명 드린 tavily등의 web search engine내용 및 tavily pay as you go 및 무료사용에 대한 설명등과 아울러 매우 news로서 참신성이 높은 다음 query에 대한 결과를 다음 동영상에서 보여 드립니다.

query: 최근 뉴욕시장에 당선된 사람은 누구인가요?

Youtube address: <https://youtu.be/LoEUUusr32zo>

동영상 제목: 16 web tool 예1

이 query는 동영상을 만든 시점에 비하여 2 ~ 3시간전에 기사로서 언론기관에 노출된 내용이므로 tavily가 상당히 최신 news내용도 검색할 수 있음을 보여 주는 내용입니다. 그리고 동영상에는 당시의 기억으로 brave는 API key가 필요 없는 것으로 언급되었지만, brave 역시 API key가 필요합니다. 다만, brave경우에는 무료의 경우에도 API key를 얻고 매우 많은 양의 웹검색을 하지 않는 한에는 사용이 가능합니다. duckduckgo의 경우는 API key를 필요로 하지 않습니다.

이 동영상에서 보이듯이 gpt-4.1-nano는 이 query에 대하여 답변이 오락가락합니다. 앞에서의 다른 query의 예에서도 이러한 특성을 보인 적이 있습니다. 이러한 행동특성은 query에 따라서 다르게 보입니다. 이 밖에는 gemini-2.0-flash를 제외하고는 모두 이 query에 대해서 잘 답변을 하였습니다.

tavily, brave, duckduckgo중에서 어느 것이 여러분의 요구에 가장 적합한지는 실제로 사용해 보시면서 판단하시기 바랍니다.

b. web tool 시용예2 – 최근 뉴스2

이제 직전의 query에 비하여 news로서의 참신성이 살짝 떨어지는 다음 query에 대해서 알아 보겠습니다.

query: 북한 전 외교부장 김영남 사망은 언제인가요?

이 query는 동영상 만드는 시점에 비하여 하루 이틀전에 언론기관에 노출된 내용입니다. 이 결과는 다음 동영상에서 보여 드립니다.

Youtube address: <https://youtu.be/9XNBa2ZaSn0>

동영상 제목: 17 web tool 예2

여기서 보시듯이 gpt-4.1-nano와 gemini-2.5-flash-lite는 제대로 답변을 못했고, 놀랍게도 gemini-2.5-pro도 답변을 못 했습니다. 이것은 성능이 좋은 LLM model은 대개 제대로 답변할 수 있을 것이라는 일반적인 예상을 크게 뒤엎는 결과이므로 이러한 특성을 기억하고 향후의 LLM model선정에 반영하시는 것이 바람직할 것으로 보입니다. 다만, 이 예로부터 어떠한 법칙을 찾기는 매우 어려워 보입니다. **성능 좋은 model이라고 항상 답변을 잘 한다고 믿지 말라**는 주의 말고는!

c web tool 시용예3 – 참신성이 거의 없는 query

이번의 query는 앞에서 보이는 query와는 달리 참신성이 거의 없는 것입니다.

query 1: FastMCP에 대해 알려주세요.

query 2: FastMCP에 대해 알려주세요. 도구 사용

이 query는 제가 개발자/engineer로서 생각하기 쉬운 주제로 골랐음을 양해해주시기 부탁드립니다. FastMCP는 앞에서 여러 번 보여드린 예로서 mcp tool을 사용했었는데, mcp는 앞에서 살짝 설명드렸지만 model context protocol의 약자입니다. 그리고 FastMCP는 이러

한 mcp tool을 사용하기 위해서 서버에서 사용하는 기법 정도로 이해하시고, 실제로 우리 agent앱에도 서버에 적용한 기법입니다. query의 답변 내용이 복잡하게 보여도 대부분의 사용자는 이점을 무시하시고 답변 내용에 model context protocol이 보이느냐 마느냐로만 판단하시면 됩니다.

이 query들 적용결과에 대한 동영상은 아래에 보입니다.

Youtube address: <https://youtu.be/nVfjpiVX1D0>

동영상 제목: 18 web tool 예3

이번에는 agent를 사용할 때에 강제 웹검색 toggle switch를 ON으로 하지 않고 OFF상태로 둡니다. 대신에 query 1으로도 우리가 원하는 답변인 model context protocol과 관련이 있는 FastMCP로 설명을 한다면, 이것은 LLM model이 우리가 원하는 FastMCP내용으로 훈련이 되었다는 것을 보여 주는 것이므로 굳이 웹검색을 하지 않아도 LLM only로 우리가 원하는 FastMCP내용으로 답변을 얻을 수 있기 때문입니다.

이 동영상 내용으로 보면 gpt-5는 최신의 model이므로 LLM only로 우리가 원하는 내용으로 FastMCP를 설명하지만, 다른 model들은 모두 LLM only로는 모두 이전의 다른 내용으로 학습된 내용을 바탕으로 답변하여 MCP를 model context protocol이 아닌 다른 것의 약자로 해석하여 답변합니다.

이것으로 gpt-5가 시간은 많이 걸리고 비용도 비싼 편이지만, LLM model 자체에 최신의 여러 data를 이용하여 훈련된 것임을 알 수

있으므로 LLM 자체 지식만으로도 많은 답변이 가능한 것을 알 수 있습니다.

그런데 만일 강제 웹검색 toggle switch를 ON으로 한다면 위와 같은 결과를 얻으려면 어떻게 해야 할까요? 이것은 굳이 동영상으로 또 보여 드릴 필요없이 제가 추정하고 확인한 내용만 말씀드리지요.

강제 웹검색 toggle switch가 ON이므로 다음 query는

query: FastMCP에 대해 알려주세요.

로 하면 web tool을 사용하여 답변하게 됩니다. 그러므로 FastMCP에 대해서 LLM only로 답변을 얻기 위해서는 다음처럼 해야 합니다.

query: FastMCP에 대해 알려주세요. 도구 사용 금지.

이것은 확인한 결과, 앞에서 강제 웹검색 toggle switch를 OFF로 하여

query: FastMCP에 대해 알려주세요.

로 한 것과 같은 결과를 보여 줍니다.

## 12. AI agent의 특성을 결정 짓는 agent code

AI agent는 LLM model과 embed model, mcp tool, web tool등 위 구성 요소와 더불어 이들을 연결시켜서 AI agent가 제대로 작동을 하도록 하는 agent code가 있어서 AI agent의 특성을 결정짓게 합니다. 당연히 필요 시에는 mcp tool을 없앤 AI agent가 있을 수 있고, web tool을 없앤 AI agent도 있을 수 있고 아예 mcp tool이나 web tool을 다 제거하여 RAG만 사용하는 AI agent도 있을 수 있습니다. 그리고, LLM model의 종류나 embed model의 종류도 agent code에 의하여 결정이 됩니다.

동영상에서 보시듯이 agent는 여러 처리 경로가 있습니다. 일반 모드/보강 모드(normal model/augmented mode)에서는 다음을 우리가 사용할 수 있습니다.

가. RAG

나. LLM only

다. mcp tool 이용

라. web search tool 이용

그리고, 이 문서의 거의 마지막 부분에서 설명드릴 반사 숙고 모드(reflection mode)도 있습니다.

일반 모드의 실제 사용 예는 앞에서 이미 보여 드린 바 있습니다. 이들의 주요 특성을 다시 요약 정리해서 설명 드리는 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/ULWhfzGbGjg>

동영상 제목: 19 agent 처리 여러 경로

그리고 이에 덧붙여서 입력창에 입력이 가능한 경우에 대한 조건을 agent 화면을 이용하여 설명드리는 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/JlDUpq2Z4eg>

동영상 제목: 20 agent 입력창에 입력가능 경우

### 13. agent 앱의 query처리 우선 순위

앞에서 agent 앱은 agent code가 특성을 정한다고 말씀 드렸습니다.

이 agent앱은 아래 보이는 query처리 우선 순위로 작동하며

## query처리 우선순위

- query와 같이 내리는 지시 사항 우선(도구 사용, 도구 사용금지, 웹검색 사용, use tool, use web search 등)
- LLM+Tools: OFF 이고 강제 웹검색: OFF면 RAG query로 처리
  - 단 이때 도구 사용등 추가지시 있으면 도구 사용으로
- Tool은 mcp tool과 web search tool
- LLM+Tools: ON이면 query에 대해서 우선 LLM only로 답을 할 수 있는지 판단 → 아니면 mcp tool 사용시도 → 아니면 web search 시도 → 아니면 답할수 없다고 답변
- 강제 웹검색: ON이면 RAG, LLM only, mcp tool 모두 건너 뛰고 web search 시도 → 아니면 답할수 없다고 답변

다음 동영상에서 설명 드립니다.

Youtube address: <https://youtu.be/GbARrDXVe-0>

동영상 제목: 21 query처리 우선순위

이것은 제가 1년이 조금 넘는 기간 동안 여러 차례 agent 앱을 사용하면서 나름 가장 최적화된 법칙으로 여겨지는 것을 정하여 구현한 것입니다.

query에 같이 도구 사용, 웹검색 사용등 추가지시 사항을 사용할 수 있는 것이 분명 더 편하므로 이것을 사용하기 위하여 query에서의 지시사항을 최우선으로 정하고, 이 것이 없다면 GUI로 도구 사용 우선 순위를 명시적으로 표시하여 agent를 효과적으로 사용할 수 있도록 한 것입니다. 이 문서, 이 agent앱에서 tool 또는 도구라 함은 mcp

tool과 web search tool을 의미합니다. 이 query처리 우선순위는 매우 중요하므로 항상 기억하시는 것이 효과적 agent사용에 크게 도움이 됩니다. 그래서 질문처리 우선순위 보이기 toggle switch가 ON이면 이 우선순위를 보실 수 있도록 한 것입니다. 이럴 경우 여러분은 agent에서 다음 block을 보시게 될 겁니다.

**작동우선순위 (LLM+Tools가 RAG경우는 OFF, 다른 경우는 ON):**

1. **RAG** - 기존문서검색 최우선
2. **LLM만에 의한 답변** - LLM의 일반 지식활용
3. **MCP tools** - 특별한 MCP도구 이용
4. **웹검색** - 추가적인 웹검색도 활용 가능

"강제 웹검색" 이 ON: 웹검색이 최우선으로 작동

이 것은 동영상에서 설명드리는 것을 압축해서 간단히 보여드리는 것입니다. 이 block의 내용은 이 agent 앱을 많이 사용하시면 저절로 익숙하시게 될 것이고 그러면 work space를 더 넓게 사용하기 위해서 OFF로 하시고 사용하시면 됩니다.

#### 14. 까다로운 웹검색 예

다음 query에 대해서 web search를 해보겠습니다.

query: 미국 제30대 대통령의 아내의 어머니가 돌아가신 날짜는 몇년 몇월 며칠인가요?

아마도 한국에서는 미국 30대 대통령에 대해서 잘 아시는 분이 얼마 안될 것으로 추정합니다. 그리고 실제로 이 query에 대해서 web search를 해 본 결과 LLM model도 맞추는 것이 소수인 것으로 보면 미국에서도 덜 알려진 대통령으로 추정이 됩니다. 게다가 아내의 어머니로 좀 더 관계가 더 벌어지는 경우이니 그만큼 정보가 덜 알려지고 web search에서나 LLM model훈련시에 사용할 data확보에 분명히 어려움이 존재할 것으로 예상은 됩니다.

실제로 대표적 chat bot을 이용하여 얻은 query결과로 정답인 1929년 10월 24일을 맞춘 결과는 아래와 같습니다.

chatgpt: 못 맞춤

gemini: 맞춤

claude: 맞춤

grok: 못 맞춤

여러 LLM model을 이용하여 test해보니 gpt-5-nano, gpt-5-mini, gpt-5, gemini-2.5 pro model만 정답을 맞추고 나머지는 모두 못 맞추었습니다. 이에 대한 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/vWK6GEHmmyo>

동영상 제목: 22 까다로운 웹검색 예

## 15. RAG의 실질적 query적용 예

앞에서 여러 RAG query의 적용 예를 보여드린 바, 지금 보여드리는 예는 앞에서 보여드린 바와는 다르게 한 개의 문서에 대하여 다양한 query를 적용하여 비교적 빠른 embed model로 vector store를 얻더라도 꽤 많은 중요한 문서내용을 파악할 수 있음을 보여 드립니다. 이것은 앞에서 보여드린 drill down에서와는 다르게 질문의 주제가 다양한 특성이 있습니다. 여기에 사용한 문서는 한국산업기술평가관리원이 2011년 발행한 문서로 제목과 저자는 아래와 같습니다.

제목: X선 의료영상기기 기술 및 산업동향

저자: 허영 의료기기 PD

최신 문서는 아니나 국내 국외의 추세에 대하여 비교적 아직도 시사점이 있는 내용을 제공하고 다양한 query에 대하여 RAG를 이용하여 실제로 주요 내용을 잘 파악할 수 있는 좋은 예로 보여 여기서 보여드립니다. 비슷한 주제로 googling을 해보니 이보다 더 오래된 문서도 보이고, 대부분 문서가 학술적 논문이라서 이런 것은 피하고, 더 찾아보니 최신 문서도 보이나 최신 문서는 제가 감당하기에 너무 비용이 많이 드는 것으로 파악되어 이 문서를 사용하게 되었으니 이점 양해해 주시기 바랍니다. 여기서 LLM model은 gpt-4.1-mini만 사용하고 embed model은 better를 사용합니다. 그리고 query는 모두 영어로 되어 있고 기술적 질문이 아니므로 많은 Biz 사용자가 관심을 가질 수 있는 것으로 판단하여 사용했습니다.

이문서는 32페이지 짜리로 실제 본문은 3페이지부터 시작하므로 본문은 30페이지만으로 생각하시면 됩니다.

여기서 사용할 query는 아래와 같습니다. 모두 영어 query인데 영어를 query로 사용하면 앞에서 보셨듯이 가벼운 LLM model도 답변을 잘 합니다. 또한, 여기서는 embed model은 better를 사용했음에도 불구하고 모든 query에 대해서 만족스러운 답변을 얻었습니다.

Q1: what is the purpose of the document?

Q2: from your oversea market study, what is the CAGR(compound annual growth rate) from 2010 to 2015 for your focused market?

Q3: for oversea market, which is expected to have the biggest size?

Q4: for oversea market, which shows the fastest growth?

Q5: what is the major difference of domestic market from the oversea market?

다음 동영상으로 확인해 보시면 가벼운 embed model과 LLM model을 사용하면서도 실질적 RAG query에 만족스러운 답변을 얻게 됨을 확인할 수 있습니다.

Youtube address: <https://youtu.be/9Ufgycoob54>

동영상 제목: 23 실질적 RAG query 사용예

## 16. web search와 연동되는 drill down예

앞에서 보여드린 9-b. drill down 예1 - 서울시 환경정책을 설명드릴 때에 이미 말씀 드렸듯이 web search tool을 연동하시면 drill down 방법으로 얻은 답변의 내용이 훨씬 풍부하고 유용하게 될 수 있습니다. LLM model을 사용해서 query를 얻은 것은 앞에서 사용한 동일한 query를 사용한다 하더라도 동일한 답변을 얻게 될 확률이 매우 낮습니다. 그것은 LLM의 특성상 안에서 처리할 때에 기본적으로 불확실성을 이용한 처리방식으로 발생이 되는 것인데요, 이것은 그만큼 다양한 답변을 내놓기도 하지만 때때로 다른 동작처럼 보이기도 해서 처음에는 혼동을 주기도 합니다. 그러나, 큰 시야로 보면 사실은 동일한 행동을 하는 것으로도 볼 수 있습니다. 이것은 일반 사용자는 LLM이나 AI가 보이는 특성이구나 생각하시고 지나가시면 됩니다.

앞에서 설명 드린 내용과 동일한 답변을 재현할 수는 없으므로, 이번에는 앞에서 설명드린 내용에 뒤이어서 끝 부분만 web search tool을 적용한 것이 아닌, 처음 시작하는 query만 동일한 것으로 시작했으며, 이어지는 심화질의는 앞에서의 drill down과는 다른 것이 되었습니다.

처음 시작하는 query는 아래와 같습니다.

query: 서울시의 환경 정책에 대해 자세히 알려주세요.

그런데 이번에는 유용한 정보를 얻는데 집중하기 위하여 추가지시 사항을 이용하는 것은 피했습니다. drill down을 하면서 느낀 것은 RAG에 사용한 문서가 서울시 환경정책에 대하여 꽤 유용한 정보를 제공한다는 것과 운도 어느정도 따라서 마지막에는 web search도 성공했다는 것입니다. drill down의 depth가 꽤 깊은데도 이 문서는 내용이 관련 심의 질의에도 잘 답변할 수 있을 정도로 내용은 알차게 있는

것으로 판단되고, 또한 우리의 drill down이 정보를 제대로 깊게 얻는 데에도 유용할 수 있는 것으로 판단되었습니다. 이에 대한 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/RUuRZ9vV17A>

동영상 제목: 24 drill down 4 서울시 환경정책 web search 연동

이번에는 9-c. drill down 예2 - 서울시 자율주행 관련계획에서 보여드린 drill down에 이어서 web search tool을 연동하여 설명드리겠습니다. 이번에는 9-c.에서 보여드리는 동영상에 그대로 이어서 web search만을 더 추가해서 설명 드리는 데요. 이것은 사실은 여기서는 agent의 내용을 그대로 이용해서 다시 web search부분만 추가해서 설명 드리기 때문에 가능한 일입니다.

참고로 9-c.에서 보여드리는 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/5K3EUwrqngk>

동영상 제목: 8 drill down 2 - 서울시 자율주행 관련계획

이번에도 drill down이 서울시 자율주행 관련계획에 대하여 충분히 정보를 얻게 함을 확인 하였고, 이것은 여기서 문서분석에 시용된 문서인 2040 서울도시기본계획이 이에 대한 알찬 내용을 포함하기 때문입니다. 새로 web search를 연동하게 된 부분을 설명하는 동영상은 아래와 같습니다.

Youtube address: <https://youtu.be/2NsP3NM4NqU>

동영상 제목: 25 drill down 5 - 서울시 자율주행 관련계획 web search 연동

## 17. 반사숙고 모드(reflection mode)

여러분의 학창 시절에 대부분은 다음과 같은 작문 연습이 있었을 겁니다. 선생님께서 여러분께 어떠한 주제로, 어떠한 형식으로 글을 써서 제출하라고 말씀하시고, 여러분은 이에 맞추어 글을 쓰시고 제출하시면, 선생님께서 검토하시고 의견을 말씀하시면 여러분은 이것을 반영하여 다시 제출하시지요. 이러한 과정을 통해서 글이 좀 더 나은 내용을 포함하게 되겠지요. 이러한 것을 모방한 것이 반사숙고 모드로 생각하시면 됩니다. 물론 이 agent는 아직은 실생활에서의 응용은 아직 좀 부족할 수 있을 겁니다. 우선 글의 내용이 여러 페이지 이상의 장문이 되기는 어렵습니다. 그리고 선생님처럼 꼼꼼히 봐주시는 내용에는 한참 못 미칠 수 있을 겁니다. 그럼에도 불구하고, 이것은 꽤 유용하게 다른 장점을 가지고 응용할 수 있다고 판단됩니다. 그래서 이것을 구현해서 여러분께 보여드리는 겁니다. 이것은 사실 여러분과 선생님처럼 여러 관계를 신경쓰면서 작성하고, 검토하고 그럴 필요없이 얼마든지 여러분이 원할 때 사용할 수 있습니다. 그리고 예를 통해서 여러분이 살짝 그능력을 보실 수 있겠지만, 꽤 LLM이 똑똑합니다. 아마도 더 좋은 글은 더 성능좋은 LLM을 사용하시면 될 것으로 추정이 되는데요. 웬만한 것도 내용이 괜찮습니다.

reflection mode는 generator와 reviewer 2개의 agent가 번갈아 가면서 사용자의 요청 및 지시에 따라서 반복적으로 처리하는 agent 처리방식을 의미합니다. 여기서 generator는 앞에서 언급한 학생 역할, reviewer는 선생님 역할을 하는 agent로 생각하시면 이해하기 쉽습니다. reflection mode는 여기서 보여드리는 글쓰기에 적용하는 것 이외에도 여러가지에 응용할 수 있다고 보여집니다. 이것은 우리의 상상력

에 따라 얼마든지 응용범위가 넓어질 수 있다고 봅니다. 여기서는 이러한 시작점으로 우선 글쓰기 분야에 적용해 보겠습니다. 그리고, reflection을 통하여 LLM의 생성형AI(generative AI)능력을 잘 확인할 수 있을 것으로 봅니다. 저도 이것을 구현하고 사용해 보면서 LLM의 능력이 정말로 여러가지가 잘만 쓰면 대단하구나 하는 것을 느꼈습니다. 물론 AI agent는 이러한 능력을 사용하는 것이니까, 저절로 커지는 능력을 그대로 사용할 수 있는 것이지요.

그러면 이번에는 LLM은 gpt-4.1을 사용하고 초기 사용자 요청, generator 지시사항, reviewer 지시사항을 아래와 같이 정해서 reflection을 사용해 보겠습니다.

사용자 요청

미국 허밍웨이의 노인과 바다의 문학적 영향에 대해서 에세이를 작성하세요

generator 지시사항

당신은 글작성 비서로서 3단락 에세이를 작성하는 임무가 있습니다. 사용자가 요청하는 것을 당신의 최선을 다해 작성하세요. 사용자가 비평을 제공하면 이전 버전을 비평을 반영하여 수정하세요.

reviewer 지시사항

당신은 당신에게 제공되는 글을 비평하는 비평가입니다. 제공되는 글을 내용, 스타일과 필요하면 역사 문화적 배경, 지역적 상황에 맞추어 구체적 비평을 제공하세요.

이것을 이용하여 3 cycle을 얻은 것을 아래 동영상에 보입니다.

Youtube address: <https://youtu.be/-ROLRzEsSBA>

동영상 제목: 26 reflection mode 노인과 바다 에세이

여기서 generator의 입력은 generator input window에 입력되고 이 window는 또한 reviewer의 output window가 됩니다. 그리고 generator output window는 reviewer의 input window가 됩니다. 그리

고 generator와 reviewer의 input history는 각각 up/down arrow, < > button으로 navigation이 됩니다. 이러한 행위를 input history navigation이라 부릅니다. 이를 통해서 agent로 reflection을 끝 마친 후에도 나중에 필요하다면 이를 통하여 내용을 다시 얻을 수 있습니다. 동영상에 보이는 word file은 이러한 방식으로 얻은 것입니다. reflection mode에 대해서도 source를 공개할 예정이니 여러분이 많이 사용해보시기 바랍니다.

제가 동영상에 보이는 주제는 노인과 바다에 관한 주제이지만, 따로 한강의 소년이 온다등 다른 주제로 해봤었는데 LLM이 reflection답을 잘 내놓았었습니다. 여러분도 여러 주제로 test해보시기 바랍니다. 이 reflection은 전적으로 LLM model의 성능에 따라서 출력의 품질이 얻어집니다. 그리고 우리가 요청하는 주제가 만일 LLM이 학습한 것에 없으면, 그에 대한 주제는 reflection에 적용할 수 없습니다. 당연한 말씀이지만.

## 18. API key적용 방법

여러분이 무료 test application을 이용하여 AI agent를 사용하시려면 다음 4개 회사의 API key가 필요하거나, 경우에 따라서는 OpenAI를 제외하고 3개의 API key가 필요하거나 또는 Google, Tavily 2개만 필요할 수도 있으며 그것은 Application에 따라서 달라지게 됩니다. 그 특정 test application의 사용방법에 안내해 드리게 될 것입니다.

OpenAI

Google

Tavily

Brave Search

이러한 회사의 API key를 얻는 방법은 chagpt, gemini등의 chat bot을 이용하시면 상세히 잘 설명하고 있으므로 여기서는 생략하겠습니다. key를 얻으신 후에는 application을 구동할 .exe나 .bat file과 같은 directory내에 .env file이나 env.txt file을 아래와 같이 만드시기 바랍니다. 이것은 text file이므로 Windows의 메모장을 이용하시면 편합니다.

--- .env or env.txt file 내용 ---

```
OPENAI_API_KEY = "openai api key값"
```

```
GOOGLE_API_KEY = "google api key값"
```

```
TAVILY_API_KEY = "tavily api key값"
```

```
BRAVE_SEARCH_API_KEY = "brave api key값"
```

이에 관한 동영상은 아래와 같습니다.

Youtube address: [https://youtu.be/1f1FjDo\\_wt0](https://youtu.be/1f1FjDo_wt0)

동영상 제목: 27 API key file 이용하는 법

위의 동영상에서는 4개의 API key로 공개된 source로 설명드리지만,

앞에서 말씀드렸듯이 공개된 source가 아니라 무료로 공개되는 test application을 의미하는 것이며 API key도 4개가 아닐 수도 있습니다.

## 19. 무료 test application

제가 앞에서 설명드린 RAG는 문서를 서버에 올리고 문서 분석을 진행하는 것으로 설명을 드렸는데, 시간이 지남에 따라 문서를 서버에 올리고 분석된 내용을 저장하여 사용자는 이 저장된 내용만 사용하는 방법이 많은 경우에 유용성이 클 수 있음을 발견하게 되었습니다. 그리고 이것은 반복적인 학습이 필요한 경우에 크게 도움이 될 수 있어서 여러 문서에 적용하여 test를 해 보았습니다. 그리고 많은 경우에 이러한 방법이 문서학습에 크게 도움이 된다는 확신도 생겼습니다. 그런데, 이것을 이용하기 위해서는 저작권이 있는 여러 partner들도 필요하다는 사실도 파악하게 되었고, 이러한 partner들과 협력하기 위해서는 기존의 source공개 정책을 변경하는 것이 필요하다는 인식도 생겼습니다. 그래서 이 보다는 test를 할 수 있게 여러 application을 만들어서 한정된 기간이지만 무료로 공개하는 것이 바람직하다는 생각이 들어서 그 것을 준비하여 이제 알려 드리게 되었습니다.

지적재산권 분쟁소지가 전혀 없는 문서를 이용하여 문서분석을 하고 이것을 미리 저장하여 일반사용자는 AI agent를 이용한 여러 query를 test할 수 있게 만든 다음 application은 Bible입니다.

제가 개신교식 성경에 대해서는 잘 몰라서 가톨릭 성경을 대상으로 하여 지적 재산권 문제로부터 매우 자유로운 영어로 된 Latin-2009-Vulgate-Bible.pdf 문서를 사용하여 문서 분석을 하고 일반사용자는 이 분석된 내용을 이용하여, 신약, 구약 성경에 대한 의미 검색을 할 수 있습니다.

이것은 성경(가톨릭) AI agent App.pdf에 자세히 설명드리고 있으니 참조하시고, bible.zip이나 bible.7z을 download 받으시고, Gemini, Tavily, BraveSearch에 대한 API key를 받으시어 이 key들을 api\_config.exe를 이용해서 입력한 다음 저장하면 env.txt가 생성이 되고, 그러면 bible.bat를 더블클릭하시면 성경(가톨릭) AI agent App을 이용하실 수 있게 됩니다. 핵심적 내용요약은 이 문서의 8.에 기술하고 있으니 이를 참조하시면 됩니다.

## 20. 끝맺음

초기에 말씀드린 대로 이 문서는 개발자나 엔지니어가 아닌 일반 사용자가 AI agent에 대해서 쉽게 이용하실 수 있도록 가능하면 기술적 내용은 피하고 핵심적 내용을 이해해서 agent를 똑똑한 내 비서로 여기고 이용하실 수 있도록 노력하여 작성하였습니다만, 아직은 부족한 점이 많을 것으로 여겨집니다. 그런데 어떠한 내용은 기술적 지원이 있어야만 사용이 가능한 것도 꽤 있어서 그러한 기술적 지원에 관한 문의는 아래 email주소로 연락주시기 바랍니다.

email: koagentcom@gmail

담당자: 공학박사 조대철

내용상 보완점이나 더 좋은 의견이 있으시면 위의 email로 연락주시면 감사하겠습니다.

그리고 아래 웹페이지에 download할 문서나 파일과 아울러 새로 알려 드릴 정보도 여기에 실을 예정이니 참고 하시기 바랍니다.

ko-agent web page 주소: [www.ko-agent.com:8080](http://www.ko-agent.com:8080)

감사합니다.