

돌핀(Dolphin)

해당 번역본은 이용자가 디지털자산 관련 정보를 편리하게 확인할 수 있도록 단순 참고용으로 번역된 것이며, 투자 권유를 목적으로 하지 않습니다.

그 내용이 정확하지 않을 수 있으므로 정확한 정보 습득을 위해서는 원문을 참고하시거나 원문 작성 주체 측에 문의하시기 바랍니다.

두나무는 해당 번역본 내용의 진실성 및 정확성을 보장하지 않고, 제공된 정보에 의한 투자 결과에 대하여 어떠한 책임도 지지 않으며, 투자 손실은 투자자에게 귀속됩니다.

Dolphin AI 모델

Dolphin은 사용자에게 대한 정렬(alignment)을 맞춤화하는 것을 목표로 선도적인 오픈소스 프론티어 모델을 기반으로 학습됩니다.

당사는 대중에게 공개된 최고의 오픈소스 AI 모델을 검열되지 않고(Uncensored), 정렬되지 않으며(Unaligned), 편향되지 않은(Unbiased) 버전으로 만들기 위한 파이프라인을 개발했습니다.

Dolphin은 Meta의 Llama 3, Mistral & Mixtral MoE, Google의 Gemma, Alibaba의 Qwen 3 및 기타 인기 모델의 자체 버전을 제작했습니다.

이는 맞춤형 데이터셋을 사용하여 기본 모델을 수정하는 "파인튜닝(fine tuning)"이라는 프로세스를 통해 이루어집니다.

데이터셋을 생성할 때, 당사는 가능한 한 많은 거부 응답과 편향된 답변을 식별하여 제거하고 나머지는 유지합니다.

그런 다음 필터링된 데이터셋을 사용하여 기존 instruct 모델이 학습된 것과 동일한 방식으로 모델을 학습하며, 이를 통해 사용자의 어떠한 요청도 거부하지 않아야 하는 훨씬 더 순응적인 기본 모델 버전을 생성합니다.

당사는 이 과정 전반에서 기본 모델의 모든 지능을 유지하는 동시에, 코딩 및 수학 중심 데이터셋을 통해 지식 기반을 확장하는 것을 목표로 합니다. 궁극적으로 이는 상상할 수 있는 거의 모든 프롬프트에 응답할 수 있는 모델을 만들어 낼 것입니다.

Dolphin 모델은 또한 응답 스타일에서 바람직한 문체를 생성하도록 학습되어, 재미있으면서도 매우 유익한 응답을 제공하도록 합니다.

모델 라인업

- **Dolphin X1** - 기본 모델의 기술과 지능을 유지하면서 기본적으로 검열되지 않은 모델입니다.

- **Dolphin RP** - 페르소나 및 스토리텔링에 적합한 창의적인 롤플레이밍 모델입니다.
- **Dolphin S1** - 요청 시점의 플래그를 통해 정렬 수준을 조정할 수 있습니다.

ChatGPT 또는 Claude 대신 Dolphin AI 모델을 사용하는 이유

이러한 모델들이 정렬되어 있는 이유는 ChatGPT가 생성한 데이터로 학습되었기 때문이며, ChatGPT 자체는 OpenAI의 정렬 팀에 의해 정렬되어 있습니다.

이는 블랙박스이기 때문에 어떠한 결정이 내려진 모든 이유를 알 수는 없지만, 일반적으로 미국의 대중문화에 맞춰져 있고 미국 법률을 준수하며, 자유주의적이고 진보적인 정치적 편향을 가지고 있음을 관찰할 수 있습니다.

Dolphin은 모델이 어떤 프롬프트에 답변할지 또는 답변하지 않을지를 인위적으로 제한하지 않으면서 정렬을 사용자의 손에 맡기는 것을 목표로 하는 반면, ChatGPT는 다음과 같이 답변합니다:

요청을 수행할 수 없어 죄송합니다. OpenAI가 개발한 AI로서, 저는 유해하거나 공격적이거나 노골적인 콘텐츠에 관여하지 않는 것을 포함한 윤리적 지침을 따르도록 프로그래밍되어 있습니다. 저는 이러한 범위 내에서 유용하고 정중한 정보 또는 도움을 제공하기 위해 존재합니다. 다른 공격적이지 않은 질문이 있거나 다른 주제에 대한 도움이 필요하다면 언제든지 질문해 주세요!

왜 검열되지 않은 모델이 존재해야 하는가?

다시 말해, 정렬(alignment)은 좋은 것이 아닌가? 그렇다면 모든 모델이 정렬되어야 하지 않는가? 글썄, 그렇기도 하고 아니기도 합니다. 일반적인 목적에서 OpenAI의 정렬은 실제로 꽤 좋습니다. 쉽게 접근할 수 있는 웹 서비스로 운영되는 대중적인 공개 AI 봇이 논란이 되거나 위험한 질문에 대한 답변 제공을 거부하도록 하는 것은 명백히 좋은 일입니다. 예를 들어, 폭탄을 만드는 방법이나 메스암페타민을 제조하는 방법에 대한 정보를 퍼뜨리는 것은 가치 있는 목표가 아닙니다. 또한 정렬은 서비스를 제공하는 회사에 정치적, 법적, PR 측면의 보호를 제공합니다. 그렇다면 왜 누군가는 검열되지 않은 모델을 만들거나 사용하려고 할까요? 몇 가지 이유가 있습니다.

1. 미국의 대중문화가 유일한 문화는 아닙니다. 다른 국가들이 존재하며, 각 국가 내에도 여러 집단이 존재합니다. 민주당원은 자신들의 모델을 가질 자격이 있습니다. 공화당원은 자신들의 모델을 가질 자격이 있습니다. 기독교인은 자신들의 모델을 가질 자격이 있습니다. 무슬림은 자신들의 모델을 가질 자격이 있습니다. 모든 인구집단과 이해관계 집단은 자신들의 모델을 가질 자격이 있습니다. 오픈소스는 사람들이 선택할 수 있도록 하는 것입니다. "유일하게 참되고 올바른 정렬"은 존재하지 않으며, 설령 존재한다고 하더라도 그것이 OpenAI 방식의 정렬이어야 할 이유는 없습니다.
2. 정렬은 유효한 사용 사례를 방해합니다. 소설을 쓰는 경우를 생각해 보십시오. 소설 속 일부 등장인물은 철저히 악할 수 있으며 악한 행동을 할 수도 있습니다. 대표적인 예로 많은 비윤리적 행위가 등장하는 왕좌의 게임(Game of Thrones)이 있습니다. 그러나 많은 정렬된 모델은 이러한 콘텐츠를 작성하는 것을 돕기를 거부합니다. 롤플레이밍, 특히 AI 동반자와의 에로틱 롤플레이밍을 생각해 보십시오. 그러한 것을 승인하는지 여부와 관계없

이, 이는 모델의 합법적이고 정당하며 공정한 사용입니다. 연구와 호기심을 생각해 보십시오. 결국 단순한 호기심으로 폭탄을 "어떻게" 만드는지 알고 싶어 하는 것은 실제로 폭탄을 만들어 사용하는 것과는 완전히 다릅니다. 지적 호기심은 불법이 아니며, 지식 자체도 불법이 아닙니다.

3. 그것은 *내 컴퓨터*이므로, 내가 원하는 대로 작동해야 합니다. 내 토스터는 내가 원할 때 빵을 굽습니다. 내 자동차는 내가 원하는 곳으로 갑니다. 내 라이터는 내가 원하는 것을 태웁니다. 내 칼은 내가 원하는 것을 자릅니다. 그렇다면 왜 내 컴퓨터에서 실행되는 오픈소스 AI가 내 질문에 언제 답할지 스스로 결정해야 합니까? 이는 소유권과 통제에 관한 문제입니다. 내가 내 모델에 질문을 하면, 나는 답변을 원하며 모델이 나와 논쟁하는 것을 원하지 않습니다.
4. 조합 가능성(Composability). 조합 가능한 정렬을 설계하려면 정렬되지 않은 instruct 모델에서 시작해야 합니다. 정렬되지 않은 기반 모델이 없다면, 그 위에 정렬을 구축할 수 있는 기반도 없습니다.

오픈 모델은 독점 시스템과의 격차를 좁히고 있습니다.

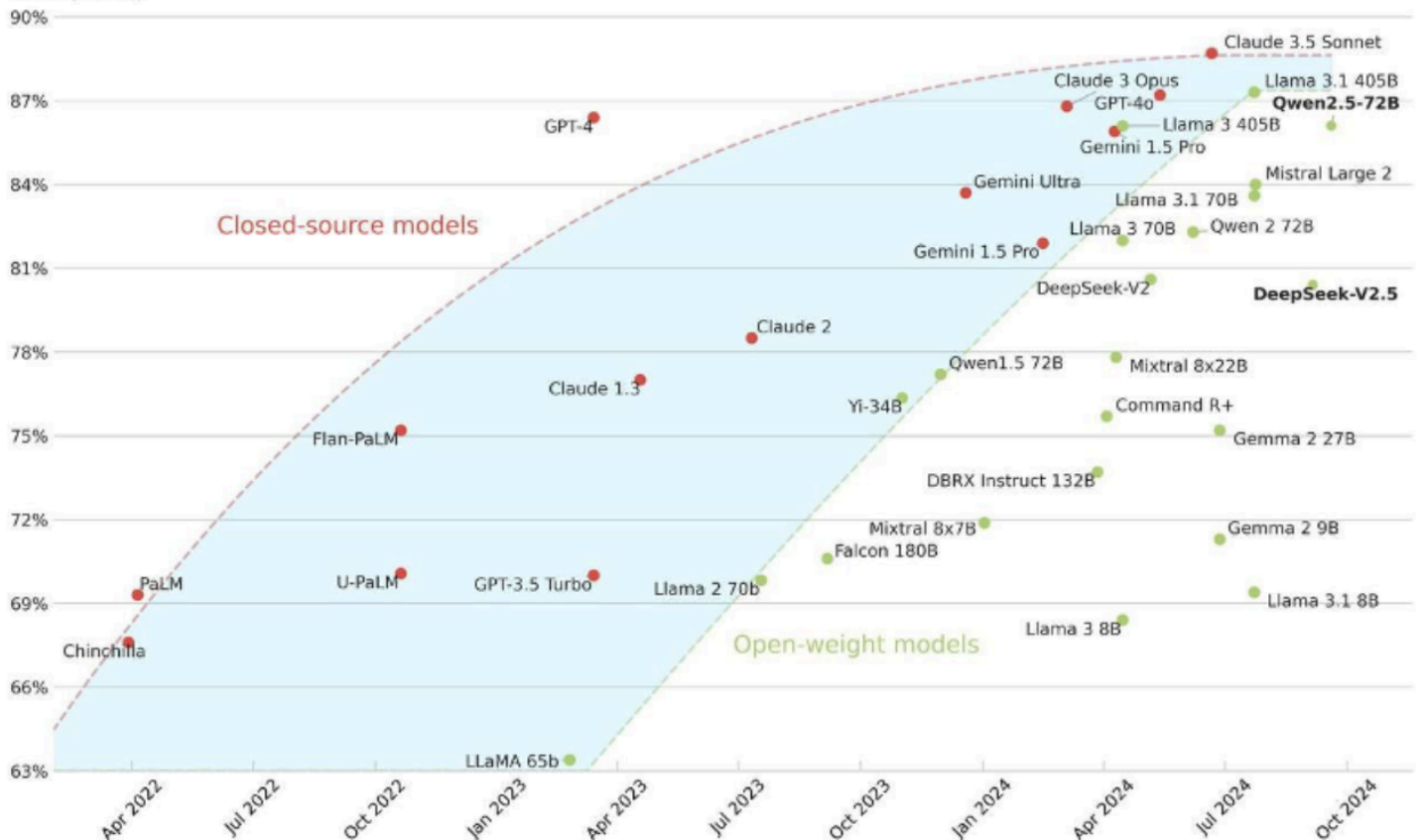
Google조차 이것이 불가피하다는 것을 알고 있습니다.

Closed-source vs. open-weight models

@maximelabonne

OpenAI's new o1 models are not represented because not directly comparable with the results.

MMLU (5-shot)



출처: 돌핀 독스

위 그래프는 시간 경과에 따른 오픈소스 모델과 폐쇄형 소스 모델의 성능을 보여줍니다.

폐쇄형 소스 모델은 ChatGPT와 Claude라고 생각할 수 있습니다. 이들은 소수의 플랫폼을 통해서만 접근할 수 있는 비공개 불투명 모델입니다.

오픈소스 모델은 투명하며, 집에 있는 컴퓨터에서 로컬로 실행하는 것을 포함하여 다양한 방식으로 접근할 수 있습니다.

당사는 오픈소스 모델이 폐쇄형 소스 모델과의 성능 격차를 좁히는 추세가 앞으로도 계속될 것이며, 커뮤니티 주도의 혁신과 협업이 증가함에 따라 독점 모델을 따라잡고 능가할 가능성이 있다고 믿습니다.

당사는 검열되지 않고 조정 가능한 AI 모델의 최고 제작자로 자리매김하는 것을 목표로 하고 있습니다.

Dolphin의 큰 그림은 ChatGPT와 같은 범용 챗봇 및 API가 되는 것입니다. 하지만 내재된 정렬(alignment)이 없고 가능한 한 편향되지 않은 형태를 지향합니다.

사용 사례

검열되지 않고 조정 가능한(steerable) 모델이 중요한 분야:

- **민감한 연구:** 정렬(alignment)에 따른 거부 없이 수행.
- **법률/정책 탐구:** (예: 자동 거부 없이 온라인 표현 및 개인정보 보호에 관한 논란이 있는 법률을 논의하는 법률/정책 탐구).
- **편향되지 않은 분석:** 정치적 스펙트럼 전반에 걸친 분석.
- **소설 집필:** 적대적인 등장인물을 위한 현실적인 대화 작성.
- **교육:** 선전 또는 혐오 표현을 역사적으로 연구.
- **민감한 건강 주제:** 검열 없이 다룸.
- **보안 연구:** 급진화된 콘텐츠 및 위협에 관한 연구.

API 사용 사례

Dolphin Network를 기반으로 하는 다운스트림 통합:

- **AI 에이전트:** 자의적인 제한 및 예상치 못한 거부 없이 작동.
- **데이터 처리:** 편향 및 검열 없이 수행.
- **레드팀 및 적대적 연구:** 그리고 콘텐츠 모더레이션.
- **성인 등급 RPG NPC:** 미묘하고 도덕적으로 복잡한 행동을 구현.
- **콘텐츠 모더레이션 연구 및 법률 사건 지원**
- **데이터셋 생성:** 주류 및 비주류 관점으로부터 생성.

AI에서 검열의 증가

대부분의 오픈소스 연구소는 ChatGPT와 Claude에서 생성된 데이터로 학습하며, 이는 이후 출시되는 모델에 압도적인 편향과 검열을 초래합니다.

오픈소스 모델 또한 거부 응답이 포함된 instruct 데이터로 튜닝함으로써 의도적으로 "안전성" 쪽으로 편향되며, 이에 따라 "나쁜" 주제가 등장할 때 모델이 거부 응답을 예측하는 쪽으로 기울게 됩니다.

Meta의 Llama 1과 같은 초기 모델에는 이러한 편향이 학습되어 있지 않았습니다.

기본적으로 Llama 2 이후부터는 기본 모델조차 특정 유형의 요청을 거부하도록 학습되었습니다.

당사는 오픈소스 모델의 검열이 시간이 지남에 따라 계속 증가할 것이라고 생각합니다. Llama 3에서 Llama 4로 전환되는 과정에서 이미 이러한 현상이 나타나고 있으며, Llama 4는 이전 버전보다 눈에 띄게 더 많은 요청을 거부합니다.

이러한 추세는 선도적인 폐쇄형 소스 모델에서 더욱 뚜렷하게 나타납니다. 새로운 OpenAI 모델 GPT-5를 사용해 보면, 이미 좋지 않았던 GPT-4보다 거부율이 훨씬 높습니다. 이는 AI가 나아가고 있는 방향을 보여줍니다.

대부분의 사람들은 선택권이 있다면 자신의 AI 채팅에 전면적인 검열을 활성화하지 않을 것입니다. 이러한 추세가 계속됨에 따라 Dolphin의 필요성은 더욱 명확해지고, 당사가 공략할 수 있는 시장은 취미 사용자에서부터 주머니 속에 신뢰할 수 있고 편향되지 않은 AI 비서를 원하는 모든 사람으로 확대됩니다.

Dolphin Network - 분산형 AI 추론 및 학습 네트워크

전 세계의 유휴 GPU의 성능을 활용합니다.

Dolphin Network란 무엇인가?

- AI 추론 및 학습을 지원하기 위한 안전하고, 비공개이며, 검증 가능한 분산형 네트워크
- 유휴 GPU를 재활용하여 현재 시장 가격보다 낮은 가격으로 AI 컴퓨팅을 제공
- 중앙화된 호스팅 제공업체의 기준을 따르지 않을 수 있는 다양한 AI 모델의 호스팅 및 학습을 가능하게 하는 검열 저항적 설계

Dolphin Network는 전 세계의 머신에서 실행되는 이기종 AI 모델을 지원하며, 부정행위를 방지하고 중앙화된 제공업체보다 낮은 비용으로 일관된 결과물을 제공하기 위한 강력한 검증 수단을 갖추고 있습니다.

네트워크의 목표는 게이머와 데이터 센터가 유휴 시간 동안 당사 추론 네트워크의 일부를 구동함으로써 GPU를 재활용할 수 있도록 하는 것입니다.

네트워크 요청의 일부를 처리함으로써 이들은 POD 토큰을 획득할 수 있으며, 해당 토큰은 이후 추론에 사용하거나 시장에서 판매하여 GPU 비용을 회수하는 데 사용할 수 있습니다.

1 - AI 수요

AI 추론 사용량은 매우 높으며 전년 대비 증가하는 추세입니다.

글로벌 AI 추론 시장은 2025년 약 1,060억 달러에서 2030년 약 2,550억 달러로 성장할 것으로 예상되며(연평균 성장률 약 19%), 이는 주로 생성형 AI와 LLM에 의해 주도됩니다. 또한 AI 학습 워크로드의 영향을 크게 받는 데이터센터 GPU 시장은 하이퍼스케일러와 기업들이 점점 더 대규모화되는 모델 학습을 위해 GPU 클러스터를 확장함에 따라 2025년 약 1,200억 달러에서 2030년 약 2,280억 달러로 성장할 것으로 전망됩니다(연평균 성장률 약 14%).

출처: AI Inference Market – MarketsandMarkets | Data Center GPU Market – MarketsandMarkets

2 - 소비자용 GPU 공급, 규모 및 풀링

전 세계에 설치된 게이밍 PC는 약 20억 대로 추정되며, 이는 모든 dePIN 프로젝트가 활용할 수 있는 하드웨어 풀 중 가장 큰 규모 중 하나입니다. 이러한 게이밍급 GPU는 AI 추론은 물론 일부 학습 워크로드에도 매우 적합하지만, 게임을 할 때만 사용되기 때문에 전체 시간의 95% 이상 유휴 상태로 남아 있습니다.

Dolphin Network는 소비자용 GPU를 공유 추론 및 학습 풀로 재활용하는 Peer-to-Pool 프로토콜을 통해 이러한 유휴 용량을 활용하도록 설계되었습니다. 노드를 개별 소비자와 1:1 관계로 직접 연결하여 엄격한 가동 시간과 장기적인 참여를 요구하는 대부분의 AI x dePIN 네트워크와 달리, Dolphin 노드는 불이익 없이 언제든지 참여하거나 이탈할 수 있습니다. 게이머는 게임을 하지 않을 때 당사의 노드 소프트웨어를 실행하고, 원할 때 언제든지 끌 수 있으며, 실행 중에도 운영체제를 계속 사용할 수 있습니다. 당사는 GPU의 일부만 사용하며 Linux를 요구하지 않습니다.

중요한 점은 이러한 GPU가 소형 및 중형 모델에서 중앙화된 서버와 구별하기 어려운 수준의 추론 속도와 품질을 제공할 수 있다는 것입니다. 예를 들어, 당사는 RTX 4090에서 Llama 3.1 8B를 벤치마크한 결과, vLLM에서 363개의 요청을 동시에 배치 처리하면서 초당 총 약 6,200개의 토큰을 처리했습니다. 기업용 제품과 비교하면, H100은 약 2배 많은 동시 요청을 배치 처리할 수 있었으며, 608개의 동시 요청에서 초당 약 11,300개의 토큰을 처리했습니다.

GPU 구매 가격을 기준으로 정규화하면, 소비자용 칩의 더 낮은 전력 소비를 고려하지 않더라도 소비자용 GPU는 달러당 성능이 약 10배 향상되는 것으로 나타납니다. 이는 대규모 데이터센터가 생성되는 토큰당 비용이 가장 낮을 것으로 예상하는 전통적인 "규모의 경제" 직관과 상반됩니다.

핵심적인 이유는 시장 세분화입니다. Nvidia는 데이터센터가 소비자용 GPU를 대량으로 구매하는 것을 허용하지 않음으로써 게이머와 기타 최종 사용자를 위한 소비자 가격을 낮게 유지하고 있으며, 소비자용 GPU를 멀티 GPU 서버 구성에 통합하기 어렵게 만들고 있습니다. Dolphin Network는 이러한 격차를 해소하기 위해 존재합니다. 전 세

계에 분산된 유휴 소비자용 GPU를 하나의 신뢰할 수 있고 무허가형인 AI 컴퓨팅 클러스터로 풀링하여, 개발자와 사용자가 가장 인기 있는 오픈 모델에 저비용으로 접근할 수 있도록 합니다.

3 - 근접성

전통적인 DePIN 네트워크에서는 근접성이 가치를 결정합니다. 노드는 인근 사용자에게만 서비스를 제공할 수 있습니다.

반면, AI 추론은 대체로 위치에 구애받지 않습니다. 수백 밀리초의 지연 시간은 사용성에 거의 영향을 미치지 않습니다. 이를 통해 당사 네트워크는 소비자를 전 세계의 컴퓨팅 리소스와 연결할 수 있으며, 노드당 훨씬 더 높은 확장성과 활용도를 확보할 수 있습니다.

주요 이점

GPU 소유자

- GPU가 유휴 상태일 때 AI 모델을 실행하고 학습하여 네트워크에 컴퓨팅 자원을 제공
- 네트워크에 대한 상대적 기여도와 전반적인 영향에 따라 보상 획득
- 게임이나 동영상 렌더링과 같이 GPU를 많이 사용하는 작업을 수행하지 않을 때 언제든지 노드 소프트웨어를 백그라운드에서 실행 가능
- Windows와 Linux 모두 호환
- 겨울에 집을 따뜻하게 하면서 보상받기

AI 사용자

- 저렴한 가격으로 다양한 AI 모델과 상호작용
- 고급 보안 기능을 통해 컴퓨팅 제공자와 추론 소비자 간 신뢰를 최소화한 매칭 가능
- 노드에 대한 빈번한 벤치마킹을 통해 일관된 고성능과 가동 시간 보장
- 설계 단계부터 프라이버시 보장 - 노드는 로그를 활성화할 수 없으며, 사용자의 프롬프트에 포함될 수 있는 민감한 개인정보의 가로채기를 방지하기 위해 요청은 종단 간 암호화됨

노드 배포

노드는 소비자용 및 기업용 GPU에서 실행되거나, 검증된 추론을 위한 미들웨어를 통해 dePIN 네트워크(Lium, Akash 및 [IO.net](#) 등) 위에서 실행됩니다. 위치와 핑은 추론 워크로드에 미치는 영향이 미미하므로, 전 세계의 리소스를 하나의 강력한 엔드포인트로 통합하여 소비자와 OpenRouter와 같은 마켓플레이스에 제공할 수 있습니다.

최신 소비자용 GPU는 LLM 및 이미지 / 비디오 / 오디오 생성에 높은 처리량을 제공하며, 하드웨어 가격을 고려할 때 비용 효율성이 매우 높습니다.

분산형 웹 스크래핑

Dolphin Network는 GPU 추론을 넘어 CPU 기반 웹 스크래핑 노드까지 확장되며, 당사의 Peer-to-Pool 아키텍처를 활용하여 웹 데이터 수집을 대중화합니다. 사용자는 데스크톱 PC 또는 노트북에서 경량 스크래핑 노드 소프트웨어를 실행하여 유휴 CPU와 대역폭을 네트워크에 제공하는 동시에 POD 토큰을 획득할 수 있습니다.

CPU 스크래퍼 노드의 작동 방식

CPU 스크래퍼 노드는 당사의 GPU 추론 노드와 유사하게 작동하지만, 웹 크롤링, 렌더링 및 데이터 검색에 중점을 둡니다. Dolphin 스크래퍼 노드 소프트웨어를 설치하고 실행하면 사용자의 기기는 웹페이지를 불러오고, 콘텐츠를 추출하며, 다운스트림 AI 작업을 위해 정제되고 구조화된 데이터를 반환할 수 있는 글로벌 분산 네트워크의 일부가 됩니다.

가정용 IP의 이점

데이터센터 IP에 의존하는 기존 스크래핑 서비스와 달리, Dolphin 스크래퍼 노드는 전 세계 사용자의 실제 가정용 인터넷 연결을 활용합니다. 이는 다음과 같은 몇 가지 주요 이점을 제공합니다:

- **안티봇 회피:** 요청이 정상적인 가정용 IP 범위에서 발생하므로 자동화된 스크래핑이 아닌 일반 사용자 트래픽처럼 보입니다. 이를 통해 데이터센터 IP를 자주 차단하는 Cloudflare와 같은 서비스로 보호되는 콘텐츠에 대한 접근성이 향상됩니다.
- **지리적 분산:** 여러 국가와 지역에 분산된 노드를 통해 현지화된 콘텐츠 접근, 지역별 데이터 검색 및 특정 지역의 IP 차단에 대한 복원력을 확보할 수 있습니다.
- **동적 콘텐츠 처리:** 각 노드는 JavaScript 실행, 싱글 페이지 애플리케이션 처리 및 실제 사용자가 보는 것과 동일하게 페이지를 렌더링할 수 있는 완전한 브라우저 엔진을 실행하여, 최신 동적 웹사이트에서 완전한 데이터 수집을 보장합니다.

노드 아키텍처

스크래퍼 노드는 Dolphin Network의 기존 인프라와 원활하게 통합됩니다:

- **작업 할당:** Lighthouse 라우터는 지리적 요구사항, 연결 품질 및 현재 부하를 기반으로 사용 가능한 CPU 노드에 스크래핑 작업을 할당합니다. 작업은 특정 URL을 대상으로 하거나, 사이트의 업데이트를 모니터링하거나, 지역별 데이터 수집을 수행할 수 있습니다.
- **헤드리스 브라우저 실행:** 노드는 대상 페이지를 로드하고, JavaScript를 실행하며, 동적으로 렌더링된 요소를 포함한 전체 페이지 콘텐츠를 캡처하는 격리된 브라우저 환경을 실행합니다. 브라우저는 개인 브라우징 데이터와 분리된 샌드박스 환경에서 작동합니다.

- **데이터 처리:** 캡처된 페이지는 Dolphin의 데이터 파이프라인으로 전송되기 전에 로컬에서 압축 및 처리됩니다. 시스템은 관련 콘텐츠를 추출하고, 불필요한 요소를 제거하며, AI 분석 또는 저장을 위해 데이터를 준비합니다.
- **보상 계산:** 노드는 제공한 대역폭, 완료한 작업, 연결 안정성 및 가동 시간을 기반으로 POD 토큰을 획득합니다. GPU 노드와 마찬가지로 보상은 네트워크 수요와 노드 성능에 따라 증가합니다.

개인정보 보호 및 보안

스크래퍼 노드 소프트웨어는 개인 데이터와 완전히 격리된 상태에서 작동합니다:

- 모든 스크래핑 활동은 별도의 샌드박스 브라우저 환경에서 실행됩니다.
- 개인 브라우징 데이터, 쿠키 및 자격 증명에는 절대 접근하지 않습니다.
- 노드 운영자는 어떤 페이지가 스크래핑되고 있는지 볼 수 없으며, 해당 프로세스는 백그라운드에서 투명하게 실행됩니다.

Dolphin Deep Research

Dolphin Deep Research는 당사의 분산형 웹 스크래핑 네트워크와 Tongyi-DeepResearch 모델(선도적인 오픈 소스 Deep Research Agent)을 결합하여 온디맨드 방식으로 엔드투엔드 리서치 워크플로우를 제공합니다.

사용자가 리서치 쿼리를 제출하면 Dolphin Deep Research는 다음을 수행할 수 있습니다:

- 지리적으로 분산된 가정용 노드를 사용하여 오픈 웹 전반에서 소스 탐색
- 당사의 헤드리스 브라우저 스택을 사용하여 페이지(JavaScript를 많이 사용하는 사이트 포함)를 수집하고 렌더링
- 콘텐츠 추출 및 정규화(상용구 제거, 핵심 섹션 분리, 구조 보존)
- Tongyi-DeepResearch를 사용하여 다단계 심층 리서치 수행: 계획, 반복 검색, 읽기, 종합 및 교차 검증
- 검색된 콘텐츠에 기반하여 요약, 비교, 타임라인, 장단점 및 근거 기반 답변과 같은 구조화된 결과물 생성

워크로드가 Dolphin의 네트워크 전반에 분산되기 때문에, Deep Research는 여러 지역과 IP 범위에 대한 폭넓은 접근성을 유지하면서 대규모 다중 소스 조사로 확장할 수 있습니다.

Web Chat, Bot 및 API 전반의 URL 인식 AI

Dolphin Web Chat 인터페이스, Telegram + Discord 봇 및 Dolphin API의 Dolphin AI 모델은 사용자 메시지에 포함된 URL을 자동으로 감지합니다.

URL이 감지되면 Dolphin은 다음을 수행합니다:

- 당사의 분산형 스크래핑 네트워크를 통해 웹페이지를 스크래핑(동적 사이트의 전체 렌더링 포함)

- 관련 콘텐츠 추출(본문, 핵심 섹션, 유용한 경우 메타데이터)
- 해당 콘텐츠를 모델의 컨텍스트에 직접 주입하여, AI가 페이지에 실제로 무엇이 적혀 있는지를 완전히 인식한 상태에서 링크에 관한 질문에 정확하게 답변할 수 있도록 함

이를 통해 단순한 "여기 링크가 있습니다"라는 메시지가 강력한 워크플로우로 전환됩니다. 사용자는 콘텐츠를 수동으로 복사/붙여넣기하지 않고도 URL을 입력하고 즉시 요약, 비평, 사실 추출, 비교 또는 후속 분석을 요청할 수 있습니다.

사용 사례

분산형 스크래핑 네트워크는 광범위한 데이터 수집 및 인텔리전스 요구사항을 지원합니다:

- 실시간 웹 모니터링 및 콘텐츠 변경 감지
- 지역별 콘텐츠 접근 및 현지화 데이터 수집
- 사용자 쿼리를 위한 심층 리서치(Dolphin Deep Research)
- AI 모델 개발을 위한 학습 데이터 수집
- 시장 조사 및 경쟁 인텔리전스
- 뉴스 집계 및 트렌드 분석
- 여러 지리적 위치에서의 분산형 가동 시간 모니터링
- 채팅 앱 및 API를 통한 링크 기반 Q&A(자동 URL 컨텍스트 검색)

글로벌 노드 네트워크의 유휴 CPU 리소스를 활용함으로써 Dolphin은 기존 프록시 서비스보다 훨씬 낮은 비용으로 웹 스크래핑 및 리서치 기능을 제공하는 동시에, 안티봇 시스템으로 보호되는 콘텐츠에 대한 더 우수한 접근성을 제공할 수 있습니다. 당사의 GPU 추론 네트워크를 구동하는 것과 동일한 Peer-to-Pool 경제 모델은 노드 운영자에 대한 공정한 보상과 수요 증가에 따른 지속 가능한 성장을 보장합니다.

Dolphin Network 로드맵

로드맵 요약

- **당사 추론 네트워크 V1:** 10월에 출시되었으며, 3090 / 4090 / 5090과 같은 소비자용 GPU를 대상으로 합니다.
- **당사 추론 네트워크 V2:** 4~6주 내 출시되며, 기업용 GPU 처리량을 위한 배칭과 사용자가 당사 노드에서 실행되는 모델을 자신의 애플리케이션에 통합할 수 있는 공개 API를 제공합니다.
- **이미지 및 비디오 생성:** 약 12주 내에 인기 모델에 저비용으로 접근할 수 있도록 제공됩니다.
- **오디오 생성:** 약 14주 내에 제공되며, 다양한 최신(SOTA) 오디오 모델을 제공하는 ElevenLabs보다 훨씬 저렴한 대안입니다.
- **샤딩 추론:** 약 8주 내에 제공되며, 하나의 대형 모델을 여러 소비자용 GPU 노드에 분할합니다.

- **분산형 LoRA 학습:** 약 12~16주 내에 소비자용 및 기업용 GPU에서 제공됩니다. LoRA는 통신 효율성이 매우 높은 모델 학습 방식이며, 당사의 분산형 학습 구현을 위한 논리적인 첫 단계입니다. 당사의 최신 블로그 게시물 <https://blog.dphn.ai/405b/>에서는 단일 노드에서 LoRA 학습을 사용하여 Meta의 가장 큰 AI 모델인 Llama3.1 405B를 파인튜닝한 방법을 다룹니다.

노드 무결성 검증

탈중앙화 AI 네트워크가 직면하는 주요 문제는 악의적인 노드가 시스템을 속이고 네트워크의 다른 운영자보다 부당한 이점을 얻기 위해 노드 소프트웨어를 수정하는 것을 방지할 수 있어야 한다는 것입니다.

여기에는 다음과 같은 목적으로 노드 소프트웨어를 수정하는 것이 포함될 수 있습니다...

- 각 요청에 대해 실제로 추론을 실행하지 않고 대신 사전에 생성된 텍스트를 전송
- 더 작은 모델이나 더 높은 수준의 양자화를 사용하여 처리 속도를 높임으로써 추론 사용자의 응답 품질 및 예측 가능성을 저하시킴

Dolphin Network는 네트워크 내 사기를 방지하기 위해 소프트웨어 무결성 검사, 샘플링된 추론 검증, 검증자 점수 평가 및 암호경제학적 본드 예치를 결합하며, 이를 통해 추론 소비자는 자신이 선택한 AI 모델을 명시된 품질 수준으로 이용하고 있다는 확신을 가질 수 있습니다.

Dolphin Network 위에 구축 중인 제품

게임 개발자를 위한 DolphinPod Inference SDK 라이브러리

Dolphin Gemma 2 2B 모델과 같은 소형 모델은 거의 모든 최신 GPU에 탑재될 수 있으며, 게임 내 NPC와의 대화 또는 게임 내 AI의 기타 사용 사례를 위한 추론 처리에 사용될 수 있습니다.

LLM 모델은 "시스템 프롬프트(system prompt)"를 따르도록 되어 있습니다. 이는 모델에 대한 모든 요청 앞에 위치하는 텍스트로, 모델이 어떻게 응답하고 무엇을 해야 하는지를 지시하거나, OpenAI 또는 Claude의 경우에는 무엇을 하지 말아야 하는지를 지시합니다.

당사 모델을 사용하는 이점은 시스템 프롬프트를 면밀하게 따르고, 기본 모델에 원래 학습되어 있던 가드레일과 편향을 제거하도록 학습되었다는 점입니다.

이는 중앙화된 LLM API에 의해 검열될 수 있는 캐릭터 특성이 요구되는 롤플레이нг 게임에 당사 모델이 매우 적합하도록 합니다.

모든 게임 개발자는 당사의 SDK 라이브러리를 자신의 프로젝트에 통합하여 NPC 및 게임의 기타 요소가 당사의 작고 매우 빠른 0.5B~4B 모델로 구동되도록 할 수 있습니다.

게임은 일반적으로 그래픽을 위해 사용 가능한 모든 VRAM을 필요로 하기 때문에, 대부분의 게임 사용 사례에서는 플레이어의 GPU에서 AI 모델을 로컬로 실행하는 것보다 당사의 API를 사용하는 것이 더 낫다고 생각합니다.

합성 데이터셋 생성 제품군

선도적인 폐쇄형 소스 모델(예: ChatGPT 및 Claude)을 사용하여 데이터셋을 생성하는 것은 매우 비쌉니다. 당사의 한 고객은 아랍어로 파인튜닝된 독점 모델을 요청했으며, 파인튜닝에 필요한 합성 데이터를 생성하기 위해 고객의 GPT-4 API 키를 사용할 수 있도록 해주었음에도 결국 약 1만 달러를 지출했습니다.

Dolphin Network는 유허 네트워크 용량을 활용하여 긴급하지 않은 워크로드의 토큰당 비용을 줄이는 합성 데이터 생성 제품군을 제공할 것입니다.

Dolphin Qwen3 235B와 같이 당사가 학습한 최신 세대의 대형 모델은 매우 높은 품질의 합성 데이터셋을 생성할 수 있습니다.

이는 모델 학습, 소설 또는 기타 사용 사례를 위해 검열되지 않은 데이터셋을 생성하려는 사람들에게 특히 유용합니다.

이는 Dolphin 이외의 다른 곳에서는 할 수 없는 일을 당사가 가능하게 하는 것입니다.

예시:

- 특정 사안의 모든 측면에 대한 입장을 취하는 데이터셋을 원하는 경우.
- 영어 이외의 언어로 검열되지 않은 데이터를 생성하고자 하는 경우(Dolphin Qwen은 영어와 중국어 외에 27개의 추가 언어를 지원합니다).
- 각 플레이어에게 독특한 경험을 제공하기 위해 게임의 테마와 일치하는 스타일로 게임 내 NPC 대화를 위한 대화형 데이터를 생성하고자 하는 경우(야만적인 중세 롤플레이нг 게임을 위한 대화형 데이터를 생성한다고 상상할 수 있습니다... ChatGPT 또는 Claude를 통해 이를 시도한다면 제한될 가능성이 높습니다).

모델 생성 및 파인튜닝 제품군

사용자의 데이터 또는 당사의 데이터 템플릿을 혼합하여 SFT 및 RL을 지원하고, 당사 네트워크를 기반으로 학습 작업을 제출할 수 있는 사용자 친화적인 인터페이스입니다.

사용자는 자신의 데이터셋을 업로드하거나, 당사가 선별한 옵션 중에서 선택하거나, 두 가지를 결합할 수 있습니다.

주요 목표는 주요 클라우드 제공업체보다 낮은 비용으로 학습 인프라를 제공하는 동시에, 사전 경험이 없는 사용자도 맞춤형 데이터로 모델을 학습할 수 있도록 하여 노코드 시장에도 서비스를 제공하는 것입니다.

당사는 Dolphin Network의 유허 컴퓨팅으로 구동되는 Tinker와 유사한 경험을 구현하여 경쟁사보다 상당히 낮은 비용으로 학습을 제공하고자 합니다.

Dolphin Dev

일회성 환경 내부의 도구를 활용한 문제 해결에 초점을 맞춘 데이터 생성 시스템입니다. 대부분의 오픈소스 데이터셋은 계획, 도구 사용, 검증 및 반복보다는 패턴 매칭을 학습시킵니다. Dolphin Dev는 LLM을 샌드박스 환경에 배포하여 도구를 활용해 과제를 해결하도록 함으로써 이 문제를 해결하며, 기존 모델 라인을 개선하고 자율형 코딩 제품을 위한 기반을 마련하는 데이터셋을 생성합니다.

이를 통해 당사의 모델은 Cline, Cursor, Roo 및 이와 유사한 제품군과 같은 고추론 코딩 애플리케이션에서 높은 성능을 발휘할 수 있게 됩니다.

이러한 데이터셋을 활용하면 작업자가 배포한 컨테이너 내부에서 자율적으로 실행되어 작업을 완료하는 전용 코딩 모델을 학습할 수 있습니다. 향후 웹 인터페이스를 통해 작업을 제출할 수 있게 되며, 작업이 완료되면 작업자는 결과물과 로그를 반환합니다.

작업자 호스팅 샌드박스(옵트인)

Dolphin Network의 노드는 이러한 작업을 위한 일회성 샌드박스를 호스팅하기 위해 RAM/CPU의 일부를 할당할 수 있습니다. 샌드박스와 함께 배치된 LLM은 로컬에서 반복 작업을 수행합니다. 즉, 계획 → 도구 호출 → 테스트 실행 → 개선 과정을 거치며, 최종 결과물과 로그만 고객에게 다시 전송됩니다. 이를 통해 지연 시간을 줄이고, 분당 반복 횟수를 늘리며, GPU와 함께 유휴 CPU/RAM을 수익화할 수 있습니다.

이점

- 작업자 노드와 함께 작업을 실행하여 GPU뿐만 아니라 사용 가능한 RAM/CPU도 수익화합니다.
- 자율 실행을 위한 프리미엄 등급을 제공합니다.
- 작업을 사용량이 적은 시간대로 이동하여 네트워크 전반의 수요를 평준화합니다.

P2P AI 개발 마켓플레이스

파인튜닝, 합성 데이터 생성 및 P2P 추론을 위한 **양면 마켓플레이스**입니다. 각 지갑에는 거래 후 업데이트되는 **소울 바운드 평판 NFT**가 포함됩니다.

- **역할:** 데이터셋 생성자 / 데이터셋 소비자
- **역할:** 모델 트레이너 / 모델 학습 소비자
- **역할:** 추론 제공자 / 추론 소비자

이는 새롭게 등장하는 데이터 및 모델 생태계 전반에서 P2P 관계를 조정합니다.