

팀명 : 또 너야

고급RAG와 LLM파인튜닝을 이용한 컴퓨터 주천 및 분석서비스

부제목 : 다나와보다 더나와~

202121069 최재우

201921001 박준혁

201921005 신동훈

contents

1. 프로젝트 선정 이유 및 목표
2. 데이터의 우수성
3. LLM 파인튜닝 과정
4. selfqueryretriever를 이용한 RAG란?
5. 성능 측정
6. 실습

검색엔진 상위 카테고리의 견적판매 사이트들의 분석

	카테고리 분류	견적의 설명	견적 비교	견적 문의
SHOP 다나와	금액, 게임별, 용도	스펙만 제공	X	게시글 형식
ASSACOM	용도	실행 가능 프로그램만	X	게시물 형식
MYPCSHOP	금액, 게임별, 용도	X	X	게시물 형식
쿨젠 컴퓨터	게임, 용도	X	X	게시물 형식
컴마왕	가격과 용도를 합쳐서 분류	X	X	게시물 형식
견적왕	게임, 용도	O	X	게시물 형식
퀘이사존	게임, 용도	O	X	게시물 형식

우리 서비스의 강점

타 사이트의 결론

	카테고리 분류	견적의 설명	견적 비교	견적 문의
타 사이트	사이트의 제공된 분류 선택	좋은 설명들이 있지만 없는 경우가 더 많음	없음	Q&A 게시물 형식

우리 프로젝트의 결론

	카테고리 분류	견적의 설명	견적 비교	견적 문의
다나와보다 더나와	사용자의 질문에 맞춰 파악	학습된 LLM이 제공	추천 견적을 서로 비교 가능	실시간 LLM 채팅

— 프로젝트 개요 - 우리 서비스의 목표점

실시간 견적 상담

RAG를 통해 제공된 견적 정보들을 파인튜닝된 LLM이 분석하여 사용자와의 상담을 주도 및 비전문가인 사람들 또한 실시간 상담을 하면서 모르는 부분을 응답받을수있음.

컴퓨터 부품의 전문성

부품에 대한 전문성 있는 데이터를 가지고 있어야함

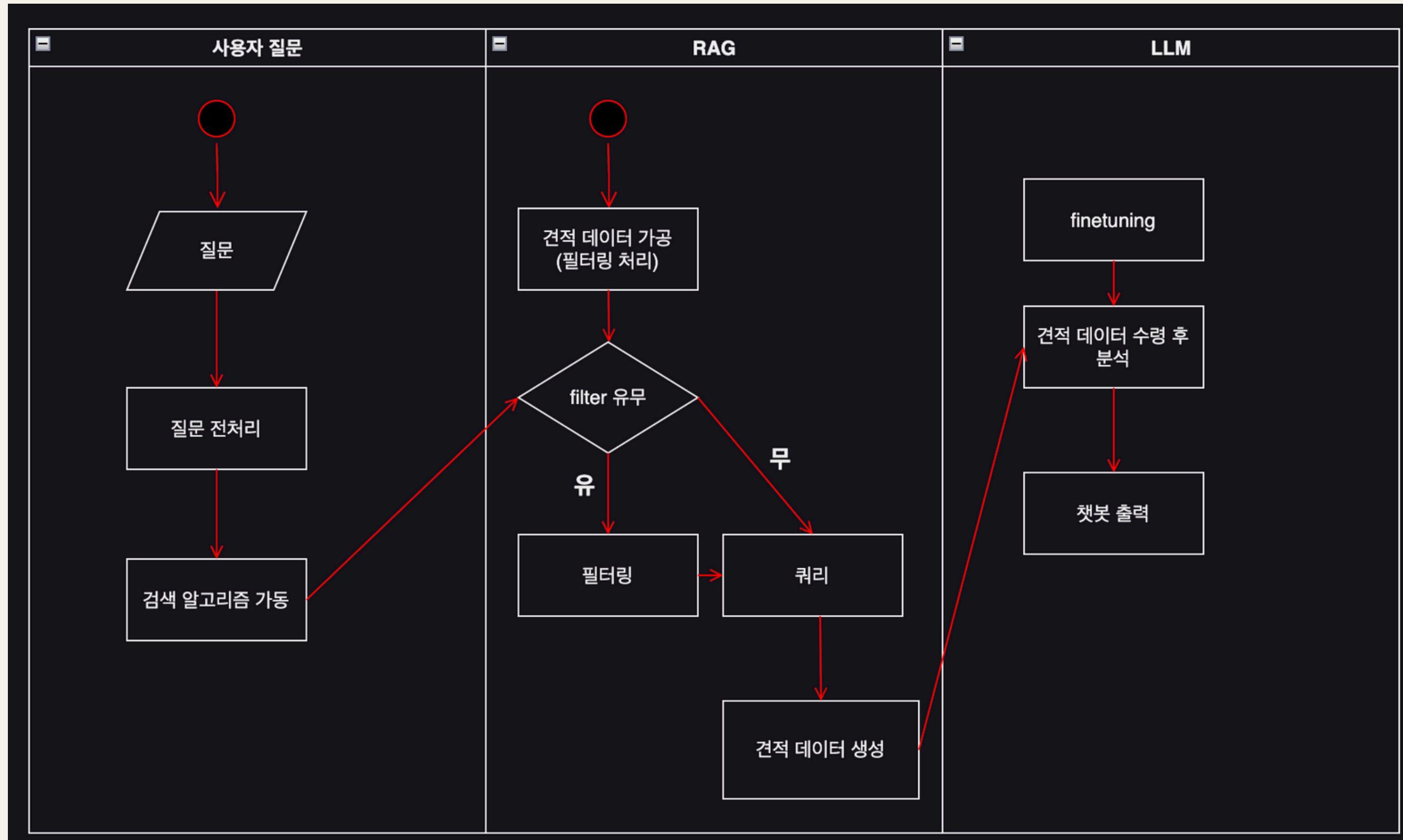
4천여개의 견적데이터

다양한 견적 사이트들의 견적 데이터를 가공하여 저장

견적 추천 알고리즘

4천여개의 견적 데이터들을 고유의 RAG기반 추천 알고리즘을 적용하여 사용자 맞춤형 견적 제공

프로젝트 순서도

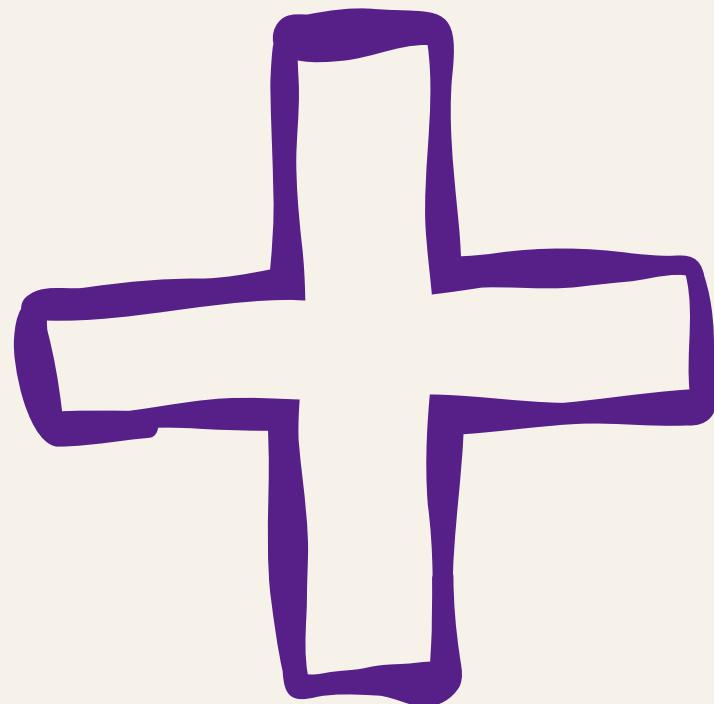


사용자 질문을 통해 RAG가 필터와 시멘틱서치를 통해 견적데이터를 뽑고
파인튜닝된 LLM이 견적에 대한 비교와 분석을 하여 사용자의 합리적인 소비를 도와줌

기술 설명 - RAG + LLM

Rag

빠른 구현
최신 데이터 적용 용이(가격데
이터, 견적데이터등)
검색 필터 활용 가능



Fine Tuning

Domain 특화 작업에 유효
(컴퓨터부품같은 전문성 데이
터)

기술 설명 - finetuning 단독 사용의 경우

01

지속적인 fine-tuning 필요
(시간문제)

02

지속적인 fine-tuning 필요
(비용문제)

03

망각 대응 어려움
(이전 데이터를
잊지못함 - ex: 견적데이터)

RAG 설명 - retriever

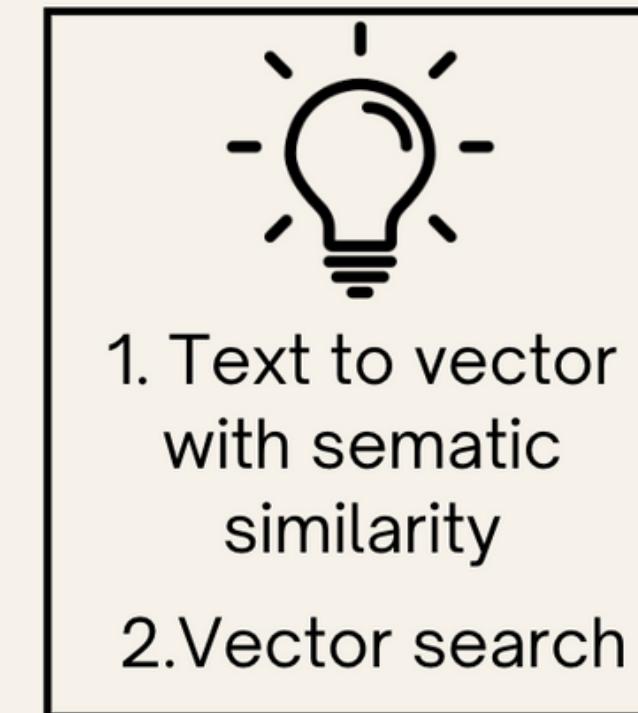
- 유사도 기반 query 검색

질의

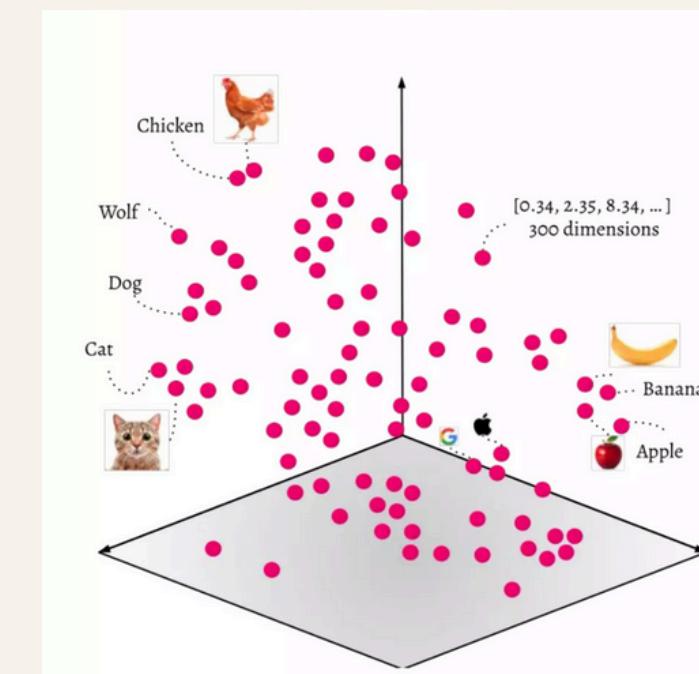
Does this work with my XBox?



시멘틱 검색

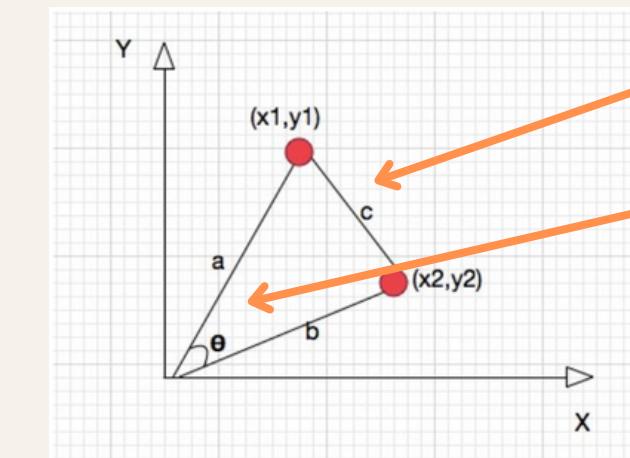


- “Hello” [0.1, 0.52, 0.2, 0.3, 0.7]
“你好” [0.1, 0.53, 0.2, 0.3, 0.7]
“안녕” [0.1, 0.51, 0.2, 0.3, 0.7]



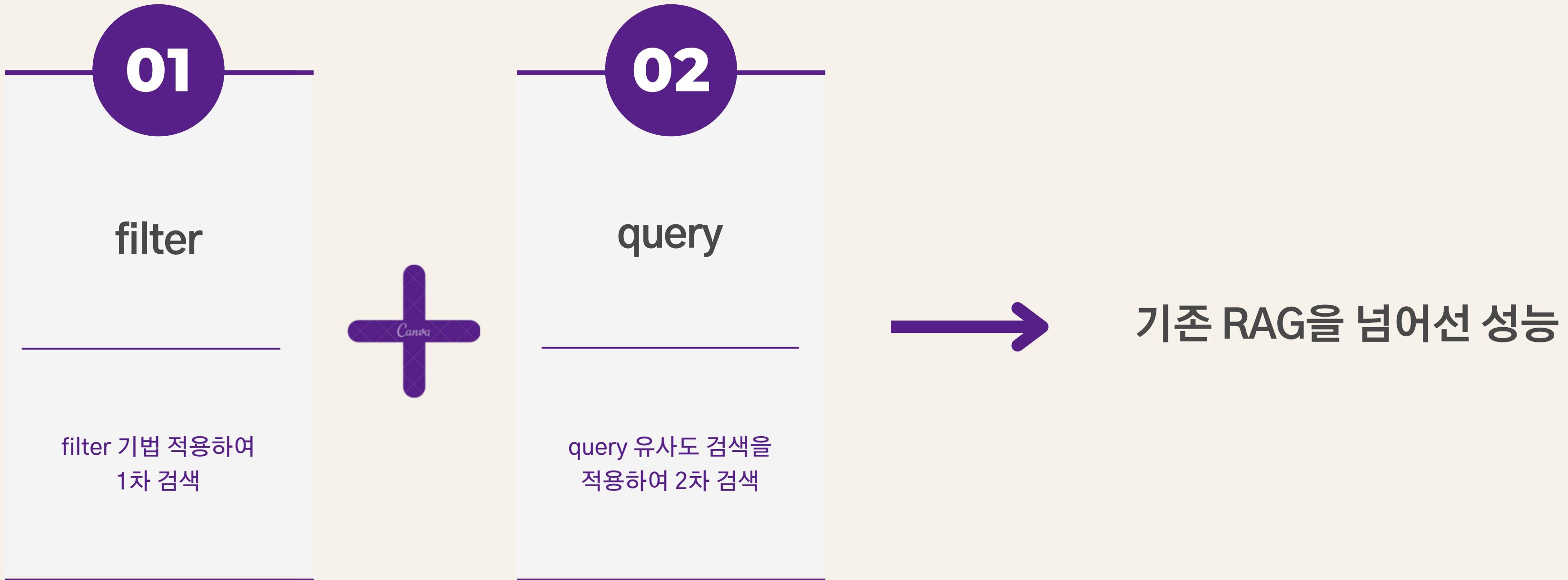
결과

- Do they work with Xbox One?
Can I use for the Xbox One?
How do I use this headset with my Xbox One?
Is this headset compatible with Xbox 1S or above?



□ Euclidean distance
Cosine distance

— 주천알고리즘을 강화시킨 retriever



— RAG 설명 - filter 적용을 위한 견적 데이터 정제

강화된 검색/추천 알고리즘

가공 항목	가공 내용
퍼포먼스	벤치마크 점수를 수집하여 데이터에 적용. 이 데이터를 근거로 기준을 삼아 고/중/저 성능을 판단
가격	쇼핑몰 사이트에서 가격데이터 크롤링하여 데이터에 적용. 이를 근거로 해당 견적 총 금액 산정. 이 데이터는 24시간 기준으로 업데이트.
주요 부품	특정 부품 포함한 견적을 사용자가 요구할 시 특정 부품을 필터링 항목에 포함시켜 검색.



filter 적용하여 검색

스키마 규칙 및 필터를 적용한 모습(예제)

When responding use a markdown code snippet with a JSON object formatted in the following schema:

```
```json
{
 "query": string \ text string to compare to document contents
 "filter": string \ logical condition statement for filtering documents
}
```

```

The query string should contain only text that is expected to match the contents of documents. Any conditions in the filter should not be mentioned in the query as well.

A logical condition statement is composed of one or more comparison and logical operation statements.

A comparison statement takes the form: `comp(attr, val)`:

- `comp` (eq | ne | gt | gte | lt | lte | contain | like | in | nin): comparator
- `attr` (string): name of attribute to apply the comparison to
- `val` (string): is the comparison value

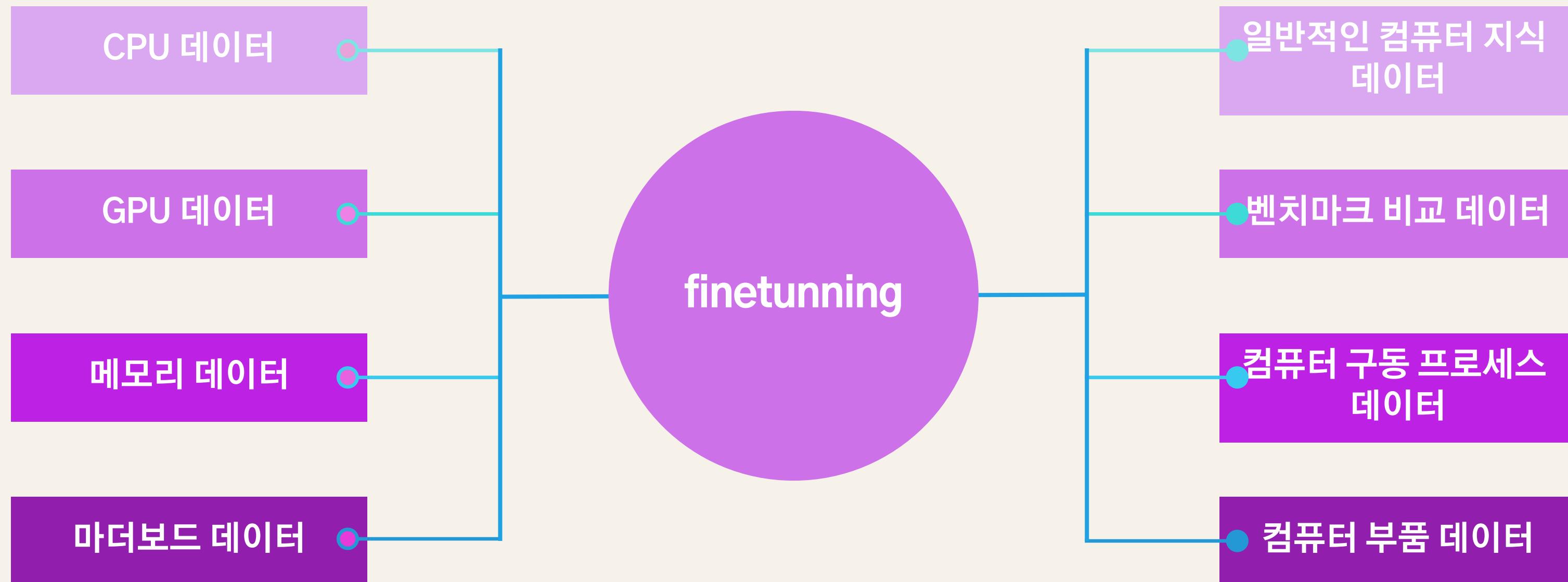
A logical operation statement takes the form `op(statement1, statement2, ...)`:

- `op` (and | or | not): logical operator
- `statement1`, `statement2`, ... (comparison statements or logical operation statements): one or more statements to apply the operation to

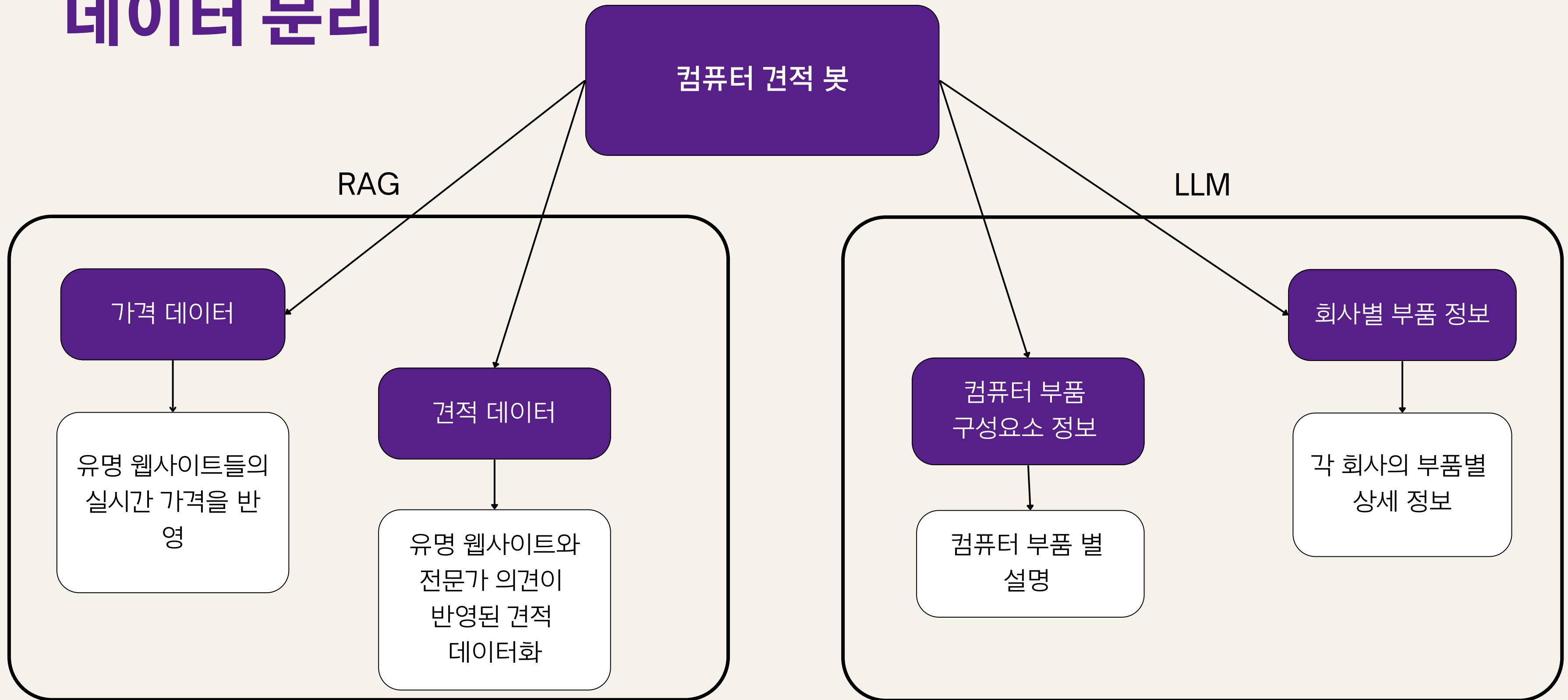
"300만원 이내의 고성능 그래픽 작업용 PC를 추천해줘.",

```
{
  "query": "300만원이내, 그래픽 작업용",
  "filter": "and(gte('total_price', 2800000), lte('total_price', 3000000), eq('performance_grade', '고성능'))"
}
```

기술 설명 - finetunning



특성에 맞춘 데이터 분리



데이터의 우수성 - 고급 RAG 데이터



영예인간17호

활동을 많이 못하지만, 최선을 다해 견적 구성 도와드리겠습니다.

추천견적

14960 LU
136 구매 0 리뷰 96 견적 6 후기

f t n



PC디자이너

화이트 狂人 PC 디자이너

성능은 기본입니다.
감성도 함께 전달합니다.

추천견적

1470 LU
13 구매 0 리뷰 523 견적 4 후기

f t

견적맨소개

안녕하세요
웨이사존, 쿨엔조이, 다나와 등에서 활동하는 커린이
영예인간17호입니다.
커린이 입장에서 영여로운 컨셉의 견적을 진행해
볼까 합니다
남들이 보기엔 언발란스한 견적, 자기만족 견적, 가
성비 견적 등 다양한 컨셉으로 진행해보도록 하…

상세보기

견적맨소개

화이트 狂人 PC 디자이너

단순한 게임기, 단순한 사무기기로써의 PC를 지양합니다.
PC는 전원을 넣지 않아도 행복을 줘야 하는 디지털
기기입니다.

성능은 기본입니다.
감성도 함께 전달합니다.

상세보기

견적맨소개

SKY WALKER

13640 LU
124 구매 2 리뷰 117 견적 5 후기

f t

스카이워커

안녕하세요! 인기 있는 활동하고 있는 Skywalkers입니다.

후원사 답변 팜피씨샵

To. user_963193 님

안녕하세요.
고려중이신 예산과 사용 용도를 기반으로 조립 서비스까지 포함하여 견적을 구성하였습니다
해당 사용 용도 기반이라면 라이젠 7950X3D + RTX 4090 조합으로 구매하시는 편을 추천드릴게요

메모리 부분은 게이밍 프레임 향상을 위해 EXPO / XMP 튜닝램으로 구매하시는 편을 추천드리고 있어요
DDR5 메모리 기준 EXPO / XMP 적용하여 사용하시는 경우라면 하이닉스 칩셋을 사용한 메모리가 불량률도 낮고 호환성이 좋으며 안정성도 높은 부분이라
SK하이닉스 칩셋을 사용한 ESSENCORE 메모리로 구성하시는 편을 추천드립니다

NZXT C 시리즈 라인업 제품이 ATX 3.0 정식으로 지원하는 제품에서 가성비도 좋은 제품이며 CWT OEC 제품이라 안정성 등 충분한 모델이라 보고있고 AS 기간이 10년이면서 AS 정책이 좋다보니 해당 모델로 구성하시는 편을 권장드립니다

From. 웨이사존 공식 후원사 팜피씨샵
견적에 대해 궁금하신 점과 부품 변경을 원하시면 아래 카카오톡으로 채팅 남겨주세요!
https://pf.kakao.com/_epfZb

데이터의 우수성 - 학습 데이터

NVIDIA GeForce RTX 4090

| AD102 | 16384 | 512 | 176 | 24 GB | GDDR6X | 384 bit |
|--------------------|-------|------|------|-------------|-------------|-----------|
| GRAPHICS PROCESSOR | CORES | TMUS | ROPs | MEMORY SIZE | MEMORY TYPE | BUS WIDTH |

Recommended Gaming Resolutions: 1920x1080, 2560x1440, 3840x2160

The GeForce RTX 4090 is an enthusiast-class graphics card by NVIDIA, launched on September 20th, 2022. Built on the 5 nm process, and based on the AD102 graphics processor, in its AD102-300-A1 variant, the card supports DirectX 12 Ultimate. This ensures that all modern games will run on GeForce RTX 4090. Additionally, the Directx 12 Ultimate capability guarantees support for hardware-raytracing, variable-rate shading and more, in upcoming video games. The AD102 graphics processor is a large chip with a die area of 609 mm² and 76,300 million transistors. Unlike the fully unlocked TITAN Ada, which uses the same GPU but has all 18432 shaders enabled, NVIDIA has disabled some shading units on the GeForce RTX 4090 to reach the product's target shader count. It features 16384 shading units, 512 texture mapping units, and 176 ROPs. Also included are 512 tensor cores which help improve the speed of machine learning applications. The card also has 128 raytracing acceleration cores. NVIDIA has paired 24 GB GDDR6X memory with the GeForce RTX 4090, which are connected using a 384-bit memory interface. The GPU is operating at a frequency of 2235 MHz, which can be boosted up to 2520 MHz, memory is running at 1313 MHz (21 Gbps effective). Being a triple-slot card, the NVIDIA GeForce RTX 4090 draws power from 1x 16-pin power connector, with power draw rated at 450 W maximum. Display outputs include: 1x HDMI 2.1, 3x DisplayPort 1.4a. GeForce RTX 4090 is connected to the rest of the system using a PCIe 4.0 x16 interface. The card's dimensions are 304 mm x 137 mm x 61 mm, and it features a triple-slot cooling solution. Its price at launch was 1599 USD.

| Graphics Processor | Graphics Card | | | | | | | | | | | | | | | | | | | | | | | | |
|---|--------------------------------|----------------------------|-----------------------------|------------------------------|------------------------|---------------------------------|------------------------|----------------------------|-------------------|------------------------------------|-------------------|-------------------|---------------|---------------------|---------------------------|------------------|-------------|-------------------------|-----------------|-------------------|-------------|-------------------------|-------------|------------|-------------------|
| GPU Name: AD102 | Release Date: Sep 20th, 2022 | | | | | | | | | | | | | | | | | | | | | | | | |
| GPU Variant: AD102-300-A1 | Availability: Oct 12th, 2022 | | | | | | | | | | | | | | | | | | | | | | | | |
| Architecture: Ada Lovelace | Generation: GeForce 40 | | | | | | | | | | | | | | | | | | | | | | | | |
| Foundry: TSMC | Predecessor: GeForce 30 | | | | | | | | | | | | | | | | | | | | | | | | |
| Process Size: 5 nm | Production: Active | | | | | | | | | | | | | | | | | | | | | | | | |
| Transistors: 76,300 million | Launch Price: 1,599 USD | | | | | | | | | | | | | | | | | | | | | | | | |
| Density: 125.3M / mm ² | Current Price: Amazon / Newegg | | | | | | | | | | | | | | | | | | | | | | | | |
| Die Size: 609 mm ² | Bus Interface: PCIe 4.0 x16 | | | | | | | | | | | | | | | | | | | | | | | | |
| Reviews: 74 in our database | | | | | | | | | | | | | | | | | | | | | | | | | |
| Relative Performance | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>GeForce RTX 3080 Ti</td> <td>60%</td> </tr> <tr> <td>Radeon RX 6950 XT</td> <td>60%</td> </tr> <tr> <td>GeForce RTX 3090</td> <td>61%</td> </tr> <tr> <td>GeForce RTX 4070 Ti</td> <td>63%</td> </tr> <tr> <td>GeForce RTX 4070 Ti S...</td> <td>68%</td> </tr> <tr> <td>Radeon RX 6950 XT</td> <td>70%</td> </tr> <tr> <td>GeForce RTX 3090 Ti</td> <td>71%</td> </tr> <tr> <td>GeForce RTX 4080</td> <td>80%</td> </tr> <tr> <td>GeForce RTX 4080 SUP...</td> <td>81%</td> </tr> <tr> <td>Radeon RX 7990 XT</td> <td>82%</td> </tr> <tr> <td>GeForce RTX 4090</td> <td>100%</td> </tr> </table> | | GeForce RTX 3080 Ti | 60% | Radeon RX 6950 XT | 60% | GeForce RTX 3090 | 61% | GeForce RTX 4070 Ti | 63% | GeForce RTX 4070 Ti S... | 68% | Radeon RX 6950 XT | 70% | GeForce RTX 3090 Ti | 71% | GeForce RTX 4080 | 80% | GeForce RTX 4080 SUP... | 81% | Radeon RX 7990 XT | 82% | GeForce RTX 4090 | 100% | | |
| GeForce RTX 3080 Ti | 60% | | | | | | | | | | | | | | | | | | | | | | | | |
| Radeon RX 6950 XT | 60% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 3090 | 61% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 4070 Ti | 63% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 4070 Ti S... | 68% | | | | | | | | | | | | | | | | | | | | | | | | |
| Radeon RX 6950 XT | 70% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 3090 Ti | 71% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 4080 | 80% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 4080 SUP... | 81% | | | | | | | | | | | | | | | | | | | | | | | | |
| Radeon RX 7990 XT | 82% | | | | | | | | | | | | | | | | | | | | | | | | |
| GeForce RTX 4090 | 100% | | | | | | | | | | | | | | | | | | | | | | | | |
| Based on TPU review data: "Performance Summary" at 1920x1080, 4K for 2080 Ti and faster. | | | | | | | | | | | | | | | | | | | | | | | | | |
| Theoretical Performance | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>Pixel Rate: 443.5 GPixel/s</td> <td>Memory: 24 GB</td> </tr> <tr> <td>Texture Rate: 1,290 GTexel/s</td> <td>Render Config:</td> </tr> <tr> <td>FP16 (half): 82.58 TFLOPs (1:1)</td> <td>Shading Units: 16384</td> </tr> <tr> <td>FP32 (float): 82.58 TFLOPs</td> <td>TMUs: 512</td> </tr> <tr> <td>FP64 (double): 1,290 GFLOPs (1:64)</td> <td>ROPs: 176</td> </tr> </table> | | Pixel Rate: 443.5 GPixel/s | Memory: 24 GB | Texture Rate: 1,290 GTexel/s | Render Config: | FP16 (half): 82.58 TFLOPs (1:1) | Shading Units: 16384 | FP32 (float): 82.58 TFLOPs | TMUs: 512 | FP64 (double): 1,290 GFLOPs (1:64) | ROPs: 176 | | | | | | | | | | | | | | |
| Pixel Rate: 443.5 GPixel/s | Memory: 24 GB | | | | | | | | | | | | | | | | | | | | | | | | |
| Texture Rate: 1,290 GTexel/s | Render Config: | | | | | | | | | | | | | | | | | | | | | | | | |
| FP16 (half): 82.58 TFLOPs (1:1) | Shading Units: 16384 | | | | | | | | | | | | | | | | | | | | | | | | |
| FP32 (float): 82.58 TFLOPs | TMUs: 512 | | | | | | | | | | | | | | | | | | | | | | | | |
| FP64 (double): 1,290 GFLOPs (1:64) | ROPs: 176 | | | | | | | | | | | | | | | | | | | | | | | | |
| Memory | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>Memory Size: 24 GB</td> <td>Memory Type: GDDR6X</td> </tr> <tr> <td>Texture Rate: 1,290 GTexel/s</td> <td>Memory Bus: 384 bit</td> </tr> <tr> <td>FP16 (half): 82.58 TFLOPs (1:1)</td> <td>Bandwidth: 1.01 TB/s</td> </tr> </table> | | Memory Size: 24 GB | Memory Type: GDDR6X | Texture Rate: 1,290 GTexel/s | Memory Bus: 384 bit | FP16 (half): 82.58 TFLOPs (1:1) | Bandwidth: 1.01 TB/s | | | | | | | | | | | | | | | | | | |
| Memory Size: 24 GB | Memory Type: GDDR6X | | | | | | | | | | | | | | | | | | | | | | | | |
| Texture Rate: 1,290 GTexel/s | Memory Bus: 384 bit | | | | | | | | | | | | | | | | | | | | | | | | |
| FP16 (half): 82.58 TFLOPs (1:1) | Bandwidth: 1.01 TB/s | | | | | | | | | | | | | | | | | | | | | | | | |
| Board Design | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>Slot Width: Triple-slot</td> <td>Shading Units: 16384</td> </tr> <tr> <td>Length: 304 mm</td> <td>TMUs: 512</td> </tr> <tr> <td>Width: 12 inches</td> <td>ROPs: 176</td> </tr> <tr> <td>Height: 137 mm</td> <td>SM Count: 128</td> </tr> <tr> <td>5.4 inches</td> <td>Tensor Cores: 512</td> </tr> <tr> <td>21 Gbps effective</td> <td>RT Cores: 128</td> </tr> <tr> <td>21 Gbps effective</td> <td>L1 Cache: 128 KB (per SM)</td> </tr> <tr> <td>12 inches</td> <td>OpenCL: 3.0</td> </tr> <tr> <td>5.4 inches</td> <td>L2 Cache: 72 MB</td> </tr> <tr> <td>2 inches</td> <td>Vulkan: 1.3</td> </tr> <tr> <td>12 inches</td> <td>CUDA: 8.9</td> </tr> <tr> <td>5.4 inches</td> <td>Shader Model: 6.7</td> </tr> </table> | | Slot Width: Triple-slot | Shading Units: 16384 | Length: 304 mm | TMUs: 512 | Width: 12 inches | ROPs: 176 | Height: 137 mm | SM Count: 128 | 5.4 inches | Tensor Cores: 512 | 21 Gbps effective | RT Cores: 128 | 21 Gbps effective | L1 Cache: 128 KB (per SM) | 12 inches | OpenCL: 3.0 | 5.4 inches | L2 Cache: 72 MB | 2 inches | Vulkan: 1.3 | 12 inches | CUDA: 8.9 | 5.4 inches | Shader Model: 6.7 |
| Slot Width: Triple-slot | Shading Units: 16384 | | | | | | | | | | | | | | | | | | | | | | | | |
| Length: 304 mm | TMUs: 512 | | | | | | | | | | | | | | | | | | | | | | | | |
| Width: 12 inches | ROPs: 176 | | | | | | | | | | | | | | | | | | | | | | | | |
| Height: 137 mm | SM Count: 128 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.4 inches | Tensor Cores: 512 | | | | | | | | | | | | | | | | | | | | | | | | |
| 21 Gbps effective | RT Cores: 128 | | | | | | | | | | | | | | | | | | | | | | | | |
| 21 Gbps effective | L1 Cache: 128 KB (per SM) | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 inches | OpenCL: 3.0 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.4 inches | L2 Cache: 72 MB | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 inches | Vulkan: 1.3 | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 inches | CUDA: 8.9 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.4 inches | Shader Model: 6.7 | | | | | | | | | | | | | | | | | | | | | | | | |
| Clock Speeds | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>Base Clock: 2235 MHz</td> <td>Memory Clock: 1313 MHz</td> </tr> <tr> <td>Boost Clock: 2520 MHz</td> <td>Memory Clock: 1313 MHz</td> </tr> <tr> <td>Memory Clock: 1313 MHz</td> <td>Memory Clock: 1313 MHz</td> </tr> <tr> <td>21 Gbps effective</td> <td>21 Gbps effective</td> </tr> </table> | | Base Clock: 2235 MHz | Memory Clock: 1313 MHz | Boost Clock: 2520 MHz | Memory Clock: 1313 MHz | Memory Clock: 1313 MHz | Memory Clock: 1313 MHz | 21 Gbps effective | 21 Gbps effective | | | | | | | | | | | | | | | | |
| Base Clock: 2235 MHz | Memory Clock: 1313 MHz | | | | | | | | | | | | | | | | | | | | | | | | |
| Boost Clock: 2520 MHz | Memory Clock: 1313 MHz | | | | | | | | | | | | | | | | | | | | | | | | |
| Memory Clock: 1313 MHz | Memory Clock: 1313 MHz | | | | | | | | | | | | | | | | | | | | | | | | |
| 21 Gbps effective | 21 Gbps effective | | | | | | | | | | | | | | | | | | | | | | | | |
| Board Design | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>Slot Width: Triple-slot</td> <td>Shading Units: 16384</td> </tr> <tr> <td>Length: 304 mm</td> <td>TMUs: 512</td> </tr> <tr> <td>Width: 12 inches</td> <td>ROPs: 176</td> </tr> <tr> <td>Height: 137 mm</td> <td>SM Count: 128</td> </tr> <tr> <td>5.4 inches</td> <td>Tensor Cores: 512</td> </tr> <tr> <td>21 Gbps effective</td> <td>RT Cores: 128</td> </tr> <tr> <td>12 inches</td> <td>L1 Cache: 128 KB (per SM)</td> </tr> <tr> <td>5.4 inches</td> <td>OpenCL: 3.0</td> </tr> <tr> <td>2 inches</td> <td>L2 Cache: 72 MB</td> </tr> <tr> <td>12 inches</td> <td>Vulkan: 1.3</td> </tr> <tr> <td>5.4 inches</td> <td>CUDA: 8.9</td> </tr> <tr> <td>2 inches</td> <td>Shader Model: 6.7</td> </tr> </table> | | Slot Width: Triple-slot | Shading Units: 16384 | Length: 304 mm | TMUs: 512 | Width: 12 inches | ROPs: 176 | Height: 137 mm | SM Count: 128 | 5.4 inches | Tensor Cores: 512 | 21 Gbps effective | RT Cores: 128 | 12 inches | L1 Cache: 128 KB (per SM) | 5.4 inches | OpenCL: 3.0 | 2 inches | L2 Cache: 72 MB | 12 inches | Vulkan: 1.3 | 5.4 inches | CUDA: 8.9 | 2 inches | Shader Model: 6.7 |
| Slot Width: Triple-slot | Shading Units: 16384 | | | | | | | | | | | | | | | | | | | | | | | | |
| Length: 304 mm | TMUs: 512 | | | | | | | | | | | | | | | | | | | | | | | | |
| Width: 12 inches | ROPs: 176 | | | | | | | | | | | | | | | | | | | | | | | | |
| Height: 137 mm | SM Count: 128 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.4 inches | Tensor Cores: 512 | | | | | | | | | | | | | | | | | | | | | | | | |
| 21 Gbps effective | RT Cores: 128 | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 inches | L1 Cache: 128 KB (per SM) | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.4 inches | OpenCL: 3.0 | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 inches | L2 Cache: 72 MB | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 inches | Vulkan: 1.3 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.4 inches | CUDA: 8.9 | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 inches | Shader Model: 6.7 | | | | | | | | | | | | | | | | | | | | | | | | |
| Power | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>TDP: 450 W</td> <td>Power Connectors: 1x 16-pin</td> </tr> </table> | | TDP: 450 W | Power Connectors: 1x 16-pin | | | | | | | | | | | | | | | | | | | | | | |
| TDP: 450 W | Power Connectors: 1x 16-pin | | | | | | | | | | | | | | | | | | | | | | | | |
| Suggested PSU | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <tr> <td>850 W</td> <td>Outputs: 1x HDMI 2.1</td> </tr> <tr> <td>850 W</td> <td>1x DisplayPort 1.4a</td> </tr> </table> | | 850 W | Outputs: 1x HDMI 2.1 | 850 W | 1x DisplayPort 1.4a | | | | | | | | | | | | | | | | | | | | |
| 850 W | Outputs: 1x HDMI 2.1 | | | | | | | | | | | | | | | | | | | | | | | | |
| 850 W | 1x DisplayPort 1.4a | | | | | | | | | | | | | | | | | | | | | | | | |
| Connectors | | | | | | | | | | | | | | | | | | | | | | | | | |

Samsung SSD 750 EVO

Data Sheet, Rev.2.1 (October, 2016)

Summary

- SATA 6Gb/s SSD for Client PCs
- 2.5 inch form factor
- Samsung NAND Flash Memory
- Samsung Magician Software for SSD management
- Samsung Data Migration Software

Bring heightened performance, reliability and security to PC storage.

Work worry-free anywhere with solid reliability

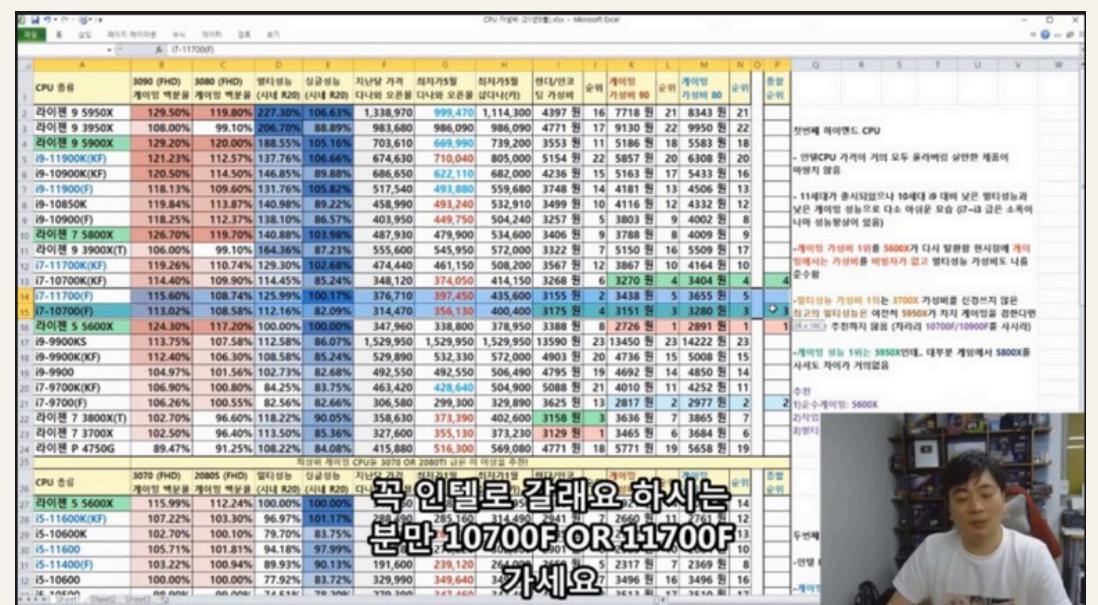
The Samsung 750 EVO boasts a fast random read/write (R/W) performance of up to 97K/88K Input/Output Operations Per Second (IOPS). With TurboWrite, write speeds for the 750 EVO are significantly accelerated during data transfers by creating a high-performance write buffer in the SSD. If a consecutive write operation (i.e. no idle time) exceeds the size of the buffer, the transfer will exit TurboWrite and be processed at "After TurboWrite" speeds. Once the buffer is cleared, TurboWrite performance will resume. However, the buffer size for TurboWrite is more than sufficient for everyday PC use, and you should experience accelerated speeds for most workloads.

With no moving parts, the NAND flash-based 750 EVO provides superb durability and is lightweight, making it the ideal choice for working on the go. It is also

DATA SHEET
Rev. 2.1, Oct., 2016

SAMSUNG

| GeForce RTX 4070 SUPER | | Average G3D Mark |
|--|---------------------------|------------------|
| Bus Interface: PCIe 4.0 x16 | Max Memory Size: 12288 MB | |
| Core Clock(s): 1980 MHz | Memory Clock(s): 1313 MHz | |
| DirectX: 12 | OpenGL: 4.6 | |
| Max TDP: 220 W | | |
| Videocard Category: Desktop | | |
| Other names: NVIDIA GeForce RTX 4070 SUPER | | |
| Videocard First Benchmarked: 2024-01-17 | | |
| G3DMark/Price: 50.13 | | |
| Overall Rank: 8 | | |
| Last Price Change: \$599 USD (2024-05-19) | | |

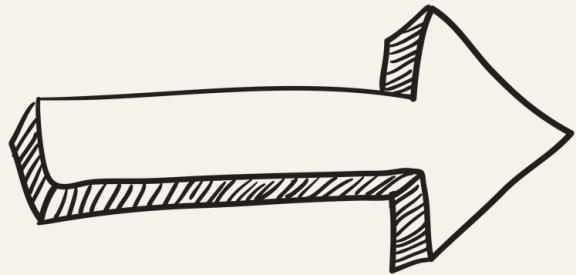
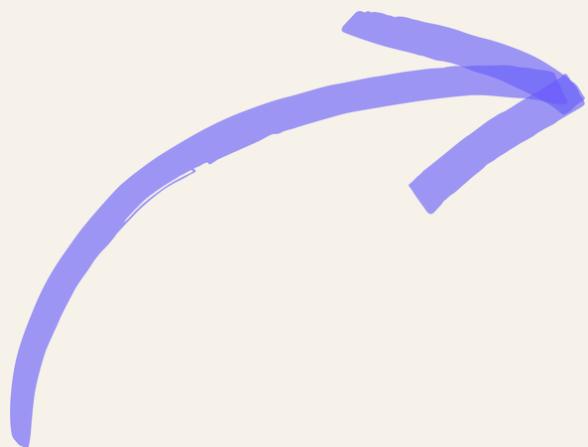
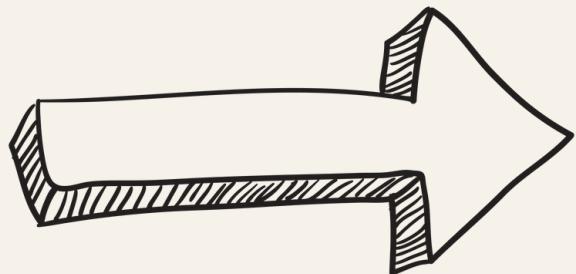


부품 정보를 알아보는 프로그램을 개발하는
techpowerup

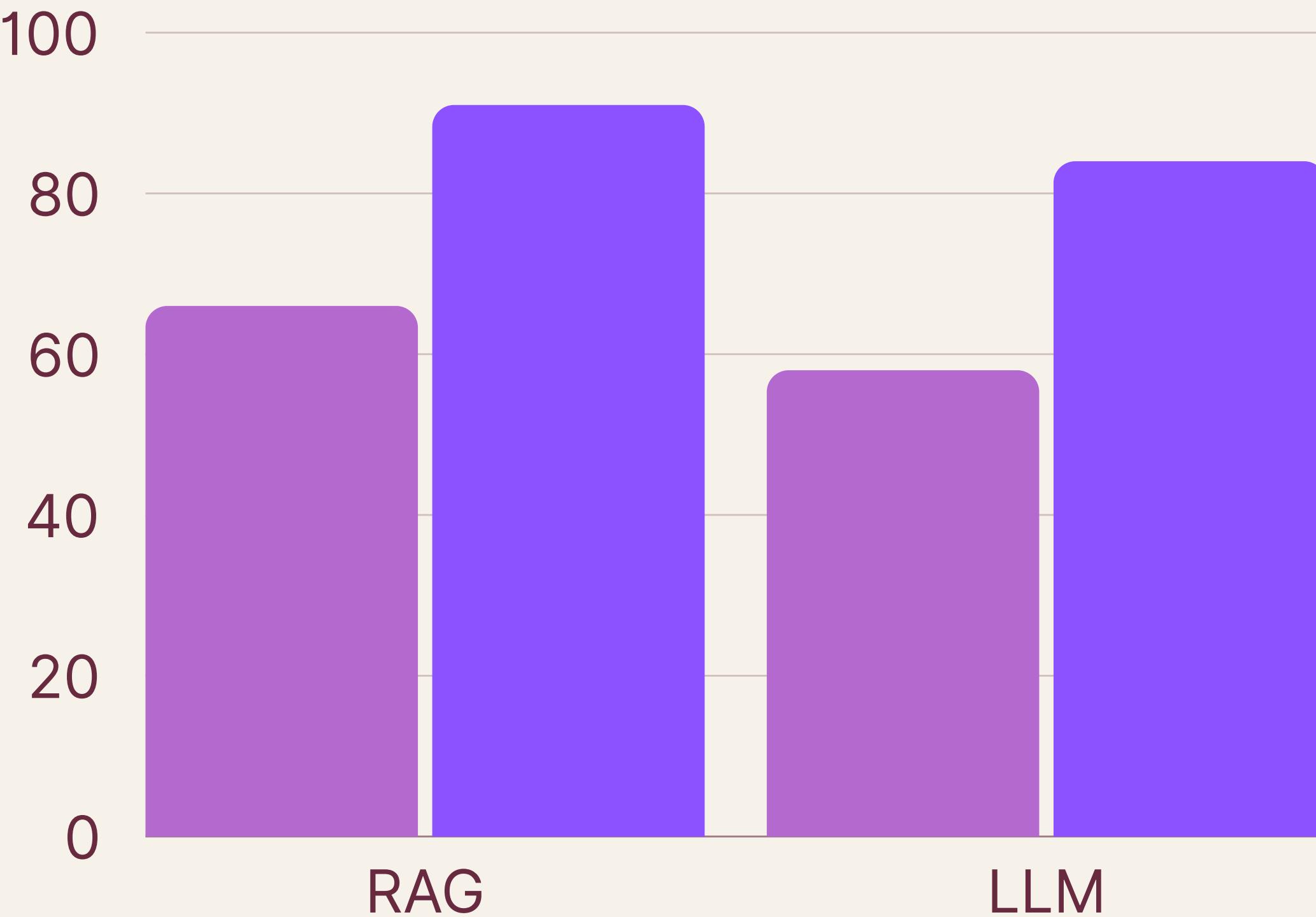
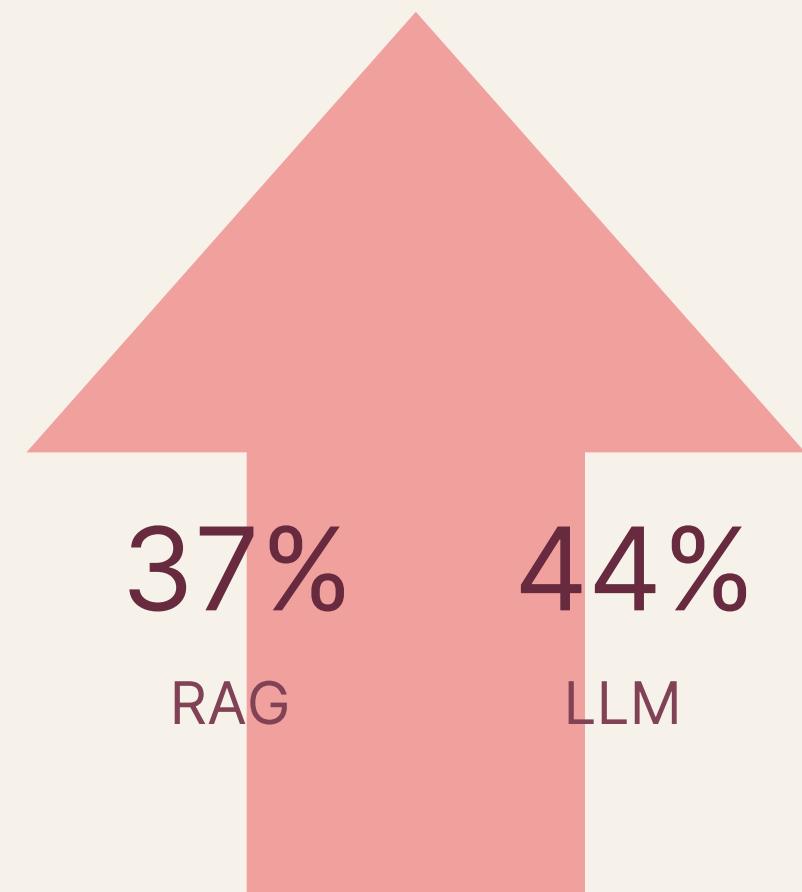
제조사의 제품 설명서

유명 견적사의 견해

— 데이터 정제 & 활용



성능 지표



성능지표 기준

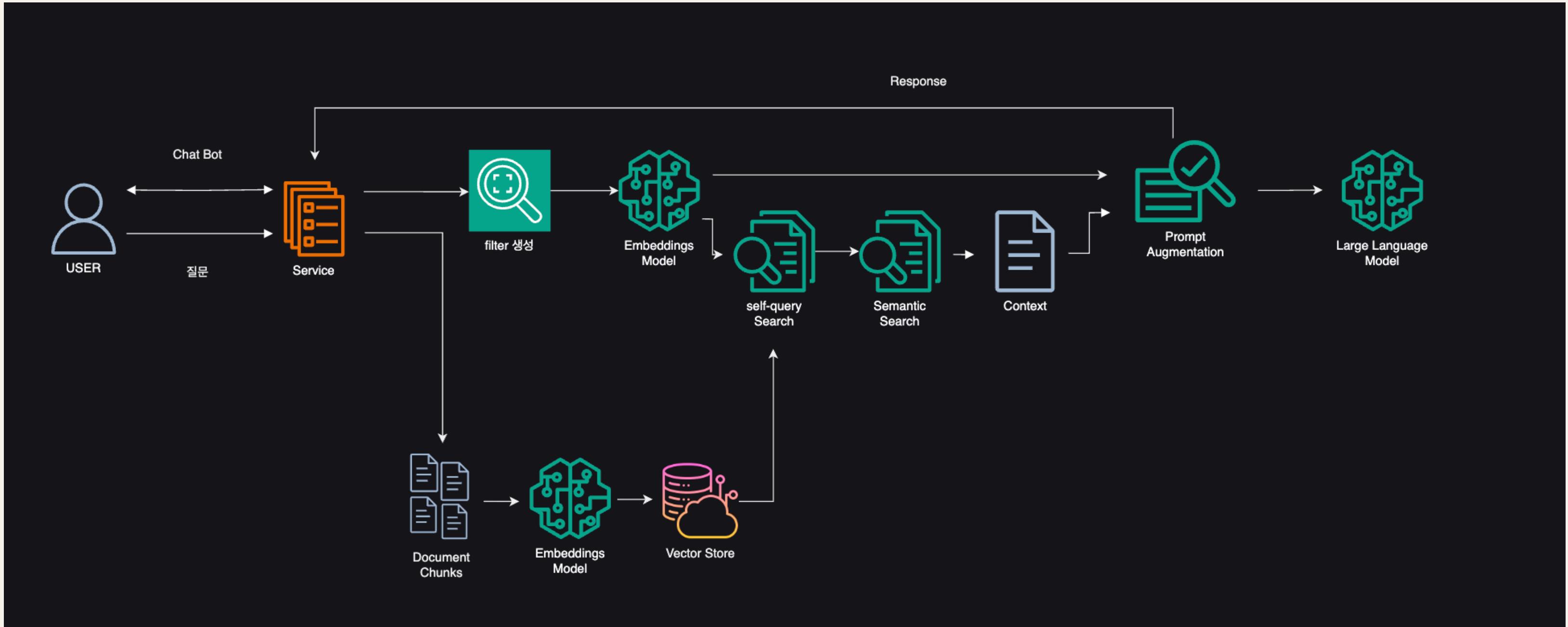
RAG 성능 평가 요소

1. 예산 입력시 총 가격이 예산에 얼추 맞아 떨어지는가
2. 주요 키워드(게임, 방송, 코딩, 영상편집등)에 적합한 견적을 추천해 주는가
3. 특정 부품이 포함되었을 때 해당 부품이 포함된 견적을 추천해 주는가
4. 질문내용과 견적 내용이 논리적으로 연관성이 있는지

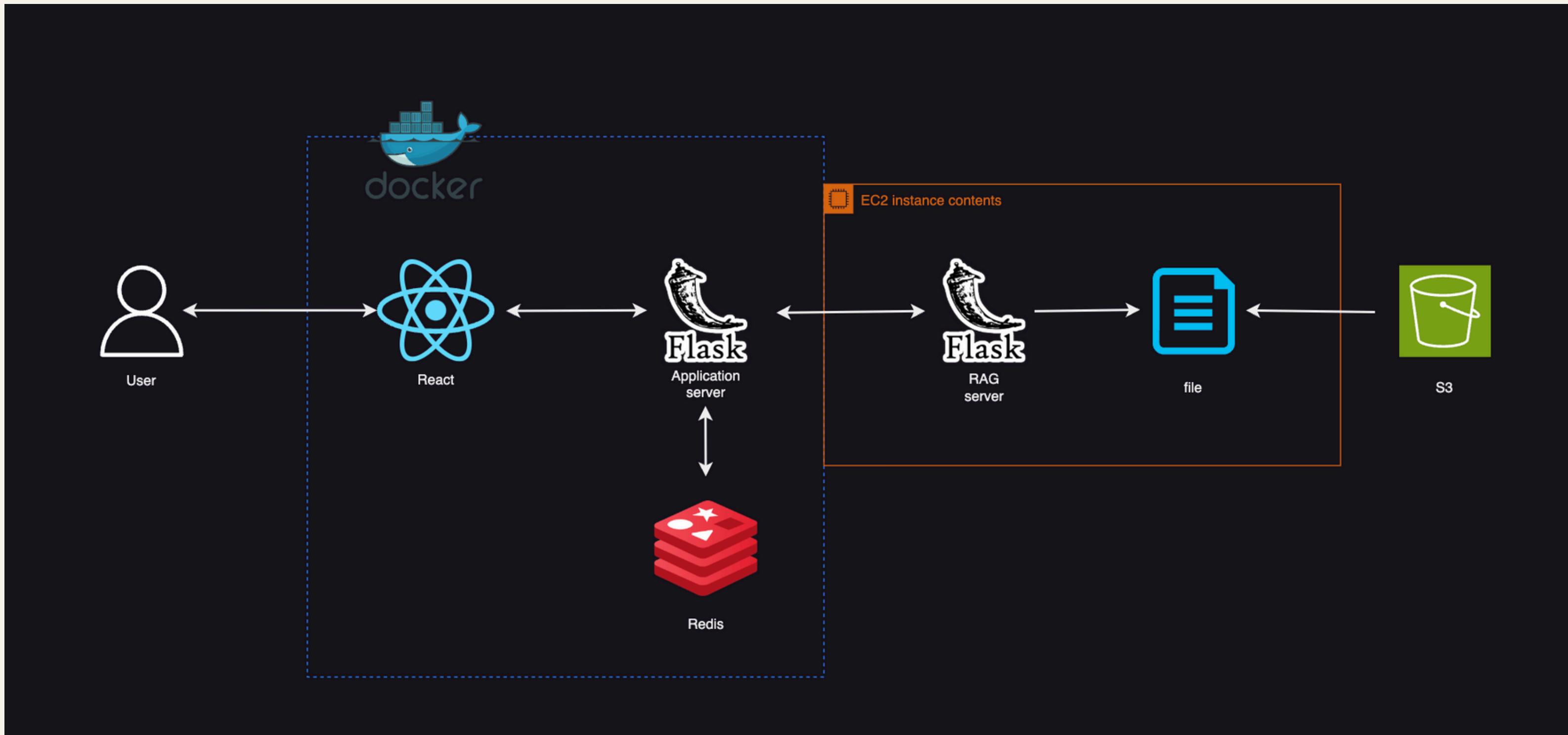
LLM 성능 평가 요소

1. gpt 3.5 대비 부품에 대해서 더 많은 정보를 보유하고 있는지
2. 원천 견적 데이터를 열람하는 것과 LLM모델을 통한 분석 진행한 데이터를 열람하는 것이 유의미한 차이가 있는지

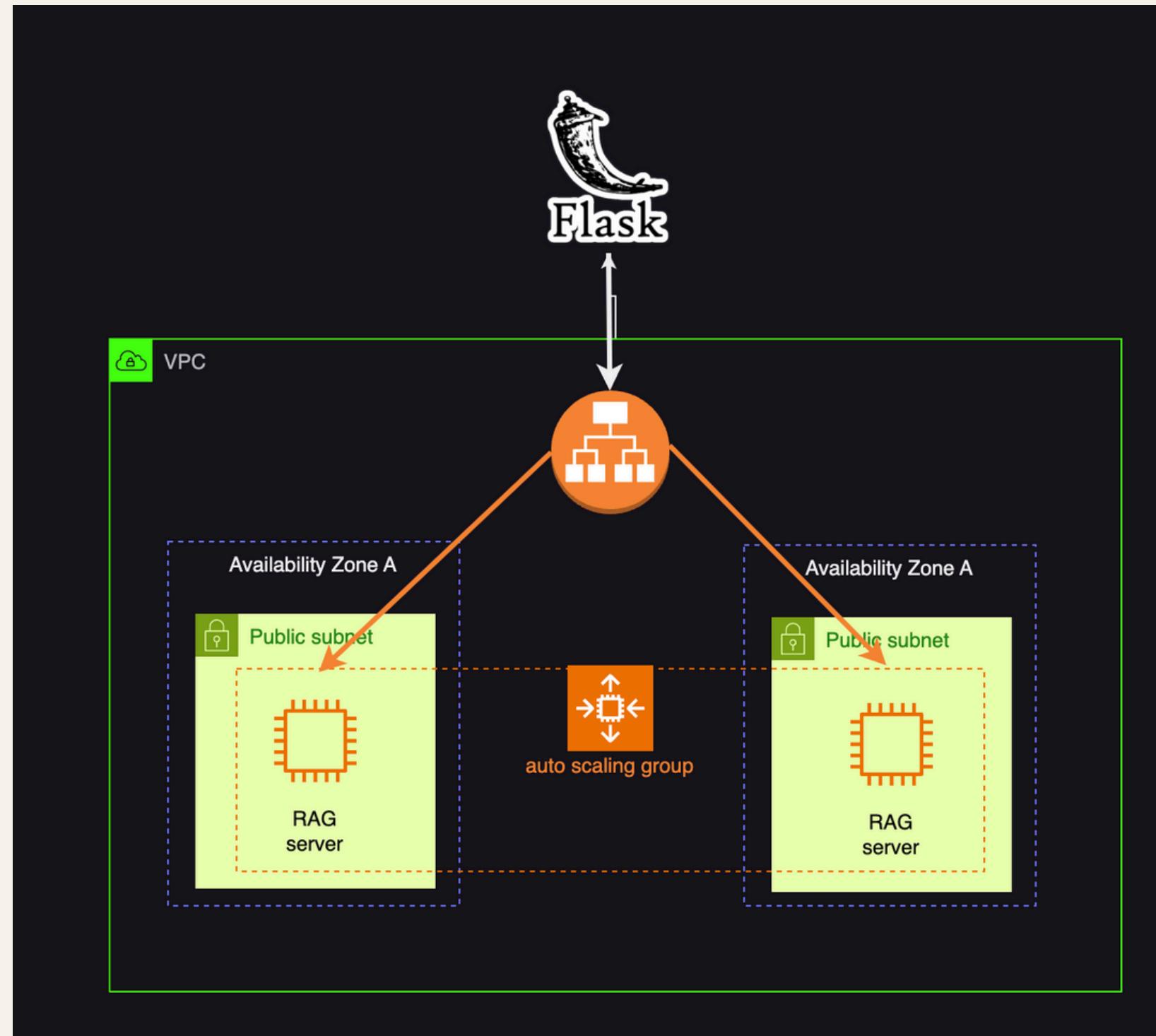
— 시스템 아키텍처 - RAG & LLM



— 시스템 아키텍처 - 웹



— 시스템 아키텍처 - 클라우드



시연

실제 예시로 시연

원하는 컴퓨터 견적에 대해 작성한다.

견적을 추천 받게 되면 견적들끼리의 비교를 하여 어떤 차이가 있는지 확인할수있다.

가격,부품등을 확인하여 마음에 드는 견적을 선택한다.

견적에 대해 자세한 상담을 하게 된다

조건부 통과 내용

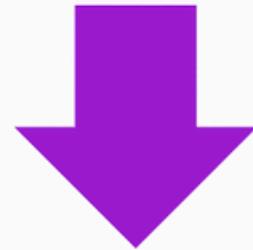
왜 느렸는지

원인

컴퓨터 부품의 상세 정보
LLM을 통해 불러오기



8번의 순차적인 반복



상세하고 신뢰성 있는 정보를
토대로 추천 이유 정리



총 9번의 LLM이 호출

조건부 통과 내용

어떻게 속도개선?

해결위한 시도 : 순차적인 반복을 동시성 기법을 사용하여 부품 상세정보
가져오는 단계 소요
시간 축소

조건부 통과 내용

속도 개선후 평균 시간

비동기 처리

평균 33초

28% 향상

스레드풀

평균 35초

24% 향상

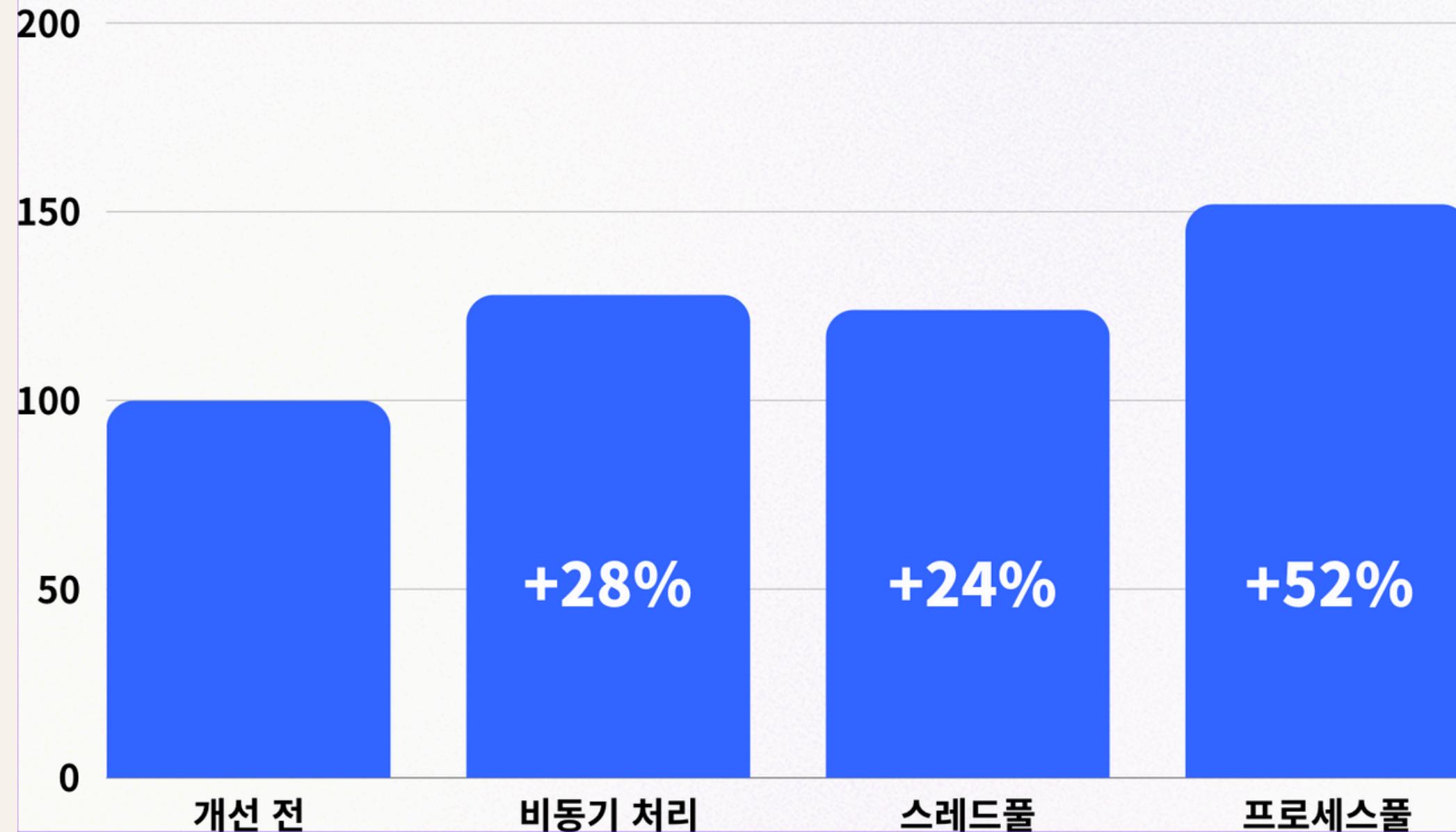
프로세스풀

평균 22초

52% 향상

조건부 통과 내용

속도 개선 향상을



기존 대비 52% 향상된
프로세스풀 체택



조건부 통과 내용

코딩 작업량 3가지

1

데이터의 작업량
rag, LLM데이터 어떻게
정제?

2

LLM작업량

3

RAG작업량



조건부 통과 내용

RAG 데이터

“

견적 사이트: 견적왕, 퀘이사존 등과 같은 신뢰성 있는 견적 사이트에서 크롤링을 통해 5000여 개의 견적 데이터를 수집

“

추가 데이터: 각 부품에 대한 벤치마크 데이터는 PassMark 사이트에서 추출하여 견적 데이터와 매칭, 카테고리를 추가함

”

”

조건부 통과 내용

RAG 작업량

“

견적데이터정제기간: 1주

RAG 구현 기간: 약 3,4주

코드줄: 약 1400줄

”

langchain_core.structured_query.Comparator

class langchain_core.structured_query.Comparator(value)

[source]

Enumerator of the comparison operators.

EQ = 'eq'

NE = 'ne'

GT = 'gt'

GTE = 'gte'

LT = 'lt'

LTE = 'lte'

CONTAIN = 'contain'

LIKE = 'like'

IN = 'in'

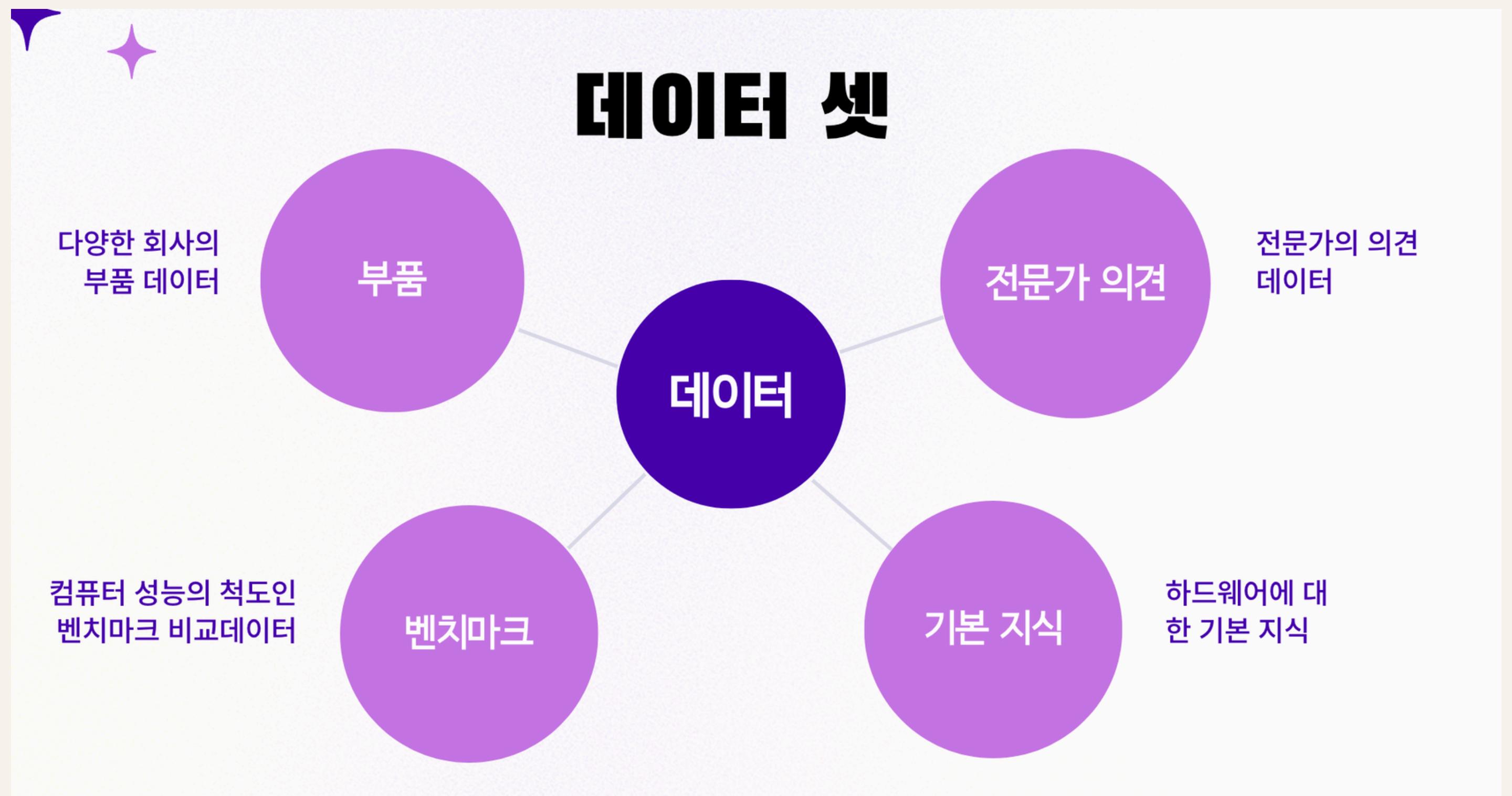
NIN = 'nin'

Examples using Comparator

• constructing-filters.md

여러가지 레퍼런스랑 랭체인 공식문서에서 참고하여 스키마 규칙을 생성 및 프롬프트를 더 정확하게 추가하여 기존 selfquery리트리버를 강화함

조건부 통과 내용



조건부 통과 내용

부품데이터



FEATURES



The latest wireless solution supports MU-MIMO and BSS color technology, delivering speeds up to 2400Mbps.



Simply use a USB key to flash any BIOS within seconds, without installing a CPU, memory or graphics card.



Personalize your PC with 16.8 million colors / 17 effects, controlled in one click!



Better EMI protection and more convenience for installation



Compatible with the latest PC chassis and perfect for external drives and other mobile devices.



Strengthened built-in M.2 thermal solution. Keeps M.2 SSDs safe while preventing throttling, making them run faster.



The latest Gen4 PCI-E and M.2 solution with up to 64GB/s bandwidth for maximum transfer speed.



MSI extended PWM heatsink and enhanced circuit design ensures even high-end processors to run in full speed.



Onboard 2.5G LAN with LAN manager, delivering the best online gaming experience.



An enhanced PCB design improves heat dissipation and performance reliability.

SPECIFICATIONS

| | |
|-------------------|---|
| Model Name | MPG B550 GAMING EDGE WIFI |
| CPU Support | Supports AM4 socket 3rd Gen AMD Ryzen™ Processors, and future AMD Ryzen™ processors with BIOS update |
| CPU Socket | AM4 |
| Chipset | AMD B550 Chipset |
| Graphics | 1x PCI-E 4.0/3.0 x16 slot + 1x PCI-E 3.0 x16 slot |
| Interface | Support 2-way AMD CrossFire™ Technology |
| Display Interface | DisplayPort, HDMI™ - Requires Processor Graphics |
| Memory Support | 4 DIMMs, Dual Channel DDR4-4866+(OC) |
| Expansion Slots | 2x PCI-E 3.0 x1 slots |
| Storage | 1x M.2 Gen4 x4 + 1x M.2 Gen3 x4 slot
6x SATA 6Gb/s ports |
| USB Ports | 2x USB 3.2 Gen 2 10Gbps (1 Type-A + 1 Type-C)
5x USB 3.2 Gen 1 5Gbps (4 Type-A + 1 Type-C)
6x USB 2.0 |
| LAN | Realtek® 8125B 2.5G LAN |
| Wireless / BT | Intel® Wi-Fi 6 module + Bluetooth 5.2 |
| Audio | 8-Channel (7.1) HD Audio with Audio Boost |

| 이름 | 시리즈 | # of CPU Cores | # of Threads | Max. Boost Clock | Base Clock | All Core Boost Speed | L3 Cache | 1kU Pricing | Default TDP | AMD Configurable TDP (cTDP) |
|-----------------|------------------|----------------|--------------|------------------|------------|----------------------|----------|-------------|-------------|-----------------------------|
| AMD EPYC™ 9754S | EPYC 9004 Series | 128 | 128 | Up to 3.1 GHz | 2.25 GHz | 3.1 GHz | 256 MB | 10200 USD | 360W | 320-400W |
| AMD EPYC™ 9754 | EPYC 9004 Series | 128 | 256 | Up to 3.1 GHz | 2.25 GHz | 3.1 GHz | 256 MB | 11900 USD | 360W | 320-400W |
| AMD EPYC™ 9734 | EPYC 9004 Series | 112 | 224 | Up to 3 GHz | 2.2 GHz | 3 GHz | 256 MB | 9600 USD | 340W | 320-400W |
| AMD EPYC™ 9684X | EPYC 9004 Series | 96 | 192 | Up to 3.7 GHz | 2.55 GHz | 3.42 GHz | 1152 MB | 14756 USD | 400W | 320-400W |
| AMD EPYC™ 9654P | EPYC 9004 Series | 96 | 192 | Up to 3.7 GHz | 2.4 GHz | 3.55 GHz | 384 MB | 10625 USD | 360W | 320-400W |
| AMD EPYC™ 9654 | EPYC 9004 Series | 96 | 192 | Up to 3.7 GHz | 2.4 GHz | 3.55 GHz | 384 MB | 11805 USD | 360W | 320-400W |
| AMD EPYC™ 9634 | EPYC 9004 Series | 84 | 168 | Up to 3.7 GHz | 2.25 GHz | 3.1 GHz | 384 MB | 10304 USD | 290W | 240-300W |
| AMD EPYC™ 9554P | EPYC 9004 Series | 64 | 128 | Up to 3.75 GHz | 3.1 GHz | 3.75 GHz | 256 MB | 7104 USD | 360W | 320-400W |
| AMD EPYC™ 9554 | EPYC 9004 Series | 64 | 128 | Up to 3.75 GHz | 3.1 GHz | 3.75 GHz | 256 MB | 9087 USD | 360W | 320-400W |
| AMD EPYC™ 9534 | EPYC 9004 Series | 64 | 128 | Up to 3.7 GHz | 2.45 GHz | 3.55 GHz | 256 MB | 8803 USD | 280W | 240-300W |
| AMD EPYC™ 9474F | EPYC 9004 Series | 48 | 96 | Up to 4.1 GHz | 3.6 GHz | 3.95 GHz | 256 MB | 6780 USD | 360W | 320-400W |
| AMD EPYC™ 9454P | EPYC 9004 Series | 48 | 96 | Up to 3.8 GHz | 2.75 GHz | 3.65 GHz | 256 MB | 4598 USD | 290W | 240-300W |
| AMD EPYC™ 9454 | EPYC 9004 Series | 48 | 96 | Up to 3.8 GHz | 2.75 GHz | 3.65 GHz | 256 MB | 5225 USD | 290W | 240-300W |
| AMD EPYC™ 9384X | EPYC 9004 Series | 32 | 64 | Up to 3.9 GHz | 3.1 GHz | 3.5 GHz | 768 MB | 5529 USD | 320W | 320-400W |
| AMD EPYC™ 9374F | EPYC 9004 Series | 32 | 64 | Up to 4.3 GHz | 3.85 GHz | 4.1 GHz | 256 MB | 4850 USD | 320W | 320-400W |
| AMD EPYC™ 9354P | EPYC 9004 Series | 32 | 64 | Up to 3.8 GHz | 3.25 GHz | 3.75 GHz | 256 MB | 2730 USD | 280W | 240-300W |
| AMD EPYC™ 9354 | EPYC 9004 Series | 32 | 64 | Up to 3.8 GHz | 3.25 GHz | 3.75 GHz | 256 MB | 3420 USD | 280W | 240-300W |
| AMD EPYC™ 9334 | EPYC 9004 Series | 32 | 64 | Up to 3.9 GHz | 2.7 GHz | 3.85 GHz | 128 MB | 2990 USD | 210W | 200-240W |
| AMD EPYC™ 9274F | EPYC 9004 Series | 24 | 48 | Up to 4.3 GHz | 4.05 GHz | 4.1 GHz | 256 MB | 3060 USD | 320W | 320-400W |
| AMD EPYC™ 9254 | EPYC 9004 Series | 24 | 48 | Up to 4.15 GHz | 2.9 GHz | 3.9 GHz | 128 MB | 2299 USD | 200W | 200-240W |
| AMD EPYC™ 9224 | EPYC 9004 Series | 24 | 48 | Up to 3.7 GHz | 2.5 GHz | 3.65 GHz | 64 MB | 1825 USD | 200W | 200-240W |

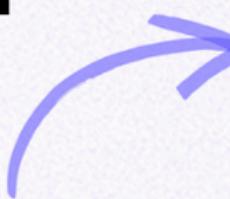
5000여개의 데이터

각 회사별 cpu, gpu, motherboard,
power, ssd, hdd 등등



조건부 통과 내용

벤치마크 데이터



| 벤치마크 데이터 | | |
|---------------------------|-----------------------------------|-----------------------------------|
| | AMD Ryzen 9 7950X | Intel Core i7-13700K |
| Price | \$489.98 - BUY | \$329.98 - BUY |
| Socket Type | AM5 | FCLGA1700 |
| CPU Class | Desktop | Desktop |
| Clockspeed | 4.5 GHz | 3.4 GHz |
| Turbo Speed | Up to 5.7 GHz | Up to 5.4 GHz |
| # of Physical Cores | 16 (Threads: 32) | 16 (Threads: 24) |
| Cache | L1: 1,024KB, L2: 16.0MB, L3: 64MB | L1: 1,408KB, L2: 24.0MB, L3: 30MB |
| TDP | 170W | 125W |
| Yearly Running Cost | \$31.03 | \$22.81 |
| Other | AMD Radeon Graphics | Intel UHD Graphics 770 |
| First Seen on Chart | Q3 2022 | Q4 2022 |
| # of Samples | 4226 | 3961 |
| CPU Value | 128.4 | 141.1 |
| Single Thread Rating | 4286 | 4365 |
| (% diff. to max in group) | (-1.8%) | (0.0%) |
| CPU Mark | 62895 | 46549 |
| (% diff. to max in group) | (0.0%) | (-26.0%) |

"AMD Ryzen™ 3 5125C에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 5125C은 Ryzen 제품군이며, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 5125C는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 16MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 7 PRO 4750U에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 7 PRO 4750U은 Ryzen PRO 제품군이며, Ryzen 5000 시리즈의 일부입니다. Ryzen 7 PRO 4750U는 8 코어 16 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 32MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 7 PRO 4750GE에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 7 PRO 4750GE는 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 7 PRO 4750GE는 8 코어 16 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 32MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 7 PRO 4750G에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 7 PRO 4750G은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 7 PRO 4750G는 8 코어 16 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 32MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4655GE에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4655GE는 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 5 PRO 4655GE는 6 코어 12 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 16MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4655G에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4655G은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 5 PRO 4655G는 6 코어 12 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 16MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4650U에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4650U은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 5 PRO 4650U는 6 코어 12 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 16MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4650GE에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4650GE는 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 5 PRO 4650GE는 6 코어 12 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 16MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4650G에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 5 PRO 4650G은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 5 PRO 4650G는 6 코어 12 스레드로, 3.6GHz의 기본 주파수와 최대 4.5GHz의 터보 주파수를 지원합니다. 16MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4450U에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4450U은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 PRO 4450U는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 8MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4355GE에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4355GE는 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 PRO 4355GE는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 8MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4355G에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4355G은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 PRO 4355G는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 8MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4350G에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4350G은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 PRO 4350G는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 8MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4350GE에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4350GE는 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 PRO 4350GE는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 8MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4350U에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 3 PRO 4350U은 Ryzen PRO 제품군의 일부로, Ryzen 5000 시리즈의 일부입니다. Ryzen 3 PRO 4350U는 4 코어 8 스레드로, 3.6GHz의 기본 주파수와 최대 4.2GHz의 터보 주파수를 지원합니다. 8MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 9 4900HS에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 9 4900HS은 Ryzen 제품군이며, Ryzen 4000 시리즈의 일부입니다. Ryzen 9 4900HS는 8 코어 16 스레드로, 3.8GHz의 기본 주파수와 최대 4.8GHz의 터보 주파수를 지원합니다. 32MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 9 4900H에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 9 4900H은 Ryzen 제품군이며, Ryzen 4000 시리즈의 일부입니다. Ryzen 9 4900H는 8 코어 16 스레드로, 3.8GHz의 기본 주파수와 최대 4.8GHz의 터보 주파수를 지원합니다. 32MB의 L3 캐시를 갖추고 있습니다."}, {"role": "assistant", "content": "AMD Ryzen™ 7 4980U Microsoft Surface® Edition에 대해 설명해줘"}, {"role": "assistant", "content": "AMD Ryzen™ 7 4980U은 Ryzen 제품군이며, Ryzen 4000 시리즈의 일부입니다. Ryzen 7 4980U는 8 코어 16 스레드로, 3.8GHz의 기본 주파수와 최대 4.8GHz의 터보 주파수를 지원합니다. 32MB의 L3 캐시를 갖추고 있습니다."}

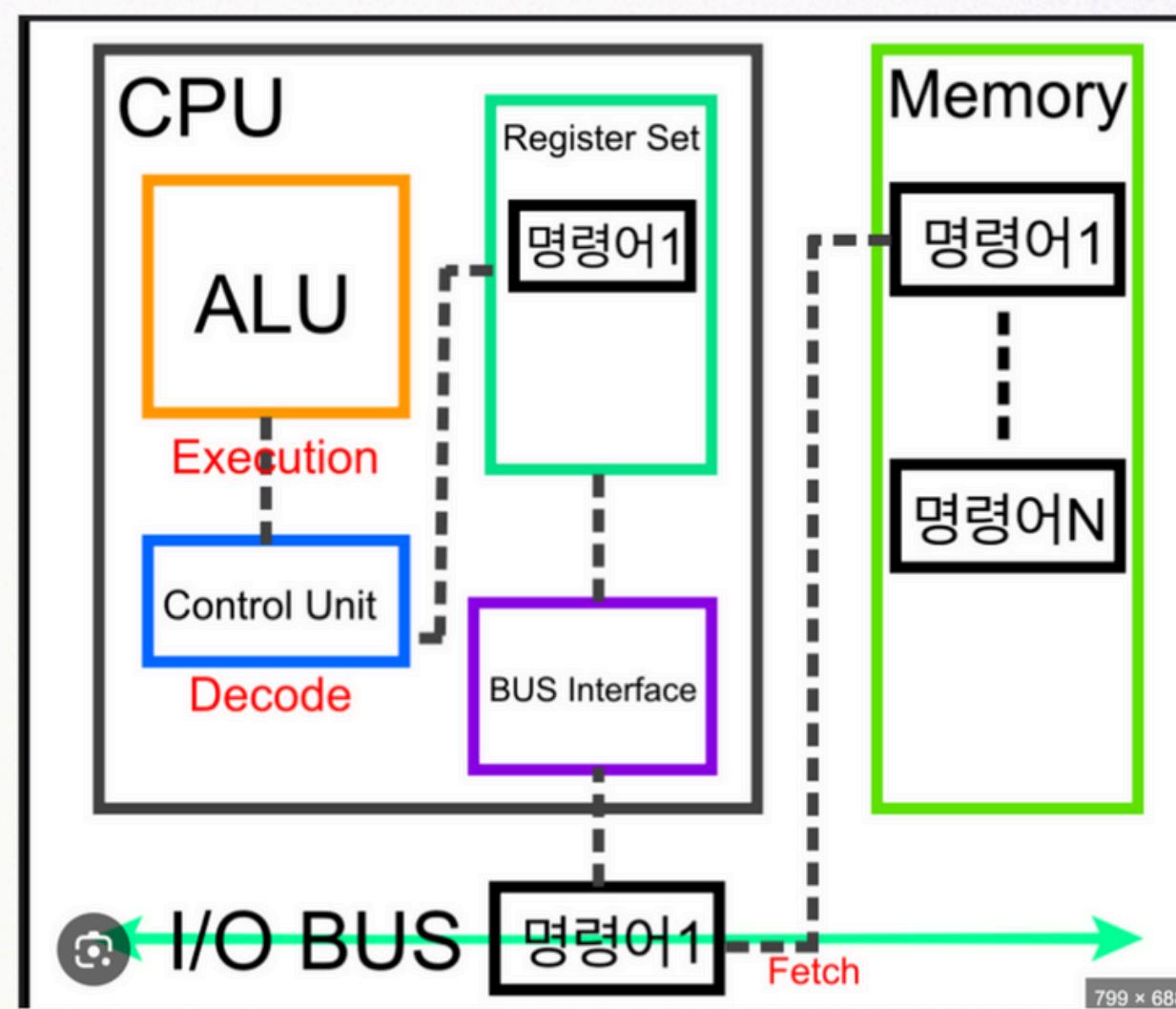
1000여개의 데이터

각 회사별 cpu, gpu, motherboard,
power, ssd, hdd 등등



조건부 통과 내용

컴퓨터 기본 지식



t": "전기로 끌어쓰는 총 전력량은 얼마인가?", {"role": "assistant", "content": "표기된 정격 출력 300W에 효율이 80%일 경우 300 x t": "그래픽 카드의 용량은 어떤 요인에 따라 결정되나요?", {"role": "assistant", "content": "파워와 그래픽 카드에 꽂는 커넥터들은 보 t": "CPU에 대한 전원 공급을 위해 파워 보조전원은 어떻게 구성되어 있나요?", {"role": "assistant", "content": "파워 보조전원은 CPU t": "어떤 경우에 그래픽카드가 225W 이상의 전력을 소모할 수 있는지에 대해 설명해 주실 수 있나요?", {"role": "assistant", "content": "t": "어떤 전원 연결 방식으로 최대 600W까지 공급할 수 있는가요?", {"role": "assistant", "content": "PCIe 전원 2, 16(12+4)핀 t": "파워 팬리스 모드를 사용하는 경우 어떤 이점이 있나요?", {"role": "assistant", "content": "파워 팬리스 모드란 발열이 적은 상황이 t": "파워 모듈러를 선택할 때 '풀 모듈러가 무엇인가요?", {"role": "assistant", "content": "풀 모듈러는 모든 선을 원하는 대로 설치 가능 t": "파워 플랫 케이블이 무엇인가요?", {"role": "assistant", "content": "파워 플랫 케이블은 주로 전원 신호를 전달하는 용도로 사용되며 t": "어떤 이유로 파워 대기전력이 없는 컴퓨터는 전원을 꺼도 전기를 사용하지 않나요?", {"role": "assistant", "content": "파워 대기전력은 t": "파워 서플라이가 PC의 다른 부품들에 전원을 공급하는 역할을 하는 것으로 알고 있는데, 파워 서플라이가 불량품일 경우 어떠한 문제가 발생할 수 t": "파워 선택 시 주의 사항이 무엇인가요?", {"role": "assistant", "content": "값싼 고품질 제품은 존재하기 어렵다. 가격대는 100W정도 t": "어떤 이유로 케이스 크기가 컴퓨터를 놓을 자리에 맞는지 궁금한가요?", {"role": "assistant", "content": "케이스 크기는 컴퓨터를 놓을 자리에 맞는지 궁금한가요?", {"role": "assistant", "content": "그리핀카드를 장착할 때 고려해야 하는 케이스의 최대 그래픽카드 수용 크기는 어떻게 확인할 수 있을까요?", {"role": "assistant", "content": "CPU 쿨러 높이 제한이 165mm 이상일 때 어떤 점에 대해 일반적으로 신경 쓸 필요가 없는가요?", {"role": "assistant", "content": "쿨러 선택에 제약이 따를 수 있는 경우가 발생하는 이유는 무엇인가요?", {"role": "assistant", "content": "케이스 번들 쿨러는 대부분

400여개의 데이터

컴퓨터 아키텍처에 대한 기본적인 이해



조건부 통과 내용

전문가 의견 테이터



1000여개의 데이터

전문가의 유튜브, 블로그, 문서
등을 읽고 질문-응답쌍으로
1000여개로 변환

조건부 통과 내용

RAG서버 온프레미스로 전환

CPU 사용시 문제

RAG서버를 EC2에서 구동시
최소 t2.Large를 사용해야함
(GPU의 부재)

t2.large

패밀리: t2 2 vCPU 8 GiB 메모리 현재 세대: true
온디맨드 RHEL 기본 요금: 0.1752 USD 시간당 온디맨드 S
온디맨드 Linux 기본 요금: 0.1152 USD 시간당 온디맨드 V

비싼 요금

EC2에서 gpu 사용시 시간당 2달러 수준 -> 배보다 배꼽이 큰 격.
오토 스케일링 적용하여 요금 감소 노렸지만 근본적 해결은 불가능

온프레미스로 전환

NVIDIA GPU 탑재된 개인 노트북에 리눅스 설치하여 온프레미스로 전환

| 개인 서버(GPU) | aws |
|-------------|------------|
| 2.797512821 | 2.27060769 |

미세하지만 속도 상승도 있었음
(동일 질문 각 50번 호출 평균)

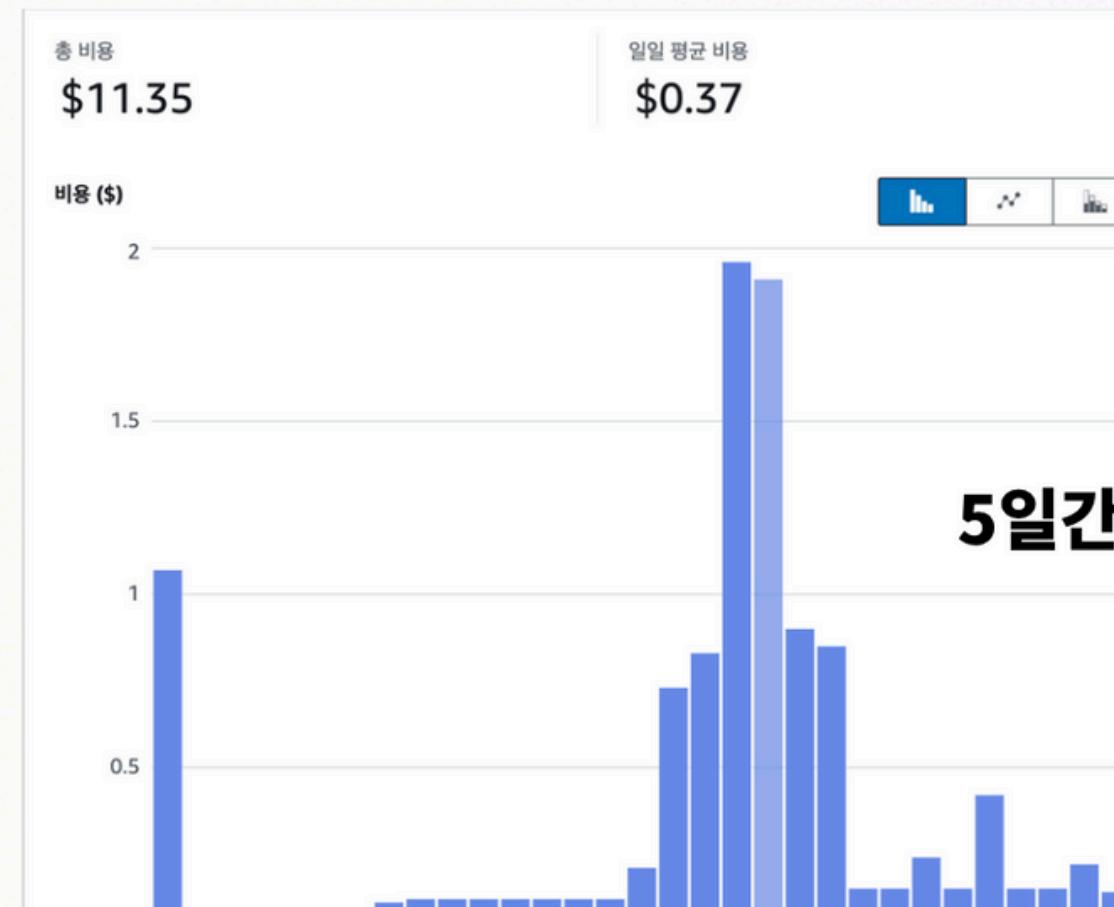


조건부 통과 내용

RAG서버 온프레미스로 전환

가장 큰 이점은 금액 절감

주 사용기간(5일)만에 약8달러 청구



Thank you.

감사합니다