EXPLORING GENERATIVE AI

Prof. Dr. Jan Kirenz

01 THE IMPACT OF GENERATIVE AI

FORRESTER

GENERATIVE AI IS THE MOST IMPACTFUL TECHNOLOGICAL ADVANCEMENT SINCE THE INTERNET

Source: Forrester (2023). Forrester's Predictions 2024.

HOCHSCHULE DER MEDIEN

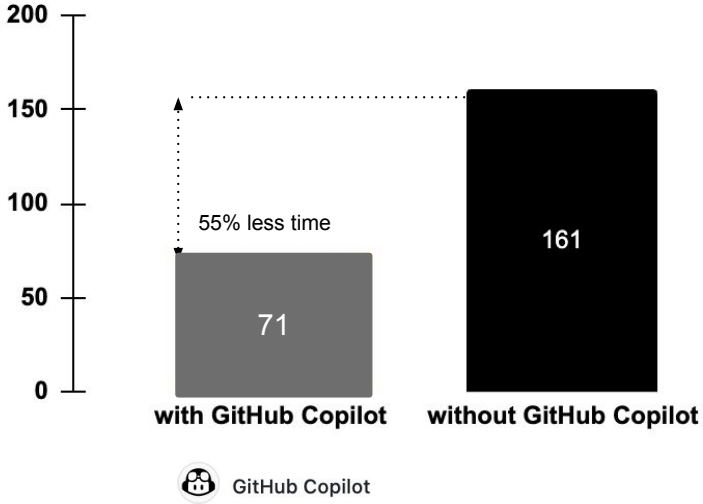
FORRESTER

GENERATIVE AI INITIATIVES WILL BOOST PRODUCTIVITY BY 50%

Source: Forrester (2023). Forrester's Predictions 2024.

HOCHSCHULE DER MEDIEN

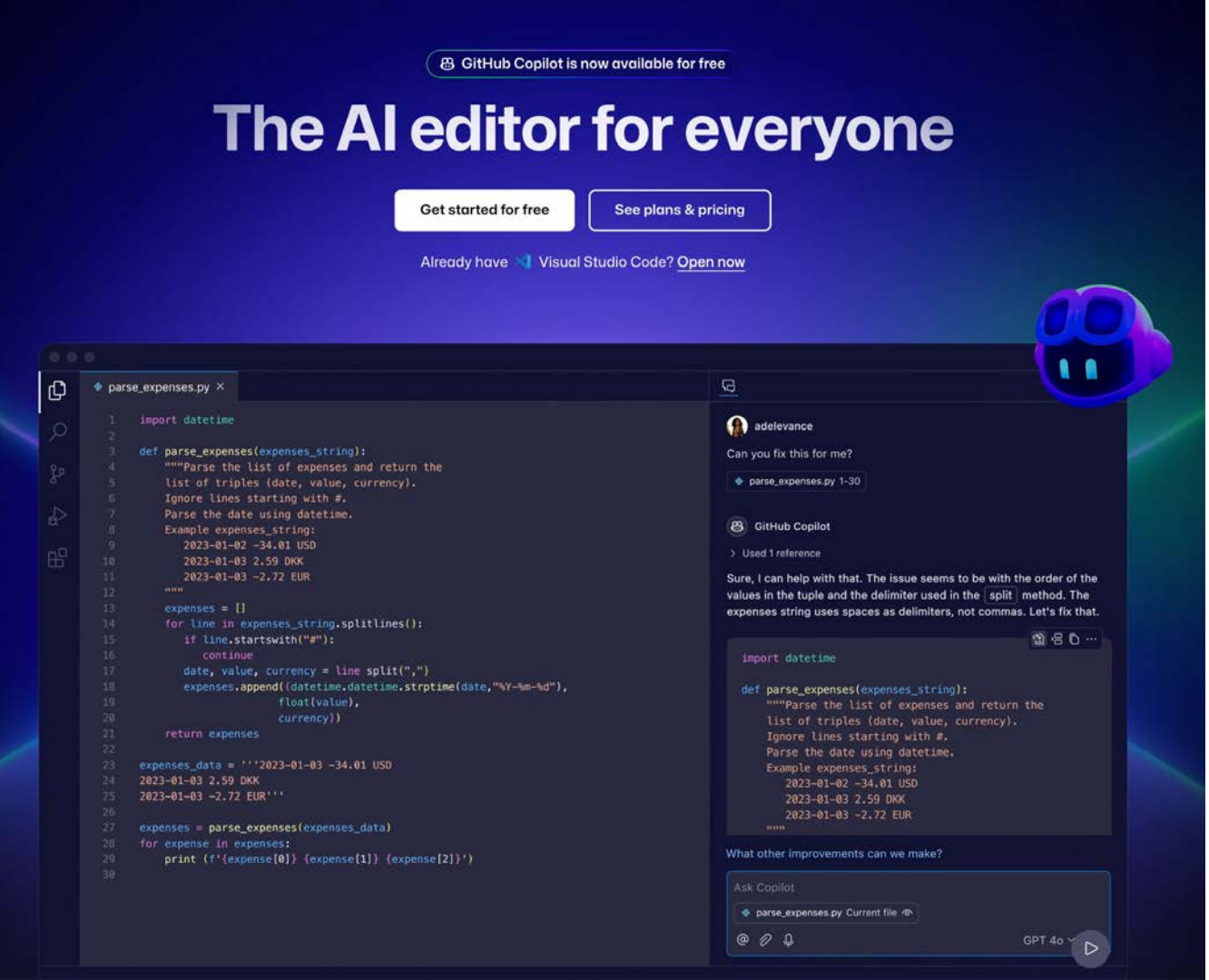
IMPACT OF GENERATIVE AI ON TASK PRODUCTIVITY



Task: Implement an HTTP server in JavaScript

Skills: Software development

Source: Peng et al. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot



GitHub Copilot

Visual Studio Code

python™

Education

ners

GitHub Copilot

Copilot for free!

GitHub Student Developer Pack

Learn to ship software like a pro. There's no substitute for hands-on experience. But for most students, real world tools can be cost-prohibitive. That's why we created the GitHub Student Developer Pack with some of our partners and friends.

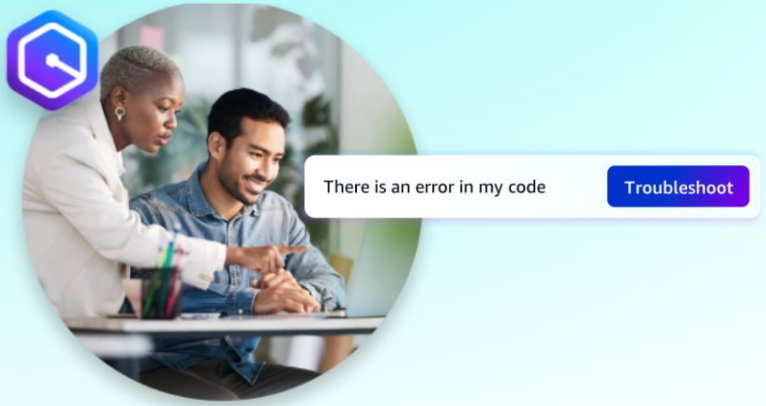
Sign up for Student Developer Pack

Love the pack? Spread the word

Post Like 78K

Amazon Q – Generative AI Assistant

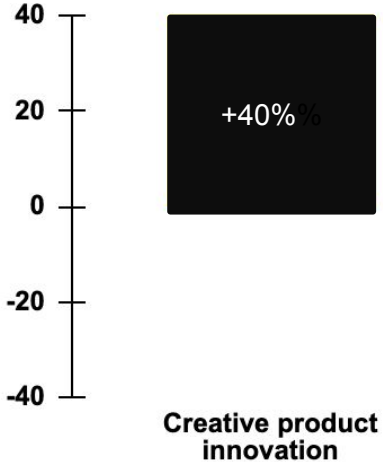
The most capable generative AI–powered assistant for accelerating software development and leveraging companies' internal data



$260 MIO IN ANNUAL COST SAVINGS

boost developer efficiency and productivity as much as 40%

IMPACT OF GEN AI ON CREATIVE PRODUCT INNOVATION



Task: Brainstorm ideas for new products

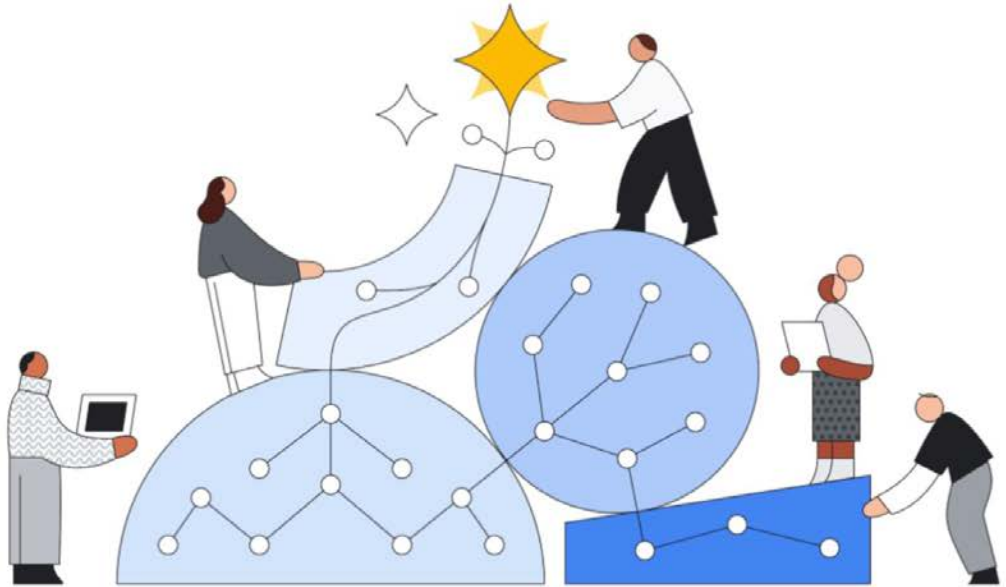
Skill: Creativity

Source: Dell'Acqua et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.

AI & Machine Learning

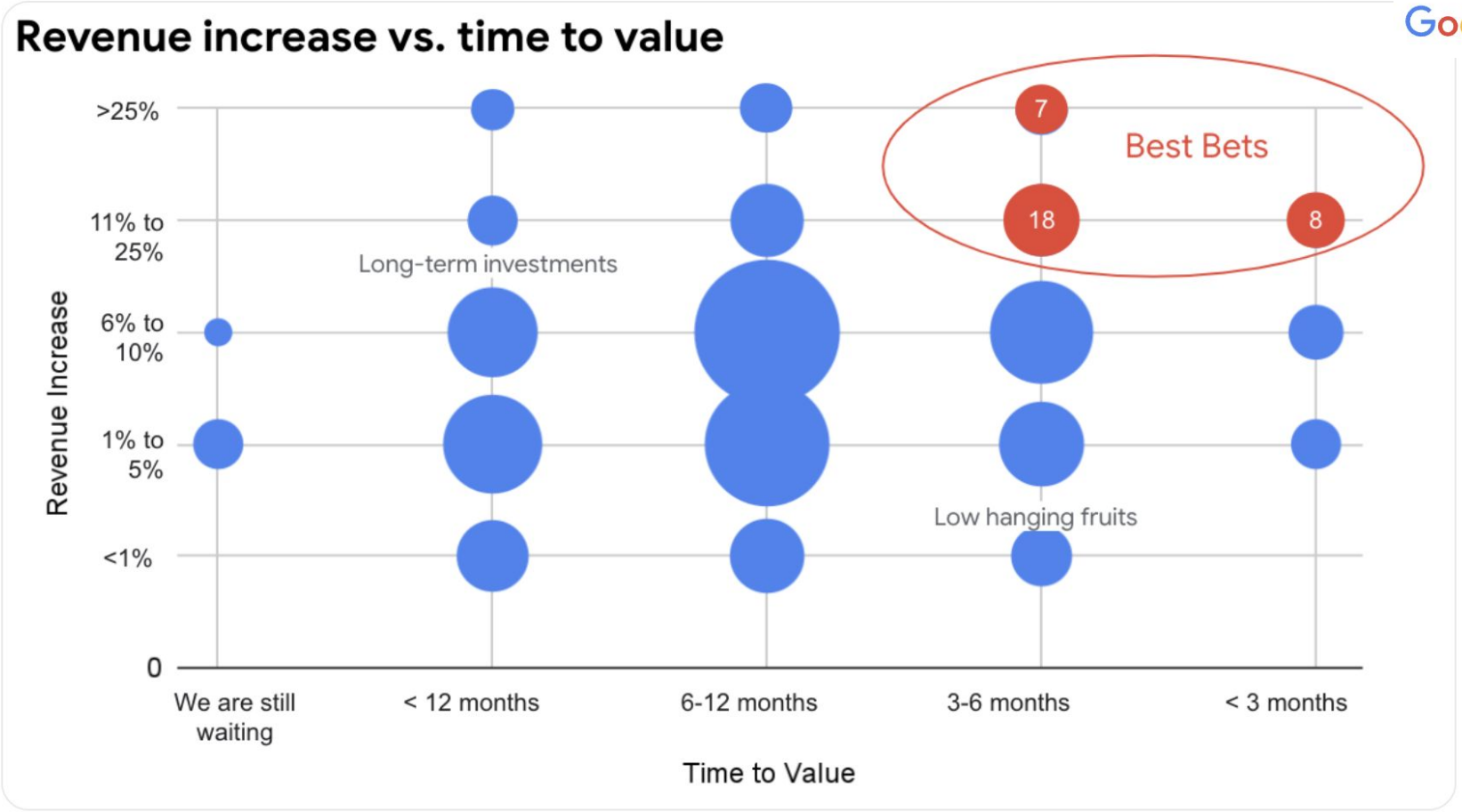
AI's Business Value: Lessons from Enterprise Success

January 13, 2025



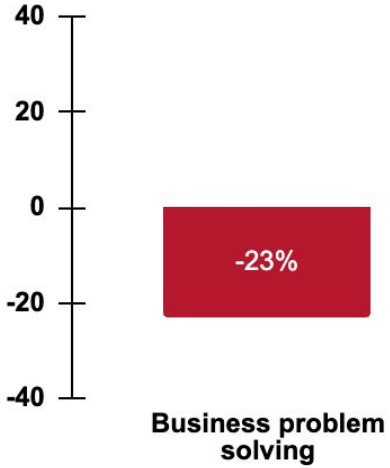
Google Cloud

HOCHSCHULE DER MEDIEN



Revenue increase vs. time to value: "Best Bet" use cases impact revenue significantly and quickly.

IMPACT OF GEN AI ON BUSINESS PROBLEM SOLVING



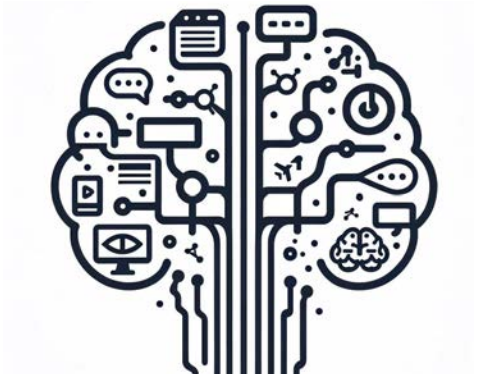
Task: Optimize revenue and profitability

Skill: Critical judgement

Source: Dell'Acqua et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.

THE LIMITS OF GENERATIVE AI

1+1=4



Factual errors: Creating information that is wrong.

Hallucinations: Generating information that is entirely made up.

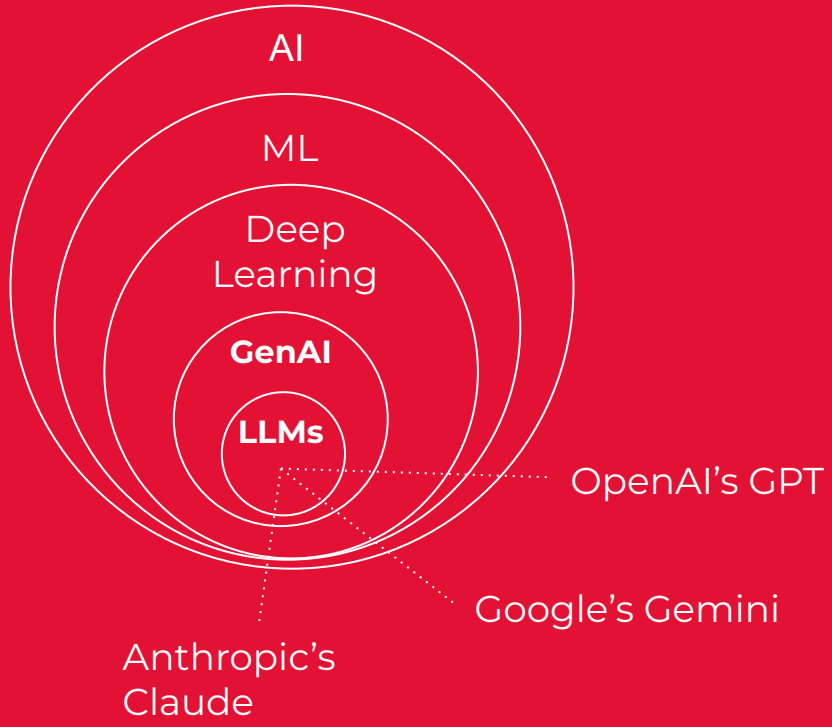
Bias: AI can inherit biases from training data.

FROM THIS POINT ON, ARTIFICIAL INTELLIGENCE WILL ONLY CONTINUE TO IMPROVE

Source: McAleese et al. (2024). LLM Critics Help Catch LLM Bugs. arXiv:2407.00215

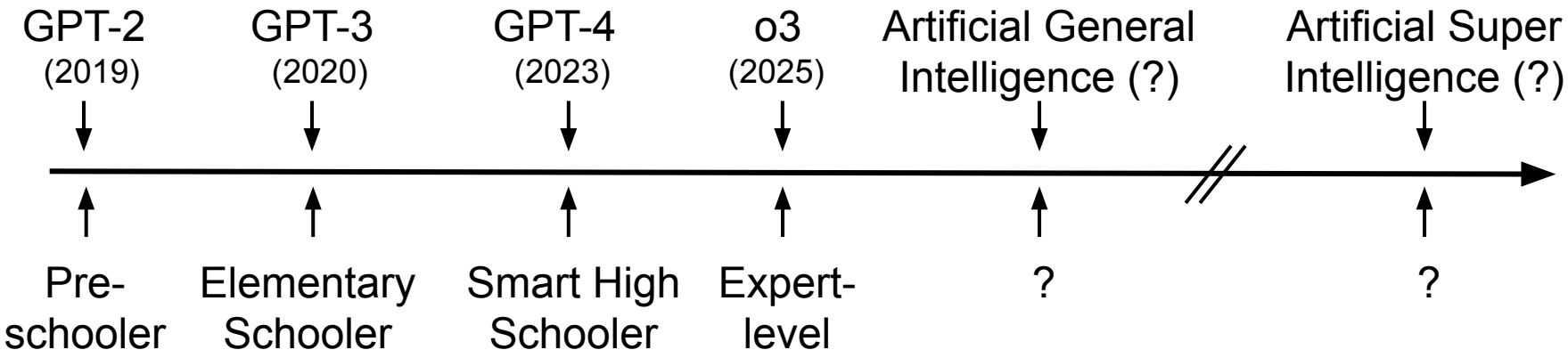
02 THE WORLD OF AI

UNDERSTANDING THE AI HIERARCHY



THE SCALE OF (ARTIFICIAL) INTELLIGENCE

Progress over just a few years



Source: based on Aschenbrenner (2024). Situational Awareness

THE TRANSFORMER ARCHITECTURE ENABLED LLMS

2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

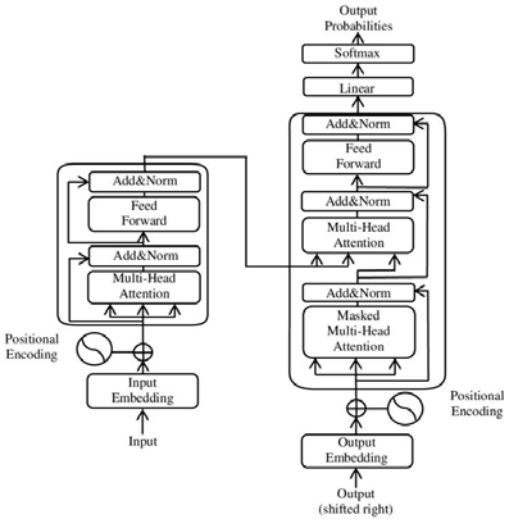
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

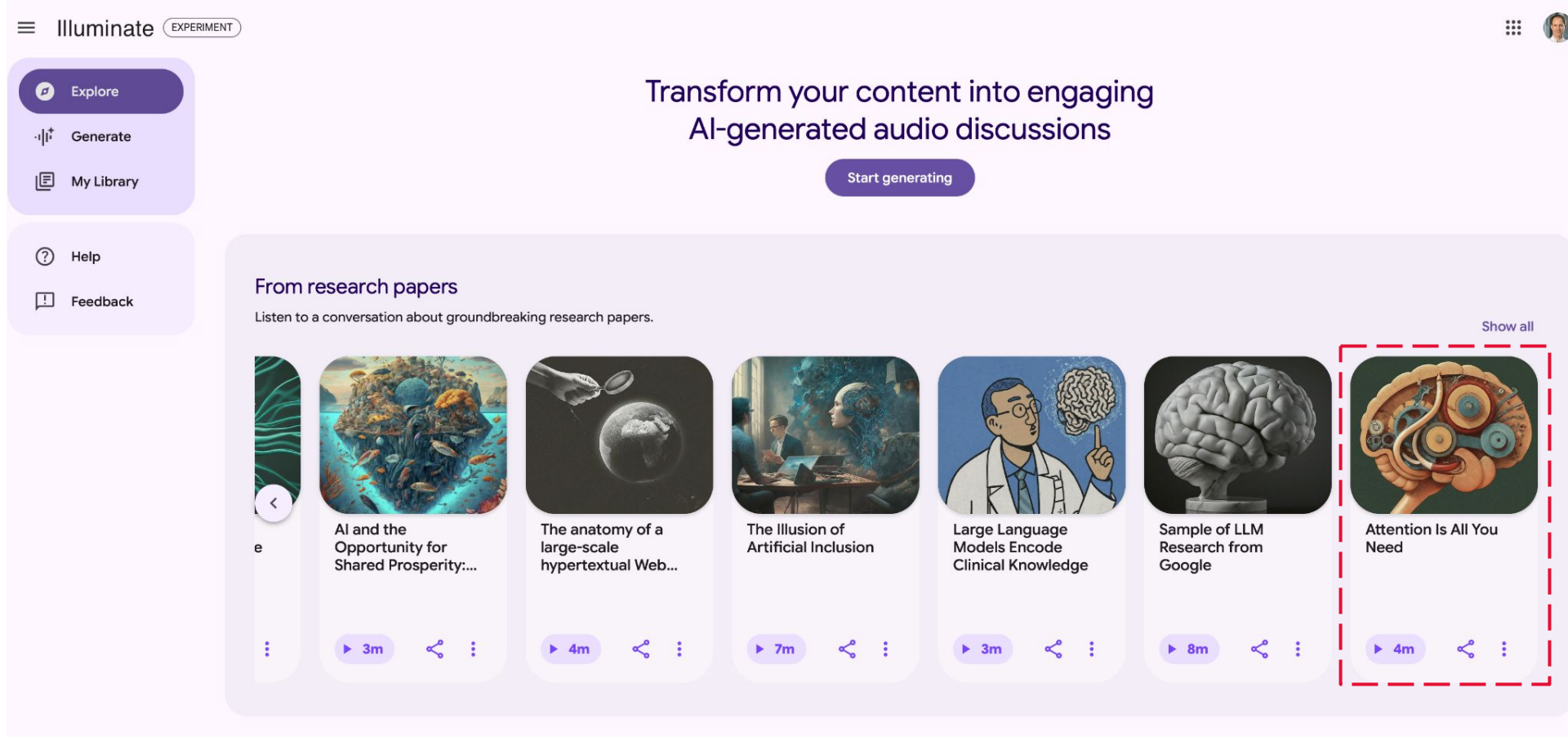
‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Source: Vaswani et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.



An illustration of main components of the transformer model



https://illuminate.google.com/explore

Illuminate EXPERIMENT

Home
Generate
My Library

Let's unpack a paper titled Attention Is All You Need. What's the core idea here?

Well, the big idea in this paper is that we can build a really effective sequence transduction model without using recurrence or convolutions - the usual suspects, right? - and instead just rely on attention mechanisms. The paper shows that in the context of machine translation, this new approach not only performs better than RNNs but also trains faster.

That's super interesting, especially considering the time this paper was published. It's from 2017, right? What was the state of sequence models back then?

Exactly, 2017! Back then, everyone was all about RNNs - recurrent neural networks - think LSTMs and GRUs. These models were the top dogs for tasks like language translation and text summarization, you know? But these RNNs had a bit of a bottleneck. They were tough to train on long sequences because they had to process everything in order, one step at a time and that was a big limitation.

How did the Attention Is All You Need paper address this sequential processing bottleneck of RNNs?

So, instead of going step-by-step like RNNs, they introduced a model called the Transformer, hence the title. The Transformer processes the entire sequence all at once by using something called self-attention. It's like giving the model the ability to look at all parts of the input simultaneously and figure out which parts are most relevant to each other. Kinda like having a bird's eye view of the entire context, which is awesome for capturing those long-range dependencies between words in a sentence.

This self-attention mechanism sounds powerful! How does self-attention work in the Transformer model?

It's a bit like how we focus on different parts of a picture to understand it. Each word in the sequence looks at other words and assigns them different weights, kind of like voting for the most important words in the context. The really cool part is that these weights are learned by the model during training!

So it's a bit like the model is learning to focus on what's important - like a spotlight for language. Can you tell us a bit more about how the model learns these weights?

Yeah, it's all about the query, key, and value concept! These are vector representations for each word in the input sequence. The attention weight for each word is calculated based on the compatibility between the query and key vectors, typically using dot-product. The value vectors, weighted by the attention scores, are then combined to produce the output.

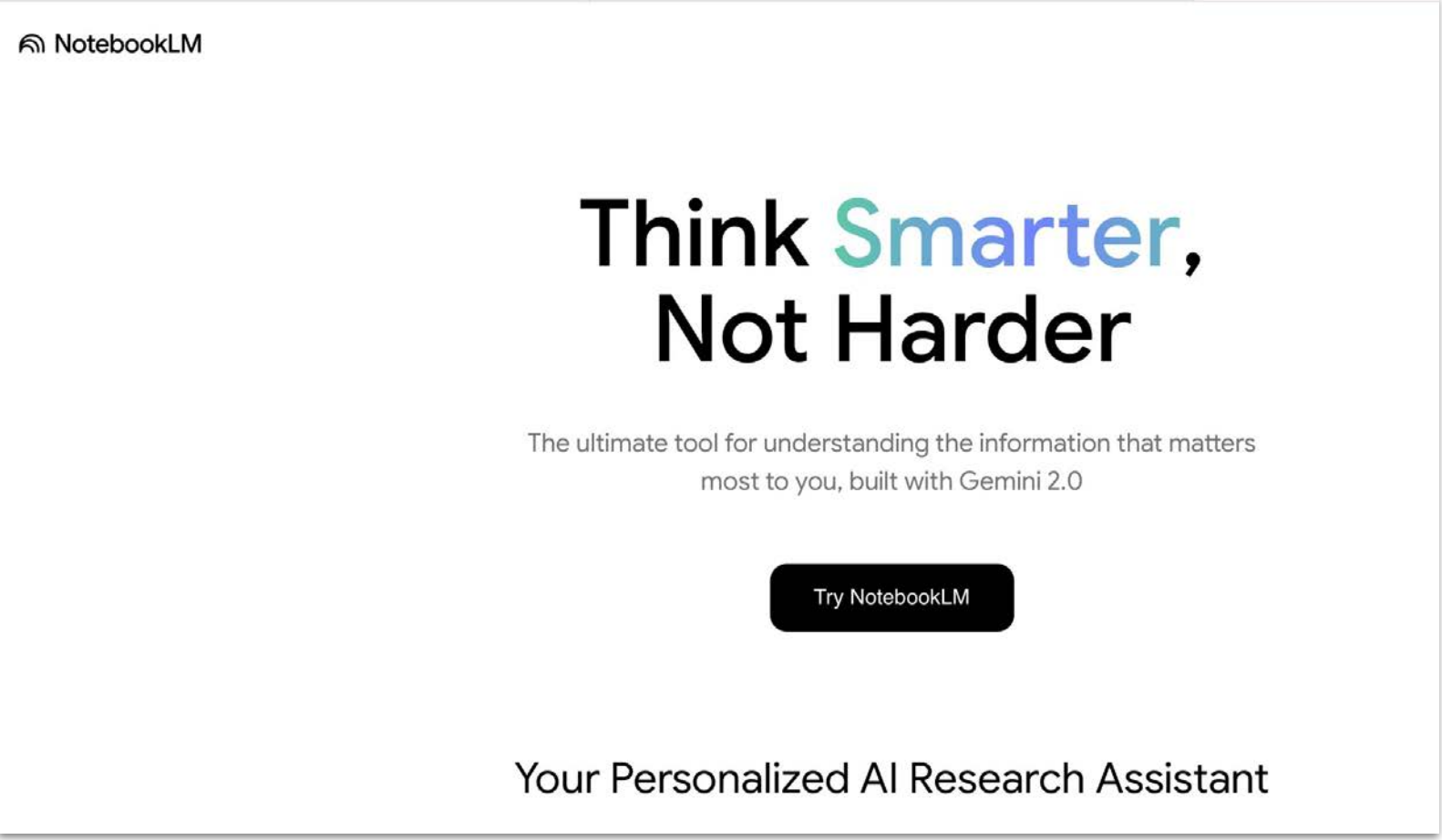
Are there any other advantages to using the Transformer model?

Attention Is All You Need

0:01 -4:24

Help
Feedback

HOCHSCHULE DER MEDIEN

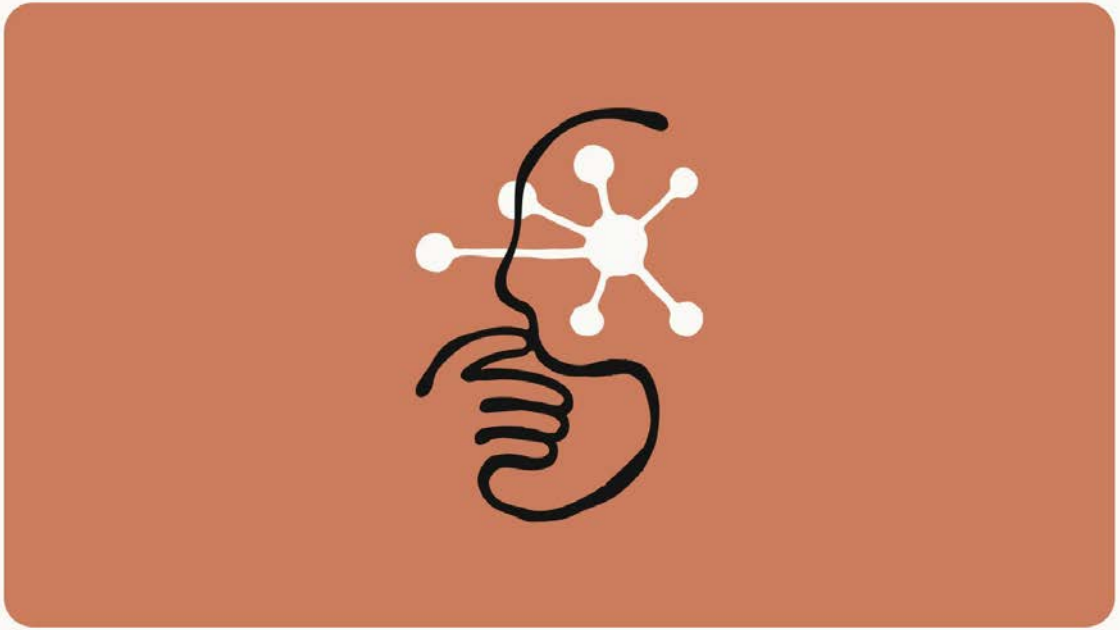


https://notebooklm.google.com/

Announcements

Claude 3.7 Sonnet and Claude Code

24. Feb. 2025 • 5 min read

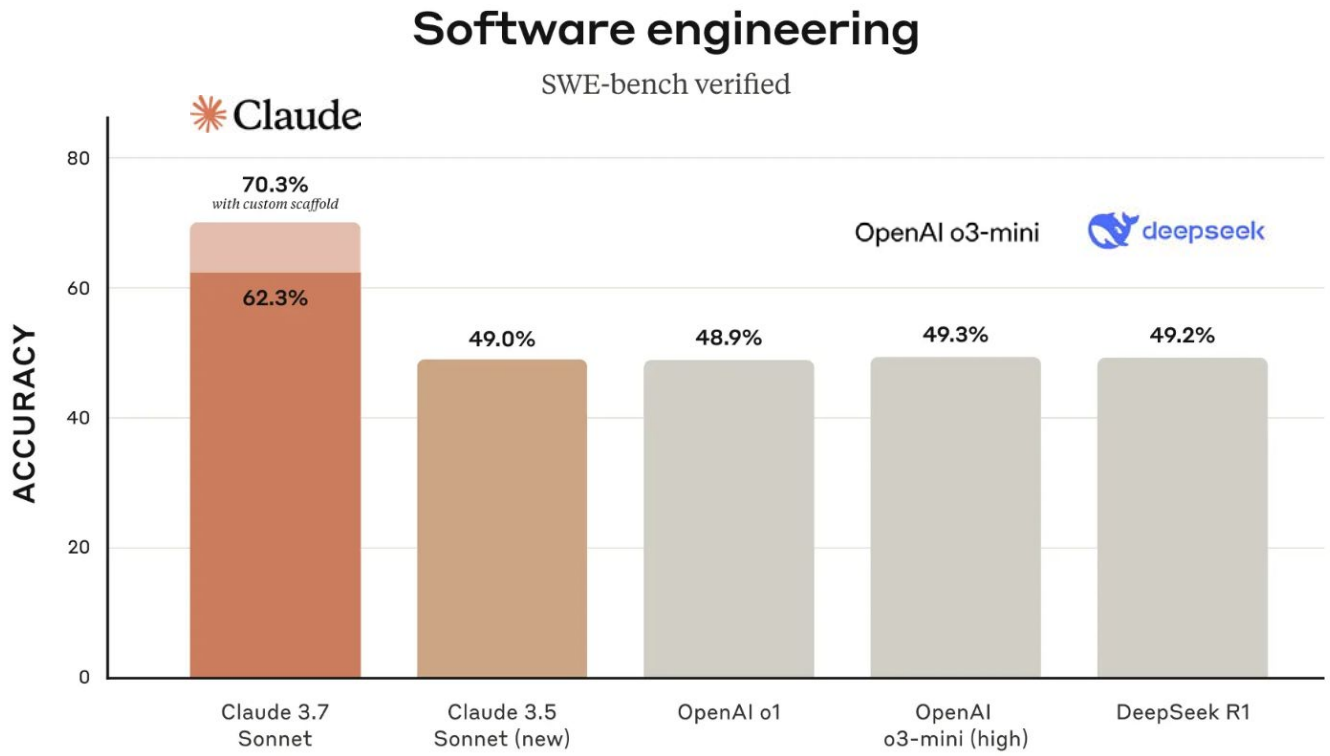


ANTHROP\C

Released Claude 3.7 Sonnet
24. Feb. 2025

SOFTWARE ENGINEERING

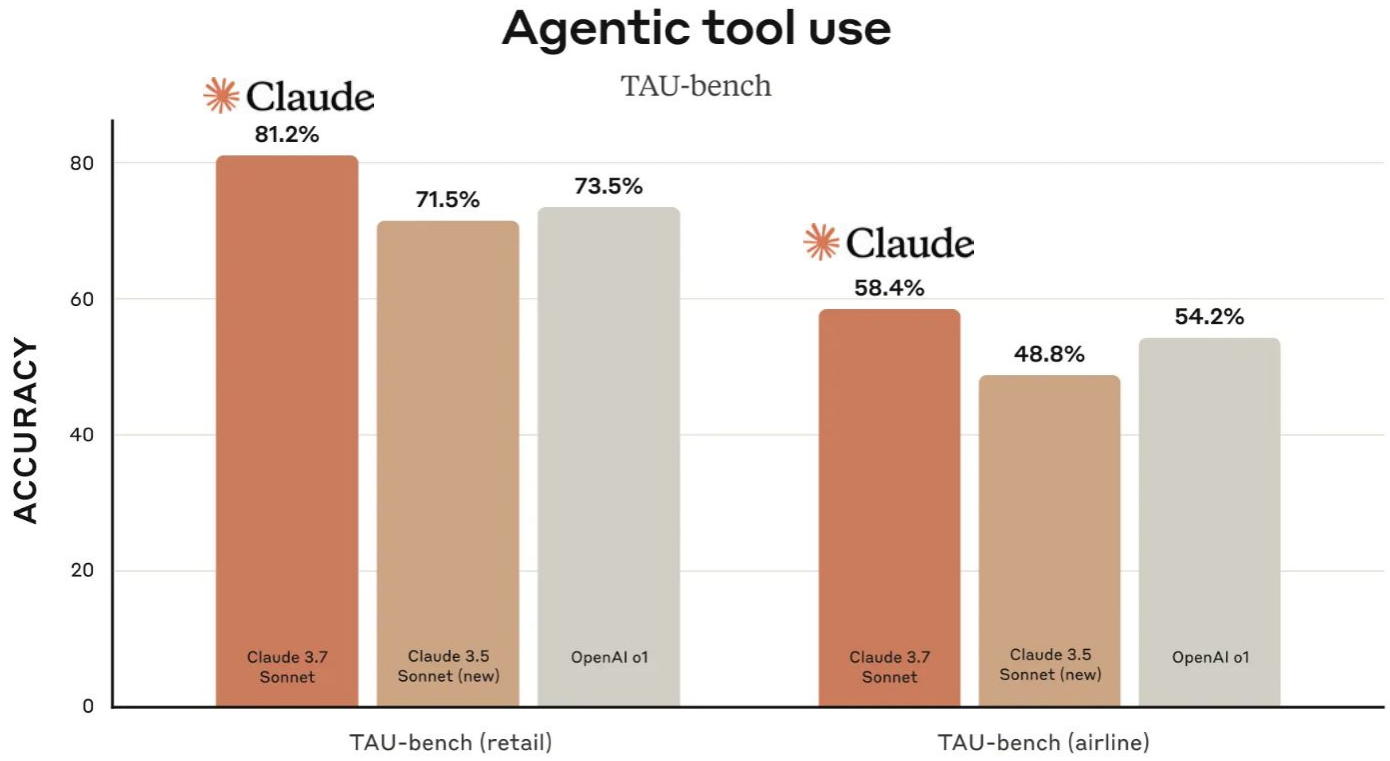
SWE-bench evaluates AI models' ability to solve real-world software issues.
SWE-bench



Source: Anthropic (2025)

AGENTIC TOOL USE

TAU-bench, a framework that tests AI agents on complex real-world tasks with user and tool interactions.
TAU-bench



Source: Anthropic (2025)