Title: Contextual AI

URL: https://en.wikipedia.org/wiki/Contextual_AI

PageID: 80182758

Categories: Category:2023 establishments in California, Category:American companies established in 2023, Category:Artificial intelligence companies, Category:Enterprise software, Category:Information retrieval systems, Category:Large language models, Category:Natural

language processing

Source: Wikipedia (CC BY-SA 4.0).

Contextual AI is an enterprise software company [1] based in Mountain View, California. It develops a platform for building [2] specialized Retrieval-Augmented Generation (RAG) agents for enterprise use. [3] The company was founded in 2023 by Douwe Kiela and Amanpreet Singh, both former AI researchers at Facebook AI Research (FAIR) [4] and Hugging Face . [5] Douwe Kiela previously led the Meta research team that introduced the Retrieval-Augmented Generation (RAG) approach in 2020. [6] [7] [8]

Contextual AI focuses on enterprise generative AI applications using RAG 2.0 technology, [9] with deployments primarily in the technology, banking, finance and media sectors. [10]

History

In June 2023, Contextual AI announced [4] it had raised \$20 million in a seed funding round led by Bain Capital Ventures (BCV), with participation from Lightspeed Venture Partners, Greycroft, SV Angel, and several angel investors. [2]

In August 2024, the company raised \$80 million in a Series A funding round [11] led by Greycroft, [12] with participation from previous investors [13] including Bain Capital Ventures, Lightspeed, and Conviction Partners. [14] The round also included new backers such as Bezos Expeditions, NVentures (Nvidia), HSBC Ventures, and Snowflake Ventures . [15]

Features

Retrieval-Augmented Generation (RAG) is an artificial intelligence framework [1] that integrates information retrieval with text generation to improve the performance of large language models (LLMs) [16] on complex, knowledge-intensive tasks. It was introduced in 2020 by researchers at Meta AI, including Douwe Kiela, Patrick Lewis and others, in their paper Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. [6] RAG enables language models to access [17] and incorporate external information, such as proprietary databases or real-time web content, at query time, instead of relying solely on pre-trained, [18] internal, static knowledge. This architecture addresses common limitations of standard LLMs, including hallucination, [19] outdated information, and lack of attribution to source materials. [20] RAG systems retrieve [6] relevant context through a variety of techniques - including vector search, keyword search, text-to-SQL - and feeds this context into the language model to generate responses. The approach improves factual accuracy, [21] supports domain-specific customization, enables citation of sources, and allows for more updated information without retraining the model itself.

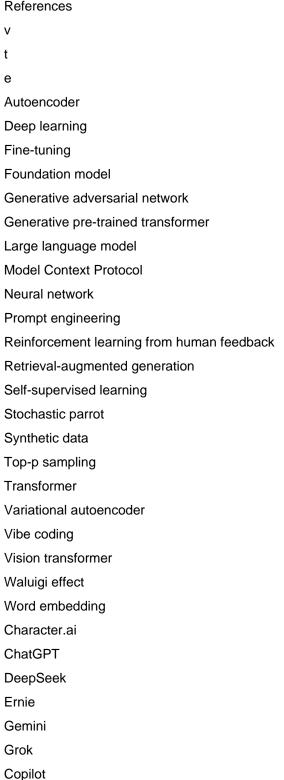
General Availability. In January 2025, Contextual AI announced the general availability of its enterprise platform for building specialized RAG agents. [22] Early adopters included Qualcomm, which used the platform for their Customer Engineering team needs.

Grounded Language Model. In March 2025, the company introduced a Grounded Language Model (GLM) [23] for factual accuracy in enterprise Al applications.

Reranker . In March 2025, Contextual AI released an instruction-following reranker [24] that allows users to influence the ranking of retrieved documents through natural language instructions, such as prioritizing recent files, specific formats, or content from designated sources.

Applications

Contextual Al's platform has been adopted across a range of industries, including finance, technology, media and professional services. Clients include Fortune 500 companies such as Qualcomm [25] and HSBC . [26] References



Claude Gemini

Gemma GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5 1 2 3 J 4 40 4.5 4.1 OSS 5 Llama о1 о3 o4-mini Qwen Base44 Claude Code Cursor Devstral GitHub Copilot Kimi-Dev Qwen3-Coder Replit Xcode Aurora Firefly Flux GPT Image 1 Ideogram Imagen Midjourney

Recraft
Seedream
Stable Diffusion
Dream Machine

Qwen-Image

Hailuo Al	
Kling	
Midjourney Video	
Runway Gen	
Seedance	
Sora	
Veo	
Wan	
15.ai	
Eleven	
MiniMax Speech 2.5	
WaveNet	
Eleven Music	
Endel	
Lyria	
Riffusion	
Suno Al	
Udio	
Agentforce	
AutoGLM	
AutoGPT	
ChatGPT Agent	
Devin Al	
Manus	
OpenAl Codex	
Operator	
Replit Agent	
01.AI	
Aleph Alpha	
Anthropic	
Baichuan	
Canva	
Cognition AI	
Cohere	
Contextual Al	
DeepSeek	
ElevenLabs	
Google DeepMind	

HeyGen **Hugging Face** Inflection AI Krikey Al Kuaishou Luma Labs Meta Al MiniMax Mistral Al Moonshot Al OpenAl Perplexity AI Runway Safe Superintelligence Salesforce Scale Al SoundHound Stability Al Synthesia Thinking Machines Lab Upstage xΑI

Z.ai

Category