



사용은 쉽게, 결과는 프로처럼

반복되는 업무는 AI에게 맡기고, 더 중요한 일에 집중하세요.
askyour.work가 매일의 일을 더 가볍게 만들어드립니다.

Contact: askyour.work@reboott.ai

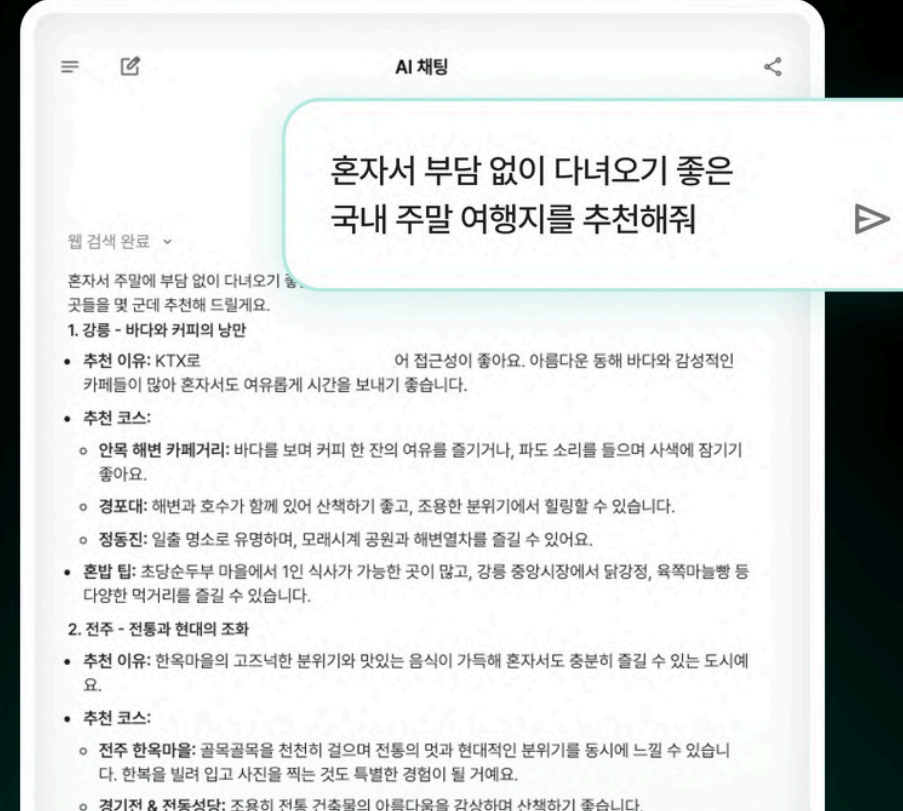
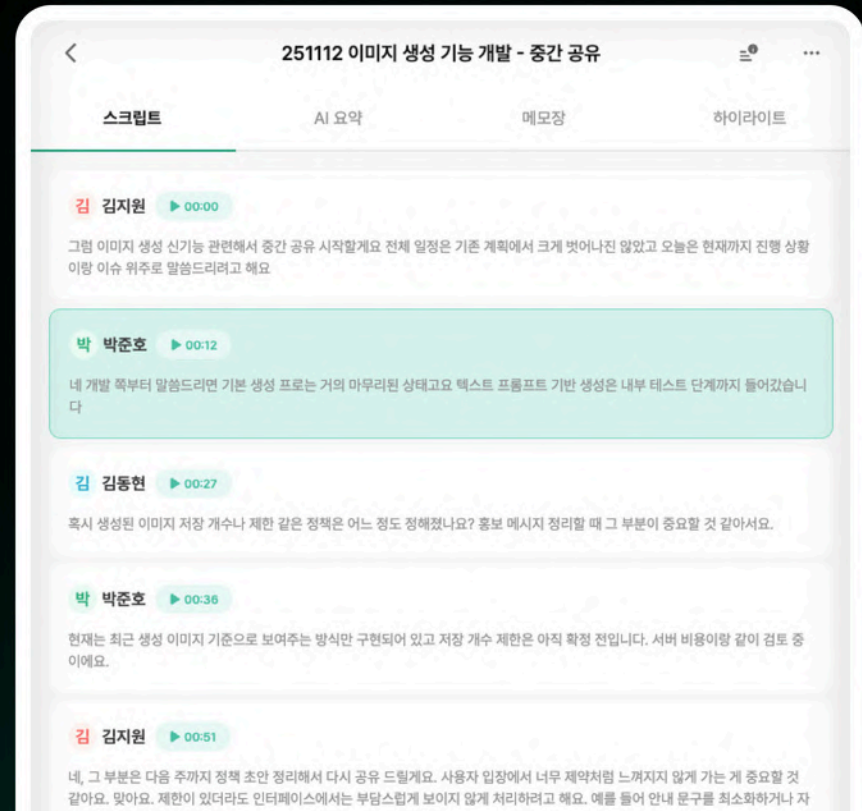




ALL-IN-ONE AI WORKSPACE

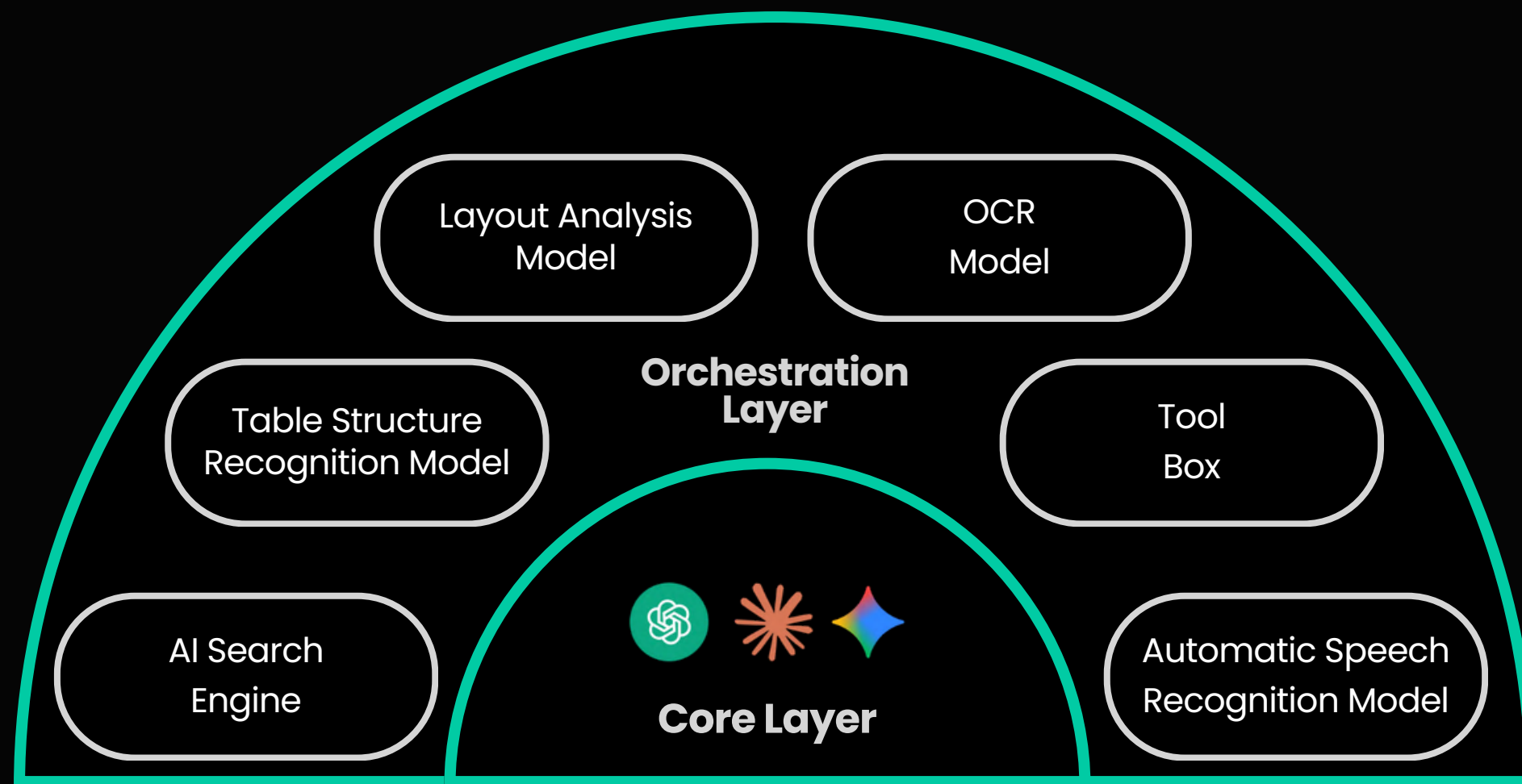
하나의 화면에서, 하나의 흐름으로. [askyour.work](#)는 흩어진 AI 작업을 자연스럽게 이어줍니다. 사내 환경에 맞춘 온프레임 도입부터 기업 맞춤형 AI 개발까지, 유연하게 지원합니다.

- LLM CHAT
- DEEP RESEARCH
- AI IMAGE
- AI SEARCH
- TRANSLATION
- MEETING NOTE
- CLOUD STORAGE
- ENTERPRISE AI



AI ECOSYSTEM

askyour.work의 AI Ecosystem은 자체 orchestration layer 위에서 다양한 AI 기능을 통합하고 연결하는 구조로, 사용 목적에 따라 자연스럽게 이어지는 AI 경험을 제공합니다. 개인은 더 쉽게 활용하고, 기업은 더 유연하게 구축할 수 있어 일상 업무부터 엔터프라이즈 환경까지 폭넓게 대응할 수 있습니다.



Awards

- **1st place** in the Justified Referral in AI Glaucoma Screening on ISBI 2024
- **1st place** in the 2nd scientific figure captioning challenge on IJCAI 2024
- **1st place** in the visually rich form document intelligence and understanding challenge on IJCAI 2024
- **1st place** in the 3rd scientific figure captioning challenge on COLM 2025

askyour.work는 매년 글로벌 AI 학회에서 축적한 연구 성과와 기술 경쟁력을 바탕으로, 개인의 일상 업무부터 기업의 복잡한 운영 환경까지 아우르는 AI Ecosystem을 만들어가고 있습니다.

LLM CHAT

BRING FRONTIER LLMs TOGETHER

프론티어 LLM을 한 플랫폼에 모아, 필요한 순간에 가장 적합한 모델을 바로 선택하고 활용할 수 있습니다.
여러 플랫폼을 오가며 비교하던 번거로움을 줄여, 작업은 더 빠르고 흐름은 더 매끄럽게 이어집니다.

 **GEMINI 3 FLASH**
 **GEMINI 3.1 PRO**

 **CLAUDE SONNET**

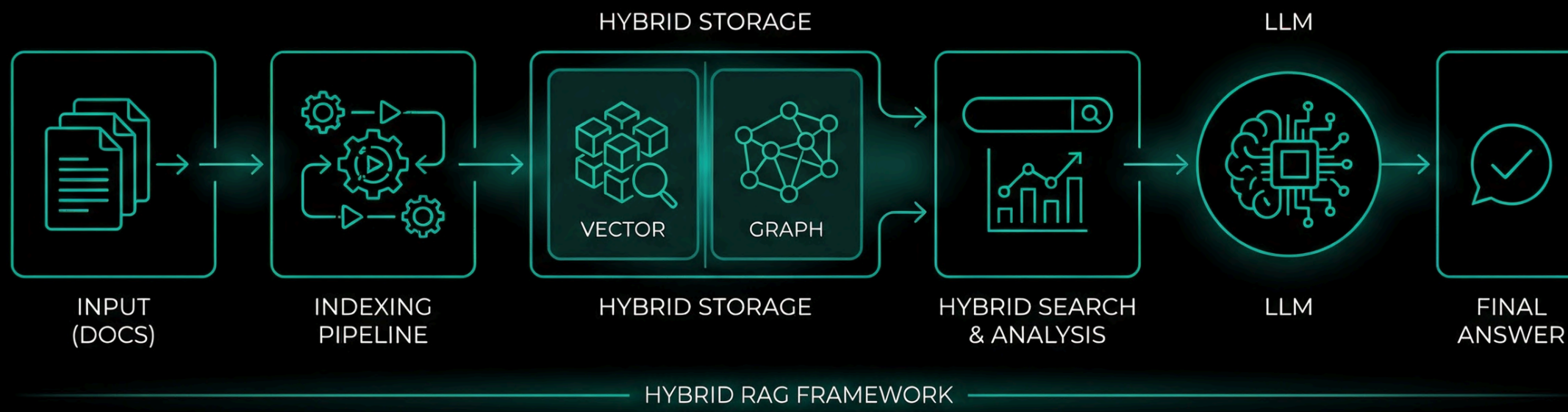
 **GPT-5.4**

 **GPT-5-MINI**

LLM CHAT

DOCUMENT UNDERSTANDING

askyour.work는 문서 청크와 그에 연결된 지식 엔티티와 비주얼 데이터를 함께 검색하여, 더 넓고 정확한 근거를 확보합니다. 이를 통해 단순 유사도 검색을 넘어 관계 기반 맥락까지 반영해 정교하고 신뢰도 높은 답변 생성을 지원합니다.



Multimodal Context

문서, 지식 엔티티, 비주얼 정보를 함께 반영해 답변의 맥락을 더 풍부하고 정확하게 만듭니다.

Visual Context Retrieval

이미지와 도표 등 시각 정보를 함께 활용해 텍스트만으로는 놓치기 쉬운 근거를 보완합니다.

Code Agent

정형 데이터 분석과 코드 실행을 통해 수치 검증과 후속 분석을 더 신속하고 정확하게 지원합니다.

DEEP RESEARCH

REPORT GENERATION

askyour.work의 딥리서치는 문서, 장표, 구조도까지 리서치 전 과정을 단번에 해결합니다. 사용자가 원하는 조건만 설정하면 AI가 복잡한 자료 조사를 자동화하여 통찰력있는 정보를 제공합니다.

Target Research

질문은 짧게, 답변은 넓고 빠르게 돌아옵니다. 기간, 국가, 키워드 등 세부 조건을 설정하여 방대한 데이터 속에서 필요한 핵심 정보에 즉각적으로 도달할 수 있습니다.

Intelligent Document Automation

보고서, 일반 문서, 전문 논문 등 사용자가 원하는 형식에 맞춰 AI가 내용을 정리합니다. 다국어 지원을 통해 글로벌 인사이트까지 놓치지 않고 하나의 문서로 통합합니다.

Integrated Visual Data Generation

단순 텍스트를 넘어 구조도, 아키텍처 이미지, 발표용 장표까지 한번에 완성합니다. 리서치 결과의 가독성을 높이고 내부 보고 및 발표 시 시각적 설득력을 극대화합니다.

전략 1. 시스템 레벨 접근: 계층적 관리와 OS 메타포

ACL 2025-2026 Focus
단순 저장을 넘어선 시스템적 자원 관리

HiAgent: 계층적 작업 메모리

하위 목표(Subgoal) 기반 정보 청킹
현재 작업과 관련된 상세 정보만 유지하고, 완료된 하위 목표는 요약하여 '능동적 망각' 수행.

장기 작업 벤치마크 성과

1.0x (Baseline) vs 2.0x (성공률)

기존 방식 vs HiAgent

평균 3.8 Step 단축

MemOS: 메모리 운영체제

메모리를 OS의 1급 자원(First-class)으로 격상

MEMORY HIERARCHY

Parametric
내재화된 장기 지식 (Weights)

Agentic Memory (RL)

강화학습 기반 능동적 제어
에이전트가 도구(Tool)를 사용하여 스스로 무엇을 기억하고 망각할지 결정.

구분: 메모리 제어 도구 (Tools)

LTM
자기 기억

ADD UPDATE DELETE

3. 전략 2. 아키텍처 접근: 뉴럴 메모리와 인지 모델링

외부 저장소를 넘어선 내부 기억의 내재화(Internalized Memory) 및 인지 과학 기반 아키텍처 혁신

3.1 Google Titans

테스트 타임 암기 (Test-time Memorization)

심층 신경망(MLP) 자체를 메모리 모듈로 활용. '놀라움(Surprise)' 지표가 높은 정보를 실시간으로 영구 기억화하여 학습.

Surprise Metric
예측 오차가 클수록(High Gradient) 중요 정보 판단 → 장기 기억 저장

초장기 문맥(2M+) 추론 성능 (BABILONG)

65점 vs 95점

기존 Transformer vs Google Titans

3.2 HippoRAG

해마 모방 인덱싱 (Hippocampal Indexing)

LLM(대뇌 피질)과 지식 그래프(해마)의 상호작용.
PageRank 알고리즘으로 숨겨진 연결 고리를 찾아내는 '패턴 완성'.

Multi-hop 추론 +20%

비용 절감 1/20

검색 속도 효율성 비교 (VS IRCOT)

IRCOT (반복 검색) 기준

HippoRAG (단일) 13배 빠름

3.3 에피소드 메모리

인지 구조 (Cognitive Architecture)

단순 정보 검색을 넘어, 경험을 시간 순으로 구조화하여 자기 진화(Self-evolution)하는 인공형 에이전트 구현.

구분	기존 RAG / LLM	인지 메모리 모델
상태	Stateless (휘발성)	Stateful (지속성)
지역 방식	벡터/외부 DB 의존	내재화 & 에피소드화
적용성	재학습 필요	실시간 경험 학습

환각(Hallucination) 최소화

Key Takeaway

2025-2026년 AI 메모리 전략은 외부 저장소 의존을 탈피하여, Titans와 같은 신경망 자체의 기억 내재화 및 HippoRAG와 같은 인간 뇌 모방 인덱싱으로 진화하고 있습니다. 이는 단순 정보 처리를 넘어 경험을 통해 스스로 학습하고 추론하는 자기 진화형 인지 시스템의 토대가 됩니다.

AI SEARCH

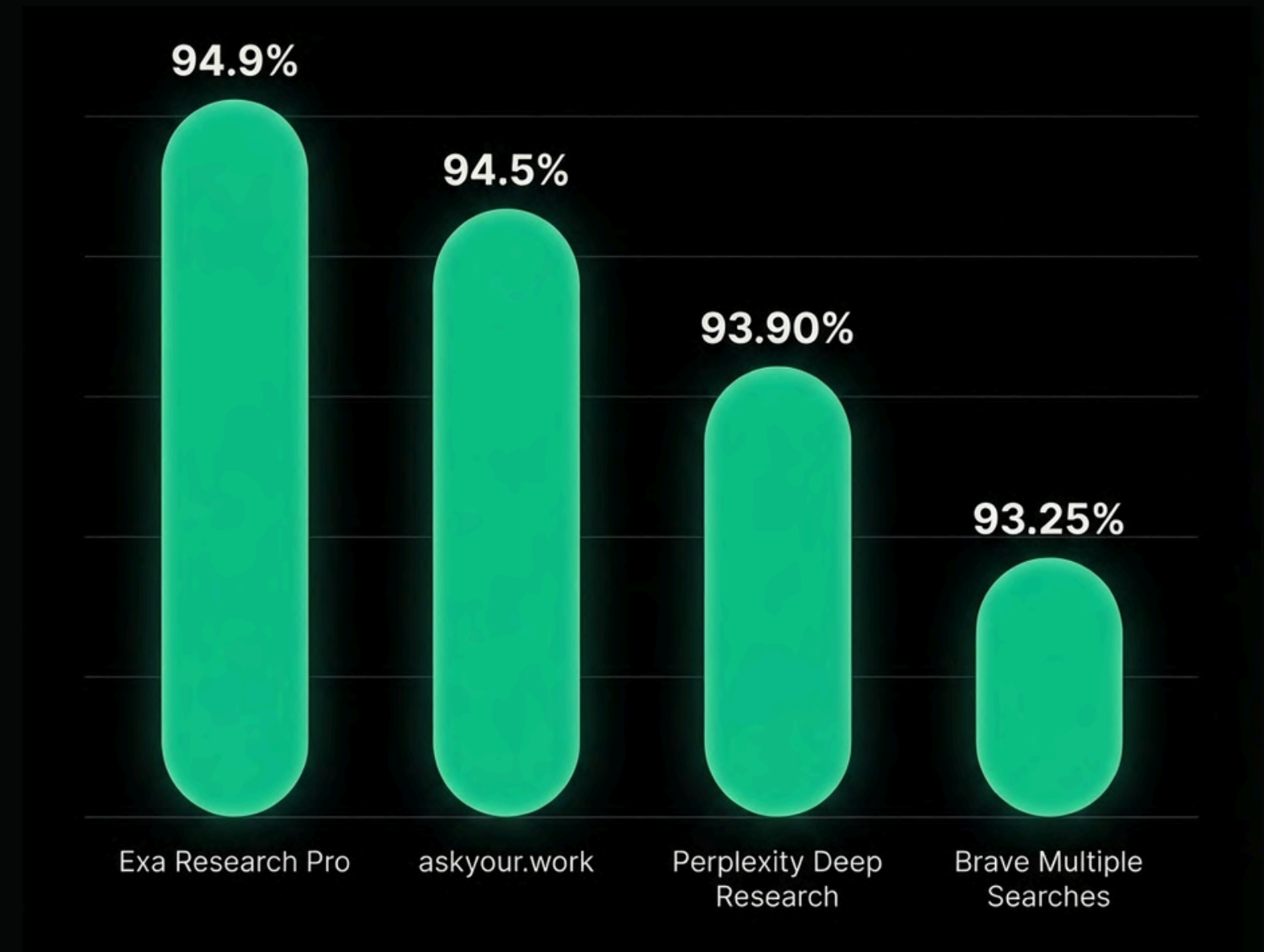
FAST AI SEARCH

askyour.work는 AI 검색을 단순한 "검색 + 생성"을 넘어, 지능형 질의 → 정보 수집 → 구조적 추론 → 생성으로 구조화된 프레임워크입니다. 이를 통해 단순 요약이 아닌, 맥락과 구조를 갖춘 인사이트를 생성합니다. 출처가 명확한 데이터를 바탕으로 5초 이내에 정확하고 신뢰도 높은 인사이트를 제공합니다.

요즘 엔비디아 주가는 어때?



4월달에 열리는 벚꽃축제 행사 일정 정리해줘



SIMPLE QA



TRANSLATION

TRANSLATION PARAPHRASE

단어가 아니라, 의도를 번역합니다. 문맥과 말투를 감지해 자연스러운 표현으로 바꿔줍니다.

Context-Awareness

논문, 일상 대화, 비즈니스 문서까지 문맥에 맞춘 번역 의역으로, 더 자연스럽게 정확한 결과를 도출합니다.

Document Translation

문서의 레이아웃을 유지한채 고품질의 번역본을 제공합니다.

arXiv:1810.04805v2 [cs.CL] 24 May 2019

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Abstract

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers. BERT alleviates the previously mentioned unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of

BERT: 언어 이해를 위한 심층 양

제이콥 데블린 밉-웨이 창
Google AI Language
{jacobdevlin, mingweichang, kentonl, kristout}@google.com

초록

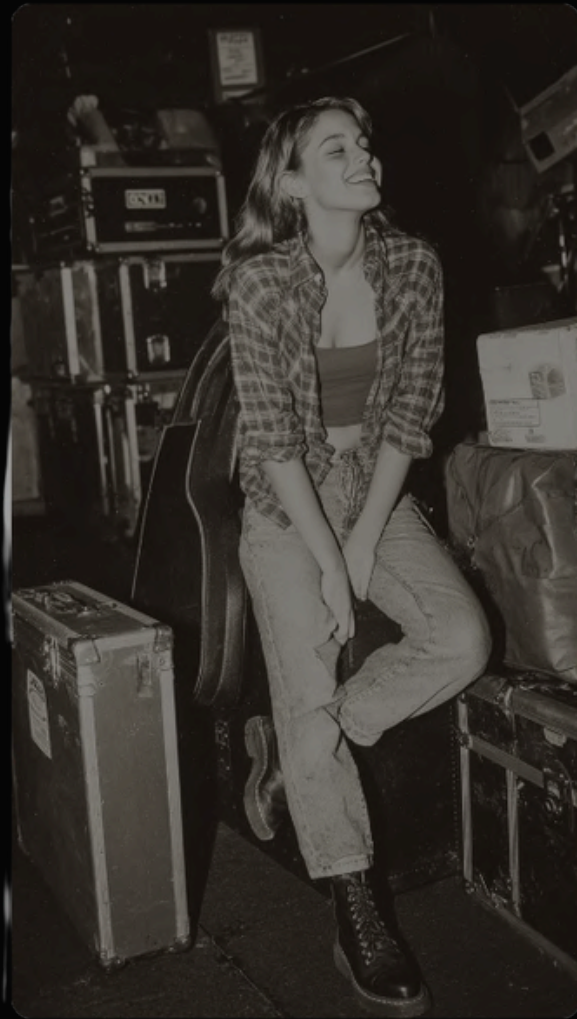
우리는 BERT(Bidirectional Encoder Representations from Transformers)라고 불리는 새로운 언어 표현 모델을 소개합니다. 최근의 언어 표현 모델(Peters et al., 2018a; Radford et al., 2018)과 달리, BERT는 모든 레이어에서 좌우 문맥을 동시에 고려하여 라벨이 없는 텍스트로부터 심층 양방향 표현을 사전 훈련하도록 설계되었습니다. 그 결과, 사전 훈련된 BERT 모델은 출력 레이어 하나만 추가하여 미세 조정함으로써, 작업별 아키텍처를 크게 수정하지 않고도 질문 응답 및 언어 추론과 같은 광범위한 작업에 대한 최첨단 모델을 생성할 수 있다.

BERT는 개념적으로는 단순하지만 실용적으로는 매우 강력합니다. 이 모델은 11가지 자연어 처리 과제에서 새로운 최첨단 성과를 달성했으며, 특히 GLUE 점수를 80.5%(7.7% 포인트 절대적 향상)로 끌어올렸고, MultiNLI 정확도를 86.7%(4.6% 절대적 향상), SQuAD v1.1 질문-응답 테스트 F1 점수를 93.2(1.5% 절대적 향상), SQuAD v2.0 테스트 F1 점수를 83.1(5.1% 절대적 향상)으로 끌어올리는 등 11개의 자연어 처리 과제에서 새로운 최첨단 성과를 달성했습니다.

1 서론

언어 모델 사전 훈련은 많은 자연어 처리 과제(Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018)의 성능을 향상시키는 데 효과적인 것으로 입증되었다. 여기에는 문장을 전체적으로 분석하여 문장 간의 관계를 예측하는 자연어 추론(Bowman et al., 2015; Williams et al., 2018) 및 의역(Dolan and Brockett, 2005)과 같은 문장 수준 작업이 포함되며, 이는 문장을 전체적으로 분석하여 문장 간의 관계를 예측하는 것을 목표로 한다. 또한 명명된 개체 인식 및 질문 응답과 같은 토큰 수준 작업도 포함되는데, 이 경우 모델은 토큰 수준에서 세밀한 출력을 생성해야 한다(Tjong Kim Sang and De Meulder, 2003; Rajpurkar 외, 2016) 등이 있다.

arXiv:1810.04805v2 [cs.CL] 2019년 5월 24일



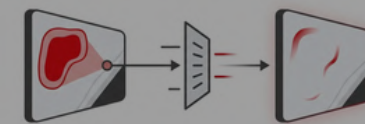
AI IMAGE

IMAGE GENERATION

다양한 스타일과 목적에 맞춰 이미지를 빠르게 생성하고, 간단한 수정부터 세밀한 편집까지 한 번에 처리합니다.

Key Feature Replacement of In-Distribution Samples for Out-of-Distribution Detection (KIRBY)

KIRBY는 외부 데이터 없이 ID 데이터의 핵심 특징을 인페인팅으로 제거하여 효과적인 OOD 데이터를 생성하는 새로운 방법론을 제시합니다.



저자: 김재영 외 (VUNO, UIUC, 성균관대학교)
AAAI 2023 발표 논문

The Proem: The Imperative of Reliable Rejection

Reliable OOD detection is crucial for safety-critical AI, addressing DNN overconfidence and the challenge of infinite OOD variations.

Deep Neural Networks (DNNs) exhibit overconfidence in predictions on out-of-distribution (OOD) samples.

The goal of OOD detection is to reliably reject inputs that deviate from the training distribution.

A key challenge lies in the immense, unrepresentable diversity of potential OOD samples.

METHODOLOGY: INTRODUCING THE KIRBY PIPELINE

KIRBY generates realistic surrogate OOD data by replacing class-specific features while preserving background context, effectively creating "background-only" samples.



KIRBY Methodology: Selective Inpainting

KIRBY uses saliency and inpainting to create semantically plausible background outliers.



ENTERPRISE AI

비정형 문서 데이터 구조화

PDF, 이미지, 표 등이 포함된 복잡한 문서에서 필요한 필드 값을 정확히 추출하여 DB화합니다.

AI 전표 및 증빙 처리

메일이나 첨부파일 내 인보이스, 영수증 등을 인식하여 기업 양식에 맞게 전표를 자동 작성합니다.

사내 맞춤형 문서 분류

외부/내부에서 유입되는 다양한 요청 문서를 AI가 자동 분류하고 담당 부서에 전달 가능한 형태로 요약합니다.

RAG(검색 증강 생성) 구축

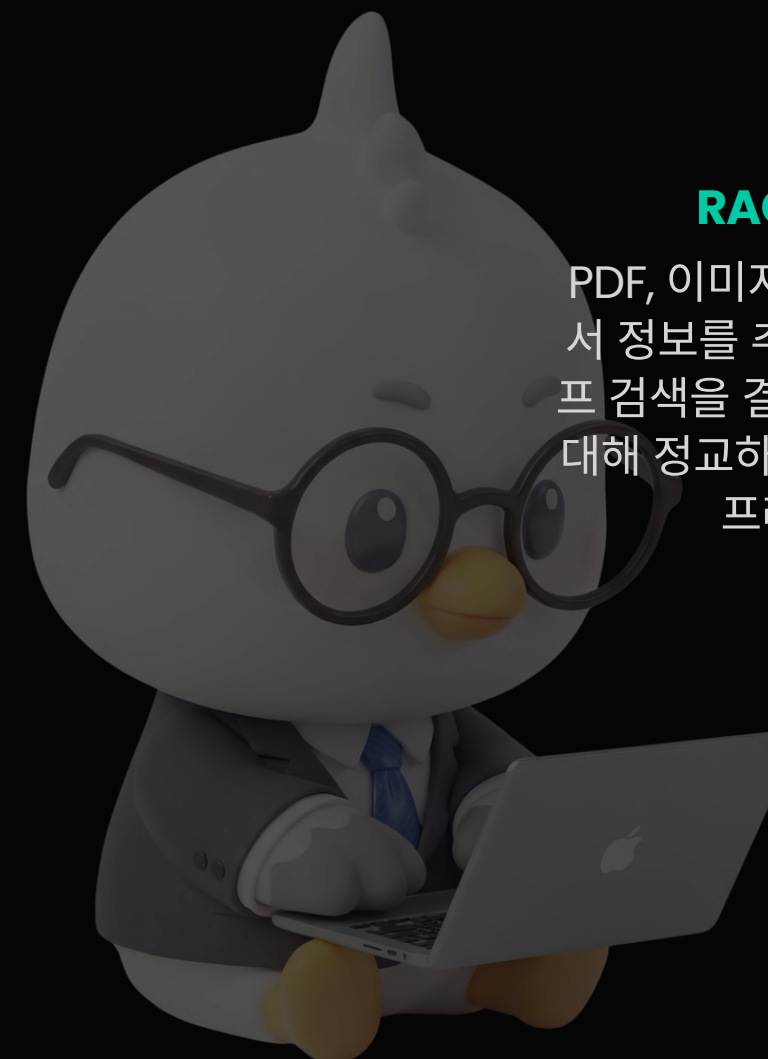
PDF, 이미지, 표 등이 포함된 복잡한 문서에서 정보를 추출. 벡터 검색과 관계 기반 그래프 검색을 결합하여, 사내 규정이나 매뉴얼에 대해 정교하고 신뢰도 높은 답변을 생성하는 프레임워크를 개발합니다.

온프레미스(On-premise) 도입

보안이 중요한 제조·산업 환경을 위해 폐쇄망(내부망) 환경에서도 동작하는 AI 아키텍처를 지원합니다.

제조 공정 문서 자동화

품질 검사서, 작업 지시서 등 제조 현장의 반복적인 문서 업무를 AI로 자동화하여 공정 효율을 높입니다.



ASK YOUR.

WORK



askyour.work@reboott.ai