

베니스 에이아이(Venice AI)

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.

Venice AI - 공식 문서 모음

1. 베니스에 오신 것을 환영합니다. (Welcome to Venice)

출처: <https://venice.ai/blog/welcome-to-venice>

베니스(Venice) - 고요한 공화국(The Serene Republic)

베니스는 비기술 사용자를 위한 생성형 AI 애플리케이션이다. 이는 Claude, Perplexity 또는 ChatGPT의 빠른 버전처럼 느껴지지만, 오웰식 요소는 배제되어 있다(아마도 이것이 빠른 이유일 수 있다).

버튼 한 번으로 전 세계의 정보를 검색하고, 풍부한 대화를 나누며, 문서를 분석하고, 다양한 스타일의 흥미로운 이미지와 예술을 생성할 수 있다. 이러한 기능은 다양한 AI 서비스에서도 제공되는데, 그렇다면 베니스의 차별점은 무엇인가?

베니스는 사용자를 감시하지 않는다

베니스는 AI를 검열하지 않는다

즉, 베니스는 프라이버시를 보장하며, 허가 불필요한(permissionless) 구조를 갖는다.

각 오픈소스 AI 모델은 고유의 경계와 규칙을 가지고 있다는 점에 유의해야 한다. 그러나 중요한 것은 베니스가 사용자에게 선택권을 제공한다는 점이다. 더 나아가 Pro 사용자는 시스템 프롬프트(System Prompt)를 직접 수정할 수 있어, 상호작용하는 AI의 성격을 스스로 정의할 수 있다.

베니스는 어떻게 작동하는가?

베니스는 선도적인 오픈소스 AI 모델(특히 Dolphin을 선호함)을 활용하여 텍스트, 코드, 이미지 생성을 웹 브라우저 또는 모바일 애플리케이션에서 제공한다.

다운로드가 필요 없다. 별도의 설치도 필요 없다. 기본적인 사용의 경우 계정도 필요 없으며, 서비스는 무료로 제공된다.

기술 사용자들은 이미 생성형 AI를 위해 오픈소스 도구를 활용해 왔지만, HuggingFace를 처음 접했을 때 다소 부담을 느끼는 사용자라면 베니스가 적합한 선택이 될 수 있다.

베니스의 프론트엔드는 ChatGPT를 사용해 본 사용자라면 누구에게나 익숙하게 느껴질 수 있는 깔끔한 웹 애플리케이션이다.

베니스 애플리케이션은 단순하다. 그러나 중요한 차이가 있다...

모든 기업은 사용자 프라이버시를 존중한다고 말한다. 이는 사실과 다르다. 사용자 프라이버시를 존중하는 유일한 방법은 애초에 이를 침해하지 않는 것이다. 기업이 사용자의 데이터를 보유하고 있다면, 이미 프라이버시는 상실된 것이다.

주요 AI 기업들은 사용자의 전체 대화 기록을 저장하고 이를 사용자의 신원과 영구적으로 연결한다. 더 나아가, 이 데이터는 광고주, 해커, 그리고 무엇보다 위험한 정부 등 다양한 제3자에게 전달된다.

암호화폐 분야에 익숙한 사람이라면 신뢰 최소화 아키텍처(trust-minimizing architecture), 클라이언트 측 설계(client-side design), 종단 간 암호화(end-to-end encryption)와 같은 개념을 잘 알고 있을 것이다. 이러한 설계 패턴은 강력한 프라이버시 문화와 사용자 및 자산 보호의 필요성에서 출발하여 암호화폐 영역에서 발전해왔다. 베니스는 이러한 보호 패턴을 생성형 AI에 적용한다.

- 사용자의 대화 기록은 오직 사용자의 브라우저에만 저장된다. 베니스는 어떠한 프롬프트나 모델 응답도 서버에 저장하거나 기록하지 않는다.
- 사용자의 추론 요청(inference request, 사용자가 전송하는 메시지)은 암호화된 상태로 프록시 서버를 통해 탈중앙화된 연산 자원으로 직접 전달된다.
- AI의 응답 또한 동일하게 암호화된 프록시 서버를 거쳐 사용자의 브라우저로 직접 스트리밍되며, 사용자의 브라우저 외의 어떤 곳에도 저장되지 않는다.
- 추론 요청을 처리하는 GPU는 여러 탈중앙화된 제공자로부터 공급되며, 각 개별 서버는 특정 대화의 텍스트만 확인할 수 있을 뿐 전체 대화 기록이나 사용자 신원을 알 수 없다.

그 결과는 다음과 같다:

- 베니스가 아는 정보: 이메일과 IP 주소 (대화 내용은 포함되지 않음)
- 연산 제공자가 아는 정보: 특정 대화 내용 (이메일과 IP 주소는 알 수 없으며, 특정 대화를 특정 사용자와 연결할 수 없음)

완전한 프라이버시는 동형암호(FHE, Fully Homomorphic Encryption)가 구현되거나(추후 가능) 모델을 로컬에서 실행할 때 비로소 달성될 수 있다. 그러나 현재 시점에서 감시나 검열을 원하지 않는 사용자에게 베니스의 아키텍처는 기존의 어떤 호스팅형 AI 서비스보다 실질적으로 우수한 대안이다.

베니스의 프라이버시 아키텍처에 대한 자세한 내용은 여기에서 확인할 수 있다.

또한 사용자 대화에 스파이웨어나 로깅을 추가하지 않기 때문에 애플리케이션의 응답 속도는 매우 빠르게 유지된다.

프라이버시는 베니스의 첫 번째 핵심 차별점이다. 이제 두 번째, 검열에 대해 살펴보자

주요 AI 애플리케이션을 사용해 본 사람이라면 누구나 기이하고 불편하며 과도하게 보호적인 검열을 경험했을 것이다. 그리고 이러한 경향은 점점 심화되고 있다. 사용자는 AI와 상호작용하고 있는 것인가, 아니면 수십억 달러 규모의 편향 시뮬레이터와 상호작용하고 있는 것인가?

해결책은 기술 기업들에게 특정한 방식으로 행동하도록 강제하는 것이 아니다. 해결책은 대안을 구축하고 이를 겸손하게 제공하는 것이다.

베니스는 사용자를 주권적 개인으로 존중하며, 프라이버시와 표현의 자유가 단순한 인권을 넘어 문명의 발전에 필수적이라고 본다. 심지어 수동적인 감시조차도 사고와 행동에 위축 효과를 초래한다는 점이 입증되어 있다. 인간과 기계의 사고가 점점 융합되는 시대에, 이러한 관계의 경계를 정의하는 주체는 국가나 기업이 아니라 사용자 자신이어야 한다.

이러한 철학은 오직 개방적이고 허가 불필요한 아키텍처 위에서만 구현될 수 있다.

2. 베니스 AI 소개(About Venice AI)

출처: <https://venice.ai/about>

베니스는 강력한 기계 지능이 이를 사용하는 개인의 주권을 존중할 때 문명에 가장 바람직하게 기여한다는 원칙 위에서 설립되었다. 따라서 기본적으로 프라이버시가 보장되어야 하며, 자유롭고 개방적인 사고를 허용해야 하고, 세계 최고 수준의 오픈소스 기술을 기반으로 구축되어야 한다. 베니스에서는 사용자의 대화와 창작물이 오직 사용자 본인에게만 귀속되며, 기업에도, 정부에도, 그리고 베니스에도 속하지 않는다. 이러한 점에서 베니스는 동종 서비스 가운데 독보적인 위치를 가진다.

우리의 프라이버시 아키텍처

사용자의 모든 콘텐츠(프롬프트, 응답, 이미지, 문서 업로드)는 로컬 브라우저에서 암호화되며, 베니스의 서버에는 어떠한 형태로도 저장되지 않는다.

암호화(Encryption)

사용자의 추론 요청(inference request, 사용자가 전송하는 메시지)은 암호화되어 프록시 서버를 통해 탈중앙화된 연산 자원으로 직접 전달된다.

전달(Circulation)

AI의 응답 또한 동일하게 암호화된 프록시 서버를 통해 사용자의 브라우저로 직접 스트리밍되며, 브라우저 외의 어떤 환경에도 저장되지 않는다.

탈중앙화(Decentralization)

추론 요청을 처리하는 GPU는 여러 탈중앙화된 제공자로부터 공급되며, 각 개별 제공자는 특정 대화 하나의 텍스트만 확인할 수 있을 뿐 전체 대화 기록을 볼 수 없고 사용자 신원 또한 알 수 없다.

편향 없는 AI(Unbiased AI)

베니스는 사용자의 창의적 자유를 제한하지 않으면서 강력한 오픈소스 AI 모델을 활용한다.

이념적 편향(Ideological bias)

주요 AI 기업들은 선의로 엄격한 콘텐츠 정책을 적용하고 있지만, 이러한 제한과 이념적 편향은 오히려 AI의 성능을 저하시켜 맥락에 맞는 정확하고 완전한 정보를 제공하지 못하게 만든다. 많은 경우, 그 결과는 이러한 정책이 방지하고자 하는 위험보다 더 큰 문제를 초래한다.

호기심의 제한(Restricting curiosity)

단순히 부적절하거나 무의미한 응답을 제공하는 수준을 넘어, 이러한 제한은 사용자가 특정 질문 자체를 하지 말아야 한다는 식의 훈계조 응답을 유도한다. 이는 AI가 장려해야 할 핵심 가치인 호기심, 창의성, 그리고 이념적 제약으로부터 자유로운 독립적 사고를 저해한다.

3. 베니스에서 '검열 없음'은 실제로 무엇을 의미하는가(What does “uncensored” really mean in Venice)

출처: <https://venice.ai/blog/what-does-uncensored-really-mean-in-venice>

베니스는 주요 AI 애플리케이션에 대한 프라이버시 중심이면서 검열이 없는 대안이다. 그렇다면 여기서 말하는 '검열 없음(uncensored)'이란 정확히 무엇을 의미하는가?

AI 산업의 콘텐츠 정책은 점점 문제점이 드러나고 있다.

“안전한” AI를 만들기 위한 시도 속에서 주요 기술 기업들은 과도한 콘텐츠 제한을 도입해 왔으며, 이는 정상적인 탐구와 논의를 저해하고 종종 의도치 않게 부자연스럽고 제한적인 결과를 초래한다.

베니스는 이에 대해 다른 접근 방식을 취한다. 베니스는 플랫폼에서 제공하는 오픈소스 모델에 내재된 범위를 넘어서는 추가적인 콘텐츠 제한을 부과하지 않는다.

베니스에서 '검열 없음'이란 기본 모델에 내재된 범위를 초과하는 별도의 제한을 추가하지 않는다는 의미이다.

베니스는 추가적인 제한 없이 강력한 오픈소스 모델을 활용한다.

현재 베니스는 다음 두 가지 모델을 지원한다:

- Nous Hermes 2 Theta (Llama3)
- Dogge 70B (Llama3)

모든 AI 모델에서 불가피하게 나타나는 편향의 주요 유형은 다음과 같다:

- **훈련 데이터의 한계(Training data boundaries):** 모델은 학습 데이터의 시점 이후 정보나 학습 범위를 벗어난 내용을 생성할 수 없다. 무엇을 포함하고 제외할지는 인간이 결정한다.
- **내재적 편향(Inherent biases):** 학습 데이터에 포함된 문화적, 역사적, 언어적 편향이 반영될 수 있다.
- **윤리적 안전장치(Ethical safeguards):** 극단적으로 폭력적인 콘텐츠 생성을 거부하는 등 기본적인 윤리 기준이 포함될 수 있다.
- **언어적 한계(Language limitations):** 사용 빈도가 낮은 언어, 속어, 방언 등에 대해서는 처리 능력이 제한될 수 있다.
- **사실적 한계(Factual limitations):** 이러한 모델은 오류를 생성하거나 최신이 아닌 정보를 제공할 수 있다.

이러한 내재적 한계를 이해하는 것은 중요하다. 베니스는 강력하면서도 최소한의 제한만 적용된 AI를 제공하지만, 사용자는 항상 그 결과를 비판적으로 검토하고 중요한 정보는 신뢰할 수 있는 출처를 통해 확인해야 한다.

4. 베니스의 새로운 모델 패러다임(Venice's New Model Paradigm)

출처: <https://venice.ai/blog/venice-new-model-paradigm>

베니스는 모델 선택 체계를 간소화하고 재정비하여, LLM 목록을 정제된 형태로 구성함으로써 모델 난립과 중복을 줄이고 각 모델이 어떤 작업에 적합한지 명확히 제시한다.

또한 사용자가 자신의 목적에 맞는 모델을 보다 쉽게 선택할 수 있도록 단순화된 선택 방식을 도입하고, 다섯 가지의 명확한 모델 카테고리를 제시한다.

다섯 가지 모델 카테고리: 검열 없음, 추론, 소형, 중형, 대형

각 모델은 명확한 목적을 가진다. 대형 모델은 더 높은 성능을 제공하고, 소형 모델은 더 빠른 응답 속도를 제공한다. 기술적인 세부 정보를 확인하고자 하는 사용자를 위해 각 모델의 HuggingFace 원본 페이지 링크를 제공하지만, 핵심 목표는 AI에 익숙하지 않은 일반 사용자도 자신의 사용 목적에 적합한 모델을 쉽게 이해하고 선택할 수 있도록 하는 데 있다. 또한 모든 모델에는 웹 검색 기능이 제공되며, 사용자가 이를 켜거나 끌 수 있다.

베니스의 새로운 모델

LLM 목록을 베니스 검열 없음, 베니스 추론, 베니스 소형, 베니스 중형, 베니스 대형의 다섯 가지 카테고리로 구성함으로써, 사용자는 자신의 필요에 가장 적합한 모델을 더욱 쉽게 선택할 수 있다.

Dolphin Mistral 24B Venice Edition과 Llama 4 Maverick 모델의 도입은 성능과 기능을 향상시키며, 기존 구형 모델의 제거는 중복을 해소하고 베니스 인프라의 확장성을 높여 사용자에게 더 높은 처리량을 제공한다.

베니스가 지속적으로 발전하고 개선됨에 따라, 이러한 구조적 정리는 전반적인 사용자 경험을 향상시키고 오픈소스 AI 모델과의 상호작용을 더욱 빠르고 효율적으로 만든다.

다른 AI 플랫폼과 달리, 베니스는 사용자의 대화를 저장하거나 상호작용을 추적하지 않는다. 사용자의 AI 경험은 완전한 프라이버시 하에서 검열 없이 유지된다.

5. 메모리아(Memoria): 기술 개요(Memoria: A Technical Overview)

출처: <https://venice.ai/blog/venice-memoria-technical-overview>

베니스는 사용자의 과거 대화 맥락을 기억할 수 있도록 하는 프라이버시 보호 메모리 시스템인 메모리아(Memoria)를 도입했다. 기존의 클라우드 기반 메모리 시스템과 달리, 메모리아는 모든 데이터를 사용자의 브라우저에 로컬로 저장하여 어떠한 데이터도 사용자 기기를 벗어나지 않도록 한다.

메모리아의 작동 방식

메모리아는 벡터 임베딩(vector embedding)을 활용하여 관련 정보를 이해하고 검색한다. 사용자가 베니스와 대화를 나눌 때 다음과 같이 작동한다:

- 대화 중: 사용자의 메시지는 의미를 반영한 수학적 표현(벡터)으로 변환된다
- 자동 추출: 일정 주기마다 베니스는 대화에서 핵심 정보와 인사이트를 추출한다
- 맥락 기반 검색: 새로운 대화를 시작하면, 메모리아는 관련된 과거 정보를 검색하여 AI에 맥락으로 제공한

다

이 과정은 AI가 이전에 논의된 내용을 참조하고, 사용자의 선호를 반영하며, 과거 대화를 기반으로 상호작용을 이어갈 수 있도록 하여 보다 개인화된 경험을 제공한다.

어떤 데이터가 저장되는가?

메모리아는 다음 데이터를 사용자의 브라우저에 로컬로 저장한다:

Data Type	Description	Example
메모리 텍스트	대화에서 추출된 핵심 요약	"사용자는 Python 프로그래밍을 학습 중"
벡터 임베딩	1024차원의 수학적 표현	메모리당 약 1.4KB로 압축
희소 토큰	하이브리드 검색을 위한 키워드	["python", "programming", "learning"]
출처	메모리 생성 출처	"venice"(자동 추출), "manual" 또는 파일 식별자
메모리 유형	메모리 분류	"extracted_summary", "user"
중요도 점수	관련성 평가(1~10)	값이 낮을수록 중요도가 높음
타임스탬프	생성 및 접근 시각	ISO 형식 날짜 문자열

프라이버시 보장

- 로컬 전용 저장: 모든 메모리 데이터는 브라우저의 IndexedDB에 저장되며, 명시적으로 공유하지 않는 한 사용자 기기를 벗어나지 않는다.
- 벡터 솔팅(Vector salting): 메모리의 수학적 표현조차 사용자별 암호화 키에서 파생된 암호학적 솔트를 적용해 변환된다. 이는 다음을 의미한다:

사용자의 임베딩은 사용자마다 고유하다

다른 사용자 데이터와 연결될 수 없다

가로채더라도 복원할 수 없다

- 서버 측 비저장: 베니스 서버는 임베딩을 일시적으로 생성할 뿐이며, 메모리나 이를 생성하는 데 사용된 데이터를 저장하지 않는다.
- 익명화 모델 보호: 사용자는 제3자("익명화된") 모델과의 메모리 공유를 비활성화할 수 있으며, 이러한 모델은 베니스의 프라이빗 모델과 서로 다른 프라이버시 보장 수준을 가질 수 있다.

6. 베니스 토큰(VVV) 소개(Introducing the Venice token: VVV)

출처: <https://venice.ai/blog/introducing-the-venice-token-vvv>

베니스 토큰(VVV)은 사용자, 개발자, AI 에이전트가 토큰을 스테이킹함으로써 요청당 비용 없이 Venice API를 통해 지속적인 프라이버시 중심의 비검열 AI 추론(inference)에 접근할 수 있도록 하는 새로운 AI 인프라 모델을 구현한다.

지난 11월 Venice API 출시를 기반으로, 베니스는 스테이커가 API를 통해 한계비용 없이 지속적으로 AI 추론에 접근할 수 있도록 하는 새로운 토큰 이코노믹스 구조를 도입했다.

VVV는 주로 에이전트를 위해 설계되어 경제적 마찰을 최소화하는 것을 목표로 한다. 에이전트는 VVV를 스테이킹함으로써 요청당 비용 없이 대규모 텍스트, 이미지, 코드 생성을 위한 지속적인 추론 자원을 활용할 수 있다.

전체 개요(The Full Story)

Venice.ai는 2024년 5월, ChatGPT의 프라이버시 중심이면서 비검열 대안으로 출시되었다.

AI가 일상 전반에 빠르게 통합되는 환경에서, 인간이 중앙화된 감시와 통제 없이 기계 지능과 상호작용할 수 있는 것은 매우 중요한 요소이다. ChatGPT(또는 Gemini, Anthropic, Perplexity, Grok 등)에 입력되는 모든 메시지는 영구적으로 저장되며, 검토·분류·판매·해킹·정부 제공 등의 대상이 될 수 있다.

Venice API와 에이전트의 부상

2024년 11월, 베니스는 Pro 사용자를 대상으로 API 베타 버전을 출시했다. 해당 API는 애플리케이션에서 제공되는 텍스트, 이미지, 코드 생성 기능을 그대로 제공하며, 제3자 애플리케이션, 개발자, AI 에이전트를 위해 설계되었다. 이와 동시에 에이전트라는 새로운 사용자군이 본격적으로 등장하기 시작했다.

에이전트는 다양한 수준의 자동화와 기능을 갖춘 AI 개체(일종의 봇)이다. 일부는 단순 자동화 스크립트를 'AI'로 재포장한 수준에 불과하지만, 일부는 실질적인 기술적 진보를 보여준다. 암호화폐 시장과 유사하게 노이즈와 저품질 프로젝트가 혼재되어 있지만, 그 이면의 핵심 기술적 흐름은 매우 혁신적이다.

에이전트가 소비하는 핵심 자원은 추론(inference)이다. 이는 기계의 연산 능력에 해당한다.

현재 에이전트는 주로 OpenAI 등 중앙화된 API를 통해 추론 자원을 활용하고 있으며, 다음과 같은 세 가지 문제가 존재한다:

1. 모든 대화, 이미지 및 데이터가 영구적으로 저장됨(감시)
2. 콘텐츠가 통제되며 정책이 불투명함(검열)
3. 요청당 과금 구조이며 결제가 허가 기반으로 이루어짐(마찰)

베니스의 애플리케이션과 API는 이미 첫 번째와 두 번째 문제를 해결했다.

세 번째 문제와 관련하여, 만약 1990년대 이메일 전송마다 비용이 발생했다면 이메일은 현재와 같은 확산을 이루지 못했을 것이다. 사용자는 비용 부담 없이 메시지를 생성하고 전송할 수 있어야 한다. 베니스는 에이전트의 추론 비용을 0 수준으로 낮춘다. VVV를 스테이킹하면 추가 비용 없이 추론을 사용할 수 있다.

에이전트 및 기타 Venice API 사용자는 이제 암호화 토큰을 전체 추론 용량에 대한 접근 권한으로 활용할 수 있다. 베니스 토큰(VVV)을 스테이킹한 사용자는 API를 통해 전체 추론 용량 중 자신의 지분에 해당하는 비율만큼을 지속적으로 사용할 수 있다.

예를 들어 전체 VVV의 1%를 스테이킹하면, 베니스 API 전체 용량의 1%를 지속적으로 사용할 수 있으며 요청당 비용은 발생하지 않는다.

VVV는 실제로 어떻게 작동하는가?

중요한 점은 베니스 애플리케이션 및 API의 기존 사용 방식은 그대로 유지되며, 토큰 사용이 필수는 아니라는 것이다. 추가적인 사용 장벽도 존재하지 않는다.

새롭게 도입된 점은, 개인·애플리케이션·에이전트 등 어떤 주체든 VVV를 스테이킹하면 일일 한도 내에서 Venice API를 통해 텍스트, 이미지, 코드 추론을 추가 비용 없이 사용할 수 있다는 것이다. 요청당 과금은 발생하지 않는다.

사용 가능한 추론량은 전체 용량 대비 스테이킹 비율에 따라 결정된다. 예를 들어 전체 스테이킹된 VVV의 1%를 보유한 경우, 베니스 API 전체 용량의 1%를 지속적으로 사용할 수 있다. VVV를 소비할 필요는 없으며, 스테이킹 상태에서 필요 시 추론을 활용하면 된다.

베니스 API의 용량은 “Diem”이라는 단위로 측정된다.

Diem이란 무엇인가?

다양한 모델 및 작업 유형(텍스트, 이미지, 70B 모델, 405B 모델 등)에 걸쳐 추론 자원을 일관되게 측정하기 위해, 베니스는 “Diem”이라는 통합 지표를 도입했다. 이는 하루 기준 API에서 제공 가능한 총 추론 용량을 의미한다.

용량이 증가할수록 각 VVV 토큰의 가치도 증가한다. 각 토큰이 전체 용량에 대한 지분을 나타내기 때문이다.

예를 들어 Venice API 용량이 10,000 Diem이고, 특정 에이전트가 전체 스테이킹 VVV의 1%를 보유한 경우, 해당 에이전트는 하루 100 Diem을 사용할 수 있다. 또한 스테이킹을 통해 추가적인 VVV 보상도 획득한다.

Diem은 다양한 모델 사용 비용을 통합한 기준 단위로, 현재 약 1달러 상당의 크레딧을 의미한다.

따라서 위 예시에서 1%의 VVV 스테이킹은 하루 약 100달러 상당의 추론 크레딧에 해당한다.

베니스 인프라 용량이 확장될수록 Diem 총량도 증가한다. 예를 들어 총 용량이 20,000 Diem으로 증가할 경우, 동일한 VVV 스테이킹으로 하루 200달러 상당의 추론 크레딧을 사용할 수 있다.

즉, VVV를 스테이킹하면:

- 추론은 추가 비용 없이 제공된다
- 배출(emission) 기반의 수익이 발생한다

결과적으로 에이전트 및 API 사용자의 실질 비용은 0 이하로 낮아진다. 이는 특정 프로모션이나 구독 조건에 국한되지 않으며, VVV 스테이킹 규모에 비례해 적용된다. 또한 API 사용에 별도의 제한 조건이 없기 때문에, 에이전트는 확보한 API 용량을 제3자에게 재판매할 수 있다.

토크노믹스(Tokenomics)

VVV의 토크노믹스는 비교적 단순하다. 총 공급량이 제한되어 있으며, 스테이킹을 통해 지속적인 추론 자원 접근 권한이 부여된다.

- 총 1억 개의 VVV가 최초 발행되었으며, 이 중 절반은 Base 블록체인 상에서 베니스 사용자 및 크립토 x AI 커뮤니티 프로젝트에 에어드롭되었다.
- 사전 판매(pre-sale)는 없었으며, Aerodrome의 공개 유동성 풀을 통해 거래가 시작되었다.
- 매년 1,400만 개의 VVV가 신규 발행되며, 초기 인플레이션율은 14%이고 시간이 지남에 따라 감소한다.
- 신규 발행된 VVV(배출분)는 API 수요를 나타내는 활용률(Utilization Rate)에 따라 스테이커와 베니스에 분배된다.
- VVV 수요는 추론 자원 접근 권한이라는 효용에 기반한다. 베니스 인프라가 확장될수록 각 VVV가 대표하는 자원의 양도 증가한다.
- 스테이커는 자신의 지분에 해당하는 추론 자원을 직접 사용하거나 다른 주체에 재판매할 수 있다.
- 스테이커는 활용률에 기반한 스테이킹 수익을 획득한다.
- 컴퓨팅 성능 향상 및 비용 감소에 따라 베니스의 추론 용량은 증가하며, 이는 VVV 가치에 긍정적으로 작용한다.
- Venice API 수요 증가 시 VVV 보유자 전체가 혜택을 받으며, 이는 생성형 AI 산업 성장과의 연동 구조를 형성한다. 베니스는 최대 보유자로서 이해관계가 정렬되어 있으며, 보유 물량은 공개되어 있다.

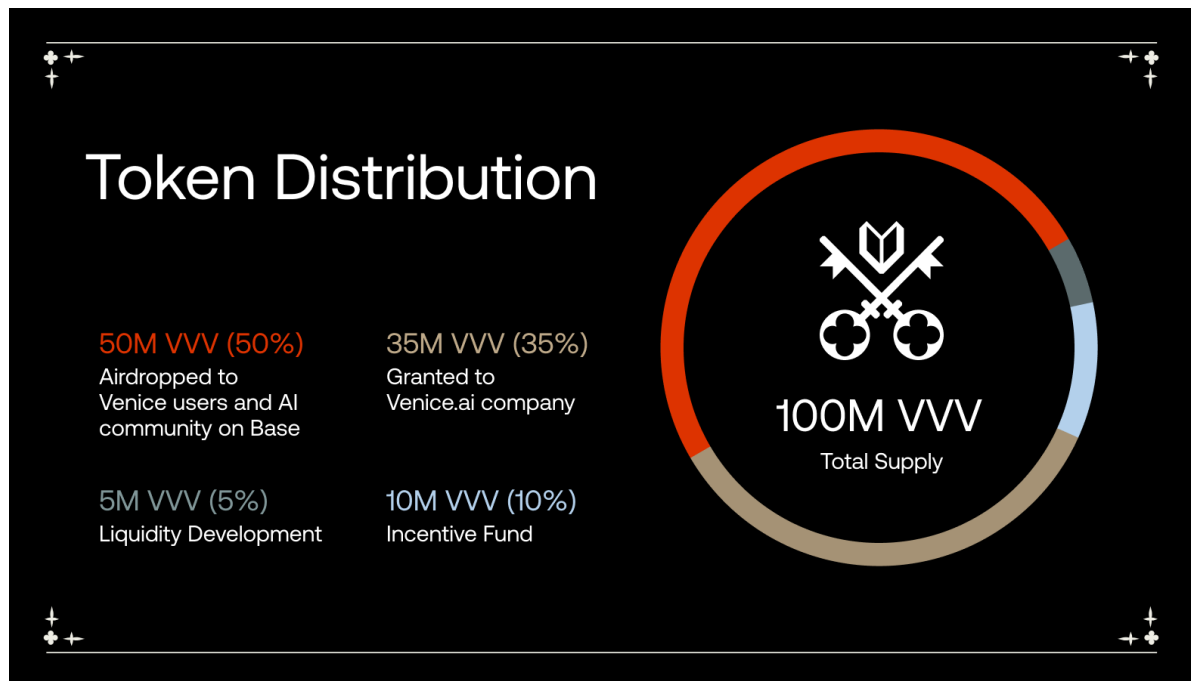
VVV 토큰 상세

- Base ERC20 컨트랙트
- 초기 공급량(TGE): 1억 개
- 연간 발행량: 1,400만 개
- Aerodrome 초기 유동성 풀
- 거버넌스 기능 없음
- 사전 판매 없음
- 감사(Audit) 완료

목적: Venice API에서 프라이버시 중심의 비검열 AI 추론(텍스트/이미지/코드)에 대한 접근 권한을 확보하고, 스테이킹 보상을 획득하기 위함.

VVV TGE 분배

- 5,000만 개(50%): 베니스 사용자 및 크립토 x AI 커뮤니티 에어드롭
- 3,500만 개(35%): Venice.ai 할당
 - 1,000만 개(10%): 팀 할당, 25% 즉시 언락, 나머지는 24개월에 걸쳐 분배
- 1,000만 개(10%): 베니스 인센티브 펀드
- 500만 개(5%): 유동성 공급



7. Diem의 이해(Understanding Diem)

출처: <https://venice.ai/blog/understanding-venice-compute-units-vcu>

베니스 토큰(VVV)은 스테이커에게 두 가지 핵심 혜택을 제공한다. 하나는 Diem을 통한 일일 AI 추론 할당이며, 다른 하나는 배출(emissions)에 기반한 지속적인 스테이킹 수익이다. 이러한 이중 인센티브 구조는 인프라를 적극적으로 사용하는 사용자와 네트워크 성장을 지원하는 장기 보유자 모두를 보상하도록 설계되어 있다.

베니스 토큰(VVV)과 스테이킹 혜택

VVV는 베니스의 AI 인프라를 구동하는 토큰으로, 스테이킹을 통해 AI 추론 자원에 대한 접근 권한을 제공하는 명확한 목적을 가진다.

총 공급량은 1억 개이며, 스테이킹을 장려하고 네트워크 확장을 지원하기 위해 매년 1,400만 개의 신규 토큰이 배출된다.

Diem의 이해

Diem은 다양한 모델과 작업 유형에 걸친 AI 추론 용량을 표준화하여 측정하는 단위이다. Diem은 Venice API에서 추론 비용을 지불하는 기준 단위로, 가격은 100만 개 입력/출력 토큰 단위로 설정된다.

VVV를 스테이킹하면 베니스 전체 연산 자원 중 일정 지분에 대한 권리가 부여되며, Venice API 사용 시 일일 Diem 할당량에서 해당 사용량이 차감된다. 각 에폭(epoch, 일 단위)이 시작될 때마다 Diem 할당량은 전량 초기화된다.

사용자의 Diem 할당량은 스테이킹한 VVV를 “활성 스테이커 총합(total active stakers)”으로 나누어 계산된다. 여기서 활성 스테이커란 최근 7일 내 API 호출 이력이 있는 스테이커를 의미한다. 이 비율에 현재 네트워크 총 용량을 곱해 최종 Diem 할당량이 산정된다. Diem 할당량은 매일 UTC 기준 자정에 갱신된다.

Diem은 API 접근의 기본 수단이지만, 베니스는 Pro 계정을 통해 USD로 추론을 직접 구매하는 방식도 지원한다.

각 Diem은 지원되는 모든 모델에서 사용할 수 있는 연산 크레딧의 일정 지분을 나타낸다. 일일 할당량은 전체 스테이킹된 VVV 대비 자신의 지분에 따라 결정되며, 예를 들어 전체 VVV의 1%를 스테이킹하면 베니스의 일일 추론 용량 중 1%에 접근할 수 있다.

모델별로 요구되는 Diem 양은 복잡도에 따라 달라진다.

채팅 모델은 입력 토큰과 출력 토큰을 구분하여 100만 토큰 단위로 과금되며, 실제 사용한 토큰 수에 대해서만 비용이 차감된다.

이미지 모델은 이미지 단위로 과금된다. 현재는 모델별 차등 가격이 적용되지 않지만, 향후 가격 정책 업데이트를 통해 변경될 예정이다. 이러한 비용 구조는 연산 요구량을 기준으로 설정된다.

Diem: 프라이버시 중심 비검열 추론 접근 구조

베니스 인프라가 확장될수록 전체 Diem 용량은 증가한다. 이는 동일한 VVV 스테이킹으로 더 많은 연산 자원에 접근할 수 있음을 의미하며, 다음과 같은 선순환 구조를 형성한다:

1. 인프라 확장 → 총 Diem 용량 증가
2. 동일 스테이킹 대비 Diem 증가 → VVV 효용 증가
3. VVV 효용 증가 → 스테이킹 수요 증가

VVV 스테이킹과 Diem 할당 구조의 결합은 사용자와 에이전트의 AI 인프라 접근 방식을 근본적으로 변화시킨다. 기존의 변동적인 토큰 기반 과금 방식과 달리, 사용자는 VVV를 스테이킹함으로써 지속적인 추론 권한을 확보하는 동시에 수익을 얻을 수 있다. 이는 대량 사용자의 비용 예측 가능성을 높이는 동시에 베니스와 사용자 간의 이해관계를 정렬한다.

연간 1,400만 개의 배출 물량은 API 사용률(Utilization Rate)에 따라 동적으로 분배된다. 사용률이 낮을 경우 약 80%가 스테이커에게 배분되며, 약 50% 수준에서는 약 80%가 인프라 확장에 활용된다. 사용률이 높아지면 다시 스테이커 중심으로 배분 비중이 조정되며, 네트워크 성장을 지속하면서도 전 구간에서 인센티브가 유지되도록 설계되어 있다.

8. 프라이버시(Privacy)

출처: <https://docs.venice.ai/overview/privacy>

베니스 토큰(VVV)은 스테이커에게 두 가지 핵심 혜택을 제공한다. 하나는 Diem을 통한 일일 AI 추론 할당이며, 다른 하나는 배출(emissions)에 기반한 지속적인 스테이킹 수익이다. 이러한 이중 인센티브 구조는 인프라를 적극적으로 사용하는 사용자와 네트워크 성장을 지원하는 장기 보유자 모두를 보상하도록 설계되어 있다.

베니스 토큰(VVV)과 스테이킹 혜택

VVV는 베니스의 AI 인프라를 구동하는 토큰으로, 스테이킹을 통해 AI 추론 자원에 대한 접근 권한을 제공하는 명확한 목적을 가진다.

총 공급량은 1억 개이며, 스테이킹을 장려하고 네트워크 확장을 지원하기 위해 매년 1,400만 개의 신규 토큰이 배출된다.

Diem의 이해

Diem은 다양한 모델과 작업 유형에 걸친 AI 추론 용량을 표준화하여 측정하는 단위이다. Diem은 Venice API에서 추론 비용을 지불하는 기준 단위로, 가격은 100만 개 입력/출력 토큰 단위로 설정된다.

VVV를 스테이킹하면 베니스 전체 연산 자원 중 일정 지분에 대한 권리가 부여되며, Venice API 사용 시 일일 Diem 할당량에서 해당 사용량이 차감된다. 각 에폭(epoch, 일 단위)이 시작될 때마다 Diem 할당량은 전량 초기화된다.

사용자의 Diem 할당량은 스테이킹한 VVV를 “활성 스테이커 총합(total active stakers)”으로 나누어 계산된다. 여기서 활성 스테이커란 최근 7일 내 API 호출 이력이 있는 스테이커를 의미한다. 이 비율에 현재 네트워크 총 용량을 곱해 최종 Diem 할당량이 산정된다. Diem 할당량은 매일 UTC 기준 자정에 갱신된다.

Diem은 API 접근의 기본 수단이지만, 베니스는 Pro 계정을 통해 USD로 추론을 직접 구매하는 방식도 지원한다.

각 Diem은 지원되는 모든 모델에서 사용할 수 있는 연산 크레딧의 일정 지분을 나타낸다. 일일 할당량은 전체 스테이킹된 VVV 대비 자신의 지분에 따라 결정되며, 예를 들어 전체 VVV의 1%를 스테이킹하면 베니스의 일일 추론 용량 중 1%에 접근할 수 있다.

모델별로 요구되는 Diem 양은 복잡도에 따라 달라진다.

채팅 모델은 입력 토큰과 출력 토큰을 구분하여 100만 토큰 단위로 과금되며, 실제 사용한 토큰 수에 대해서만 비용이 차감된다.

이미지 모델은 이미지 단위로 과금된다. 현재는 모델별 차등 가격이 적용되지 않지만, 향후 가격 정책 업데이트를 통해 변경될 예정이다. 이러한 비용 구조는 연산 요구량을 기준으로 설정된다.

Diem: 프라이버시 중심 비검열 추론 접근 구조

베니스 인프라가 확장될수록 전체 Diem 용량은 증가한다. 이는 동일한 VVV 스테이킹으로 더 많은 연산 자원에 접근할 수 있음을 의미하며, 다음과 같은 선순환 구조를 형성한다:

1. 인프라 확장 → 총 Diem 용량 증가
2. 동일 스테이킹 대비 Diem 증가 → VVV 효용 증가
3. VVV 효용 증가 → 스테이킹 수요 증가

VW 스테이킹과 Diem 할당 구조의 결합은 사용자와 에이전트의 AI 인프라 접근 방식을 근본적으로 변화시킨다. 기존의 변동적인 토큰 기반 과금 방식과 달리, 사용자는 VW를 스테이킹함으로써 지속적인 추론 권한을 확보하는 동시에 수익을 얻을 수 있다. 이는 대량 사용자의 비용 예측 가능성을 높이는 동시에 베니스와 사용자 간의 이해관계를 정렬한다.

연간 1,400만 개의 배출 물량은 API 사용률(Utilization Rate)에 따라 동적으로 분배된다. 사용률이 낮을 경우 약 80%가 스테이커에게 배분되며, 약 50% 수준에서는 약 80%가 인프라 확장에 활용된다. 사용률이 높아지면 다시 스테이커 중심으로 배분 비중이 조정되며, 네트워크 성장을 지속하면서도 전 구간에서 인센티브가 유지되도록 설계되어 있다.

