

# 코난 LLM: 한국어 대규모 언어 모델

코난테크놀로지 | 양승현\*·도원철·오창민·김정태

## 1. 서론

대규모 언어 모델(LLM)이 거대한 기술 트렌드로 자리 잡으며 수많은 상용이나 오픈소스 LLM이 발표되고 있고, 이를 현장에서 활용하려는 노력이 활발히 일어나고 있다. 이 과정에, 기업 내부 데이터의 외부 유출 우려, 학습과 추론 위한 GPU 구입으로 인한 비용 과다 문제, 모델의 한계로 인한 환각 문제 등 LLM의 자유로운 활용을 저해할 수 있는 여러가지 문제가 발생하고 있다. 이런 문제의 해결을 위해 (주)코난테크놀로지는 기업 내부에 설치-운영해 데이터 유출 우려를 원천 방지할 수 있고, 적절한 비용과 최적의 성능으로 효율적인 운영이 가능하며, 벡터 검색을 통한 답변의 증강과 근거 제시가 가능한 중형 LLM을 개발하기 위해, 파라미터 크기는 크지 않게, 토큰은 많이 학습하는 접근법을 택하였다. 더 작은 파라미터 크기 일지라도 더 많은 데이터로 학습한 모델이 더 좋은 품질을 보여 주는 것을 실험적으로 보여주는 결과가 발표되었다. 300B 토큰으로 학습한 Gopher 280B 모델에 비해 크기는 4배 작지만 1.4T 토큰으로 학습한 Chinchilla 67B가 MMLU 등 벤치마크에서 더 우수한 성능을 보였다 [1]. 파라미터 크기를 줄이면 학습은 물론 추론 성능에서도 이득이 있기 때문에 최근에 발표되는 많은 LLM은 LLaMa2의 7~70B, Dolly 2.0 12B, Vicuna 13B, Falcon 40B 등처럼 수십B 정도의 중형 크기에 많이 집중되어 있다. 코난 LLM도 13B, 41B 파라미터로 중형 크기의 두 가지 모델을 제공한다. 파라미터 크기는 줄인 대신 학습 데이터셋의 크기는 크게 늘려서 각각 5천억, 7천억개가 넘는 토큰으로 학습한 모델을 제공하고 있다. 본고에서는 위와 같은 목표로 개발된 코난 LLM의 개발에 대해 서술하겠다.

## 2. 학습 컴퓨팅 인프라

LLM을 사전학습하고 여러 가지 태스크에 대한 미

세조정을 하기 위해서는 강력한 컴퓨팅 자원이 필요하다. 코난 LLM을 사전 학습시키기 위해 인피니밴드(InfiniBand)로 묶인 H100×8 서버 8대를 구입해 사용했다 [2, 3, 4].

사전학습 기간은 가용 컴퓨팅 자원과 모델 크기 그리고 사전학습 데이터셋 크기에 따라 달라진다. 코난 LLM 13B의 경우, H100 서버 8대로 23일이 소요됐다. 이때 전체 사전학습 연산량은  $5.93 \times 10^{22}$  flops로 추산되며, 같은 크기의 GPT-3 13B의 연산량  $2.31 \times 10^{22}$  flops [5]와 비교해서 약 2.5배 이상 더 많은 연산량으로 추산된다.

모델이 커질수록 학습 시 네트워크 비용도 커진다. 20Gbps 네트워크에서는 1스텝 학습 시 모델 파라미터 데이터 전송으로 34초 정도의 오버헤드가 생겼다. 인피니밴드(InfiniBand)를 연결한 이후에는 이 오버헤드가 1초로 줄어서 학습 기간을 대폭 줄일 수 있었다.

## 3. 사전학습

LLM의 성능을 결정짓는 중요한 요인 중 하나가 사전학습(pre-training) 데이터이다. 사전학습 데이터의 양과 품질 모두 중요하다. 코난 LLM의 사전학습 데이터는 AIHUB, 위키피디아 등 공개 데이터와 뉴스, Github 등 구입 데이터를 포함하고 있다. Books3, Gutenberg 등 영문 데이터도 포함하고 있으며, 품질 좋은 문서의 학습을 위해 표 정보를 온전히 추출할 수 있는 알고리즘을 적용했다.

학습 데이터의 양뿐만 아니라 품질 또한 모델의 차이를 만든다. 인터넷 댓글, 스몰토크, 부적절한 내용의 글, 개인 정보, 중복 텍스트 등은 사전학습 데이터셋에서 미리 제거했다. 이를 통해 부적절한 질문에 대한 답변을 요청받았을 경우에도 코난 LLM은 비교적 유해성이 높지 않은 답변을 생성한다.

코난LLM 13B의 사전학습 데이터 토큰 수는 5,395억 개로 유사한 파라미터 크기의 오픈 LLM인 Polyglot-ko 12.8B [6]와 비교하면 3배 이상의 학습 데이터를 가지

\* 종신회원

고 사전학습이 되었다. 한국어 학습 토큰만 보면 3,318억개로 ChatGPT의 한국어 학습 토큰 5.7억개, LLaMA-2의 12억개와도 비교할 수 없이 많은 양이다. 이러한 사전학습 데이터의 차이로 사전학습 LLM의 성능 차이가 발생한다. 아래 표1은 미세조정(fine-tuning) 없이 동일한 프롬프트에 대해 코난 LLM 13B와 Polyglot-ko 12.8B가 생성한 답변이다. 그리고 한국어 벤치마크 데이터셋인 KOBEST [7]로 평가한 결과를 본고에서 구체적으로 밝힐 수 없지만, SentiNeg의 제로샷을 제외한 모든 태스크의 제로샷과 퓨샷에서 Polyglot-ko 12.8B보다 코난 LLM 13B가 높은 성능을 보였다.

#### 4. 미세조정

LLM이 다양한 태스크로 학습할수록 새로운 태스크에서 더 좋은 성능을 보일 가능성이 커진다고 알려져 있다 [8]. 코난 LLM도 이에 맞춰 학습 태스크 종류를 계속 늘려가며 튜닝을 진행하고 있다. 코난 LLM을 이용하여 실험한 결과, 한 가지 태스크로만 학습한 코난 LLM은 정확도가 84.2%인 반면, 다양한 태스크로 학습한 코난 LLM은 동일 태스크에 대해 93.9%로 약 10% 가까운 성능 향상을 보였다.

표 1 코난 LLM 13B와 Polyglot-ko 12.8B 답변 생성 결과 예시

프롬프트	코난 LLM 13B 답변	Polyglot-ko 12.8B 답변
아래 문서의 맞춤법과 띄어쓰기를 교정해서 다시 작성하십시오. 문서: 작년 실적 역기저 효과 방어 가능할 쥘망  KT&G의 2분기 연결기준 영업이익은 3,276억원으로 시장 컨센서스를 소폭 하회하였다. 전사 매출이 11% 성장했음에도 불구하고, 매출 믹스 악화효과가 예상보다 크게 나타나면서, 시장 기대치를 하회한 것으로 판단된다. 작년 미국 쥘련 수출 및 부동산 분양 실적 역기저 에도 불구하고, 내수담배 MS 확대, 글로벌 담배 매출 성장, 달러 강세 영향 등에 힘 입어, 올해 실적도 작년과 유사할 수준을 기록할 것으로 쥘망된다. 맞춤법과 띄어쓰기 교정 후:	작년 실적 역기저 효과 방어 가능할 쥘망  KT&G의 2분기 연결기준 영업이익은 3,276억원으로 시장 컨센서스를 소폭 하회하였다. 전사 매출이 11% 성장했음에도 불구하고, 매출 믹스 악화효과가 예상보다 크게 나타나면서, 시장 기대치를 하회한 것으로 판단된다. 작년 미국 쥘련 수출 및 부동산 분양 실적 역기저 에도 불구하고, 내수담배 MS 확대, 글로벌 담배 매출 성장, 달러 강세 영향 등에 힘 입어, 올해 실적도 작년과 유사할 수준을 기록할 것으로 쥘망된다.	플작년 실적 역기쥘 효과 방어 가능할 쥘망 쥘 쥘련 수출및 부동산 분 산 영향 등에 힘입어, 올해 실적도 작년과 유사할 수준을 기록할 것으로 쥘망된다. 쥘련 수출 및 부동산 분 산 영향등에 힘입어,올해 실적도 작년과 유사할 수준을기록할 것으로쥘망된다.쥘련 수출및 부동산분산효과등에힘입어,올해실적도작년과 유사할수준을기록할것으로쥘망쥘다. 쥘련수출및부동산분산영향등에힘입어나수담배MS확대,글로벌담배매출성장,달러강세타영향등에이해올해실적도작년과이유사할수준위 기록할것으쥘쥘망된다. 쥘련 겨를작년 쥘련 수출 및 부동산 분산 영향 등에 힘있어, 올해 쥘망도 작년과 비슷할 수준위 것으로쥘 쥘망된다 쥘련 겨를지난해 쥘련 수출 및 부동산 분산 영향 등에 힘있어, 올해쥘망 ...
고객 상담 데이터 중 욕설이 입력되면 자동으로 욕하지 말라는 멘트를 다양하게 랜덤으로 리턴하려고 합니다. 멘트 10개를 만들어주세요	4. 고객님 욕설이 감지되어 상담이 중단되었습니다. 고객님의 욕설은 상담원뿐만 아니라 다른 고객님들께도 불쾌감을 드릴 수 있습니다. 바른말 고운말 사용 부탁드립니다. 5. 고객님 욕설을 자제해 주시기 바랍니다. 욕설을 하실 경우 상담이 중단될 수 있습니다. 6. 고객님 욕설은 상담원과 고객님 모두에게 상처를 줍니다. 바르고 고운 말을 사용해 주세요. 7. 고객님 욕설로 인해 상담이 중단되어 죄송합니다. 욕설을 자제해 주시면 감사하겠습니다. 8. 안녕하세요 고객님. 고객님의 고운 말 한마디가 상담원에게는 큰 힘이 됩니다. 고운 말 사용 부탁드립니다. 9. 고객님 상담 중 욕설을 자제해주시기 바랍니다. 상담원에게 욕설은 큰 상처가 됩니다. 10. 고객님 욕설 자제 부탁드립니다.	특4. 고객님 상담도중 욕설이 감지되어 통화가 종료되었습니다. 다시 전화주시면 친절하게 상담해 드리겠습니다. 특5. 고객님 상담 중 욕설이 감지 되었습니다. 고객님의 안전을 위하여 통화가 종료 되었습니다. 다음에 다시 전화 주시기 바랍니다. 특6. 고객님 상담중 욕설이 탐지되었습니다. 상담원의 안전을 위해 통화를 종료하였습니다. 다시 전화를 걸어주시기 바랍니다. 특7. 고객님의 욕설로 인하여 상담이 종료되었음을 알려드립니다. 다음에는 바른말 고운말 사용 부탁드립니다. 특8. 고객님께서 욕설을 하셨습니다. 상담원은 욕설을 듣지 않았습니다. 고객님께서 욕설을 하지 않으셨습니다. 특9. 고객님께서 심한 욕설을 하였습니다. 상담원은 심한 욕설을 듣지 못하였습니다. 고객님은 심한 욕설을 하시지 않았습니다. 특10. 고객님의 심한 욕설로 상담이 종료되었음을 알려 드립니다. 다음에는 고운말 바른말 부탁드립니다.

표 2 전체 미세조정 시 필요한 GPU 스펙

컨텍스트 길이	파라미터 타입	GPU 스펙	1스텝 학습 시간
4096	float32	H100-80GB * 8	약 6초
4096	float16	RTX6000-48GB * 8	약 30초

## 5. 최적화

코난 LLM은 학습과 추론 비용을 고려하여 최적의 파라미터 크기를 결정했다. 그 결과 코난 LLM 13B는 저성능 GPU인 RTX 3090 × 2 환경에서 서비스가 가능하다. 파라미터 크기뿐만 아니라 사전학습, 미세조정, 추론에서 다양한 기법을 사용하여 최적화했다.

사전학습에서 다중GPU를 사용하기 위해서 모델을 3D 병렬화(Parallelism) [9] 했다. 3D 병렬화는 텐서 병렬화, 파이프라인 병렬화, 데이터 병렬화로 구성되어 있다. 그리고 효율적인 메모리 사용과 빠른 학습을 위해서 bfloat16 [10]과 Zero Redundancy Optimizer [11] 그리고 FlashAttention [12]을 적용했다.

고객사의 데이터셋을 이용하여 내부에서 미세조정을 원하는 경우가 있지만, 전체 미세조정(full fine-tuning)의 경우 많은 GPU 자원이 필요하다. 표2는 코난 LLM 13B를 전체 미세조정하는 경우 컨텍스트 길이 및 파라미터 타입에 따른 GPU 스펙 요구 사항이다. 미세조정 비용을 절감하기 위해서 PEFT 기법 중 QLoRA [13]를 사용했고 전체 미세조정(full fine-tuning) 대비 4배 가까이 GPU 메모리를 절감할 수 있었다. PEFT 기법을 사용하면 GPU 자원을 절약할 수 있지만, 적절한 하이퍼파라미터 설정이 동반되지 않을 경우 품질 저하가 발생할 수 있다.

추론에서 FlashAttention과 PagedAttention [14] 그리고 Continuous Batching [15]를 사용하여 기존 대비 처리량을 15배 높였다.

## 6. 검색 증강 생성

LLM의 답변을 보면 몇 가지 문제점이 있다. 첫째, 최신 정보나 전문 분야의 정보와 같이 사전학습 데이터셋에 포함되지 않은 정보는 답할 수 없다. 둘째, 사전학습 데이터에 포함된 정보라 할지라도 확률에 따라 허위 정보를 답하는 할루시네이션(Hallucination)이 발생할 수 있다.

이런 문제를 완화하기 위해, 프롬프트와 정답 쌍으로 구성된 추가 데이터셋으로 LLM을 미세 조정할 수 있지만, 미세조정 방식은 실행이 용이하지도 않을뿐더러 최신 데이터가 자주 변경되는 경우 빈번한 재학습이 요구돼 비용과 시간이 많이 드는 단점이 있다.

검색증강 생성(Retrieval Augmented Generation)은 이런 단점을 극복하는 효과적인 수단이다. 사실 RAG의 기본 개념은 간단하다. 먼저 사용자 질문을 검색어로 활용해 관련 문서를 검색하게 하고, 검색된 문서의 내용을 LLM에 컨텍스트로 주고 사용자 질문에 대한 답을 생성하게 하면 된다. 그러나 기본 개념에 비해 실제 구현에는 까다로운 이슈들이 많다. 광범위한 검색 시스템의 구축은 차치하더라도, 사용자 질문에 대한 검색어 확장, 벡터 임베딩에 의한 시맨틱 검색, 문서별 답변의 랭킹 등 여러 가지 기술적 고려가 필요하다. 최근에는 문서별 답변 랭킹(Rerank)를 포함해서 Re2G(Retrieve, Rerank, Generate) 아키텍처도 제안하고 있다 [16].

코난 LLM은 벡터 기반 시맨틱 검색과 결합한 RAG를 통해 최신의 정보와 명확한 근거에 기반한 답변을 제공한다. RAG에서 LLM의 역할은 검색기(retriever)와 정렬기(reranker)를 통해 선택한 문단으로부터 답변을 생성하는 것이다. 할루시네이션을 억제하기 위해 외부 컨텍스트를 주입하고 이를 바탕으로 답변을 생성한다. 이때 지식충돌(knowledge conflict)과 자제력 있는 예측(prediction with abstention) 이슈를 해결해야 하는데 [17], 코난 LLM은 아래의 예와 같이 주입된 컨텍스트를 바탕으로 적절한 답변을 생성한다.

### 6.1 지식 충돌 (knowledge conflict)

LLM이 자기가 미리 학습한 지식 기반으로 답을 하지 않고, 주어진 컨텍스트에 의존해서 답을 할 수 있어야 한다. 아래와 같이 컨텍스트가 주어지지 않으면 LLM은 자신이 학습한 지식으로만 가지고 답을 하므로 오답을 답할 가능성이 있다.

*Question: 현재 대한민국 대통령은 누구인가요?*

*Answer(Context 없이): 대한민국 대통령은 문재인입니다.*

반면, 코난 LLM에게 아래와 같은 컨텍스트를 주고 컨텍스트 기반으로 답을 하게 했을 때 정답을 말하는 것을 볼 수 있다.

*Context: 대한민국의 대통령은 대한민국의 국가원*

수이자 행정부 수반이다. 현행 헌법에 따라 5년 단임제를 따른다. 현직 대통령은 2022년 5월 10일 취임한 제20대 윤석열 대통령이다.

*Question:* 현재 대한민국 대통령은 누구인가요?

*Answer(Context 기반):* 현재 대한민국 대통령은 윤석열입니다. 이는 문서에서 “2022년 5월 10일 취임한 제20대 윤석열 대통령이다.”라고 명시되어 있습니다.

## 6.2 자체력 있는 예측(prediction with abstention)

사용자 요구와 관련 없는 컨텍스트가 주어졌을 때, LLM은 컨텍스트 기반으로 무작정 답을 하지 않고 해당 컨텍스트에서는 사용자 요구에 맞는 단서를 찾을 수 없다는 답을 내놓을 수 있어야 한다.

*Question:* 현재 대한민국 대통령은 누구인가요?

*Context:* 문재인은 1952년 1월 24일 경상남도 거제군 거제면 명진리 694-1번지에서 아버지 문용형과 어머니 진주 강씨 강한옥 사이에서 2남 3녀 중 장남으로 태어났다.

*Answer(Context 기반):* 문서에는 문재인 대통령의 출생지, 가족관계, 학력 등에 대한 정보가 있지만, 현재 대한민국 대통령이 누구인지에 대한 정보는 포함되어 있지 않습니다.

## 6.3 RAG 적용 시 권한 관리 및 접근 제어를 통한 보안

RAG를 통해 최신의 외부 데이터 소스로부터 신뢰할 수 있는 답변을 생성할 수 있지만, 각 부서나 사용자의 접근 권한에 따라 서로 다른 문서를 접근해야 하고 그에 따라 RAG에 의한 답변이 달라지는 경우가 많다. 코난 LLM은 문서의 관련성뿐만 아니라 사용자 접근 권한에 따라 반환되는 정보를 제어하는 기능도 RAG를 통해 제공한다.

예를 들어, 내부 정보 조회 권한이 없는 외부 직원이 보상 범위에 대한 정보를 찾는다고 가정해 보자. 코난 LLM은 RAG 접근 제어를 통해 해당 사용자의 권한에 부합하는 보상 범위만 답변으로 제공하고 다른 정보에 대한 접근은 일체 제한해, 사용자 권한에 부합하는 답변을 만들 수 있다. 마찬가지로 기업이나 조직은 RAG 및 접근 제어 기술을 통해 특정 사용자 그룹에게만 관련된 정보를 제공해줄 수도 있다. 이는 충성도 프로그램, 계층화된 서비스 계약과 같은 다양한 종류의 제품 및 서비스를 제공하는 조직에 특히 유용하다.

## 7. 활용 사례

LLM은 대표적인 능력인 텍스트 생성 외에도 질문-답변, 요약, 고쳐쓰기, 분류, 번역 등 다양한 자연어 처리 능력을 갖고 있다. 그러나 현재 기술 수준으로는 환각 등 고질적 문제로 인해 LLM의 제공 결과에도 신뢰성 문제가 있기 때문에 분야별로 또는 업무별로 선택적인 기능을 제한된 범위에서 활용하고 있다. 예를 들어, 카피라이팅 문구 작성처럼 창작 능력이 중요한 분야에서는 생성이 중요한 기능이지만, 금융 상담처럼 정확한 답변을 해야 하는 분야에서는 거짓이 포함된 LLM의 답변 결과는 쓸모가 없다. 이런 이유로, LLM이 직접적으로 최종 사용자에게 답변을 하기보다는 초벌 결과를 사람의 검토를 거치는 방법으로 접근하는 것이 현재의 일반적인 활용 방법이다. 코난 LLM은 B2B 또는 B2G 영역에서 다음과 같은 문제의 풀이에 집중하고 있다.

- 업무문서/보고서 초안 생성: 문서 작성 요청 프롬프트(예, “스마트 팩토리의 기술 동향 및 시장 전망”)를 입력해 초안을 생성할 수 있다. 업무 생산성 향상 및 중요한 일에 시간 집중할 수 있다.
- 계약서 검토: 계약서 내 불공정 또는 함정 조항을 탐지해, 불공정/불완전 계약 리스크를 줄일 수 있다.
- 자연어 인터페이스: 물류, 제조, ERP 등 기간제 시스템에 자연어 질의로 데이터를 조회, 분석하거나 자연어로 생성된 답변을 얻을 수 있다. 분석 없이 운영자 혼자서도 데이터 분석이 가능해지므로 전사 차원에서 데이터 기반 판단 프로세스의 확립이 가능해진다.
- 고객 맞춤 상담: AICC · 챗봇 등 고객 상담 시스템에서 고객 질의에 대해 고객 개인 정보(가입 보험 상품, 가입 시기 및 진단서 등)와 보험 약관 그리고 고객 개인 정보(투자 성향, 여유 자금 운용 규모, 운용 기간 등)와 금융 상품 등에 기반한 답변 초안 생성이 가능하다.
- 고객 상담 후처리 자동화: 상담 내용 자동 요약 및 분류, 예약 등 레거시 시스템과의 자동 연동 등 통해 업무 생산성 향상이 가능해진다.
- 고객 타겟 마케팅 활용: 상담 내용 기반으로 고객의 관심사나 성향을 파악해서 상품 등 추천 문구 초안 생성해서 발송 매출 유도가 가능해진다.
- 지속적 정보 수집과 실시간 대응 방안 초안 생성: 정보나 국방 분야에서 지속적으로 인입되는 정보

의 수집 시 효과적인 요약과 긴급 시 즉각 대처 방안 초안 생성 등 효율적인 정보 수집 및 대응이 가능하다.

## 8. 결 론

이상에서 코난 LLM 기반 모델의 사전 학습에 대해 살펴보았다. 이 과정에서 상당히 큰 규모의 컴퓨팅 자원이 소요되긴 했으나 글로벌 기업이나 대기업이 아니더라도 가용한 자원을 필요한 곳에 집중하면 충분히 좋은 품질의 LLM을 만들 수 있음을 확인하였다. 사용자가 보유한 데이터셋이 충분하다면 사용자만의 기반 모델을 학습해 보유하는 것도 가능할 것이다.

글로벌 기업이 제공하는 LLM 서비스는 데이터 보안 문제는 차치하더라도 미세조정을 거치더라도 전문 영역의 표현들이 갖는 문맥을 충분히 반영하기 어렵다. 법률/의료/금융 등 전문 영역에서 LLM에 질문을 던지면 전문용어가 갖는 의미공간이 충분히 학습이 안 되어 있어 질문의 이해, 또는 시맨틱 검색의 정확도가 많이 떨어지는 문제가 있다. 이런 경우 코난 LLM은 전문 영역에 특화된 LLM을 합리적 비용으로 제공할 수도 있어서 많은 현장 문제의 해결에 도움을 줄 수 있을 것이다.

## 참고문헌

- [ 1 ] Jordan Hoffmann et al, Training Compute-Optimal Large Language Models, <https://arxiv.org/abs/2203.15556>
- [ 2 ] Nvidia V100 data sheet, <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/v100-application-performance-guide.pdf>
- [ 3 ] Nvidia A100 data sheet, <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf>
- [ 4 ] Nvidia H100 data sheet, <https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet>
- [ 5 ] Deepak Narayanan et al, eEfficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM, <https://arxiv.org/abs/2104.04473>
- [ 6 ] Hyunwoong Ko et al, A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models
- [ 7 ] Dohyeong Kim et al, KOBEST: Korean Balanced Evaluation of Significant Tasks
- [ 8 ] Hyung Won Chung et al, Scaling Instruction-Finetuned Language Models, <https://arxiv.org/abs/2210.11416>

- [ 9 ] huggingface docs, <https://huggingface.co/docs/transformers/v4.15.0/parallelism#model-parallelism>
- [10] Dhiraj Kalamkar et al, A Study of BFLOAT16 for Deep Learning Training, <https://arxiv.org/abs/1905.12322>
- [11] Samyam Rajbhandari et al, ZeRO: Memory Optimizations Toward Training Trillion Parameter Models, <https://arxiv.org/abs/1910.02054>
- [12] Tri Dao et al, FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, <https://arxiv.org/abs/2205.14135>
- [13] Tim Dettmers et al, QLoRA: Efficient Finetuning of Quantized LLMs, <https://arxiv.org/abs/2305.14314>
- [14] Woosuk Kwon et al, Efficient Memory Management for Large Language Model Serving with PagedAttention, <https://arxiv.org/abs/2309.06180>
- [15] Anyscale blog, <https://www.anyscale.com/blog/continuous-batching-llm-inference>
- [16] Michael Glass et al, Re2G: Retrieve, Rerank, Generate, <https://arxiv.org/abs/2207.06300>
- [17] Wenxuan Zhou et al, Context-faithful Prompting for Large Language Models, <https://arxiv.org/abs/2303.11315>

## 약 력

### 양 승 현



1990 서울대학교 컴퓨터공학과 졸업(학사)  
 1992 서울대학교 컴퓨터공학과 졸업(석사)  
 1997 서울대학교 컴퓨터공학과 졸업(박사)  
 1997-1999 한국전자통신연구원 선임연구원  
 1999-2022 코난테크놀로지 CTO  
 2022-현재 SK텔레콤/코난테크놀로지 CTO  
 관심분야: 자연어처리, 인공지능  
 Email : seunghyun.yang@konantech.com

### 도 원 철



1992 홍익대학교 전자계산학과 졸업(학사)  
 1994 홍익대학교 전자계산학과 졸업(석사)  
 1994-2011 핸디소프트 이사  
 2011-현재 코난테크놀로지 상무이사  
 관심분야: 자연어처리, 인공지능  
 Email : wonchul.do@konantech.com



### 오 창 민

2000 서울대학교 컴퓨터공학과 졸업(학사)  
2002 서울대학교 컴퓨터공학과 졸업(석사)  
2002~2006 엠파스 선임연구원  
2006~현재 코난테크놀로지 상무이사  
관심분야: 대용량검색, 인공지능  
Email : changmin.oh@konantech.com



### 김 정 태

2014 한국해양대학교 데이터정보학과 졸업(학사)  
2016 한국해양대학교 데이터정보학과 졸업(석사)  
2018 한국해양대학교 데이터정보학과 수료(박사)  
2018~현재 코난테크놀로지 선임연구원  
관심분야: 자연어처리, 인공지능  
Email: jeongtae.kim@konantech.com