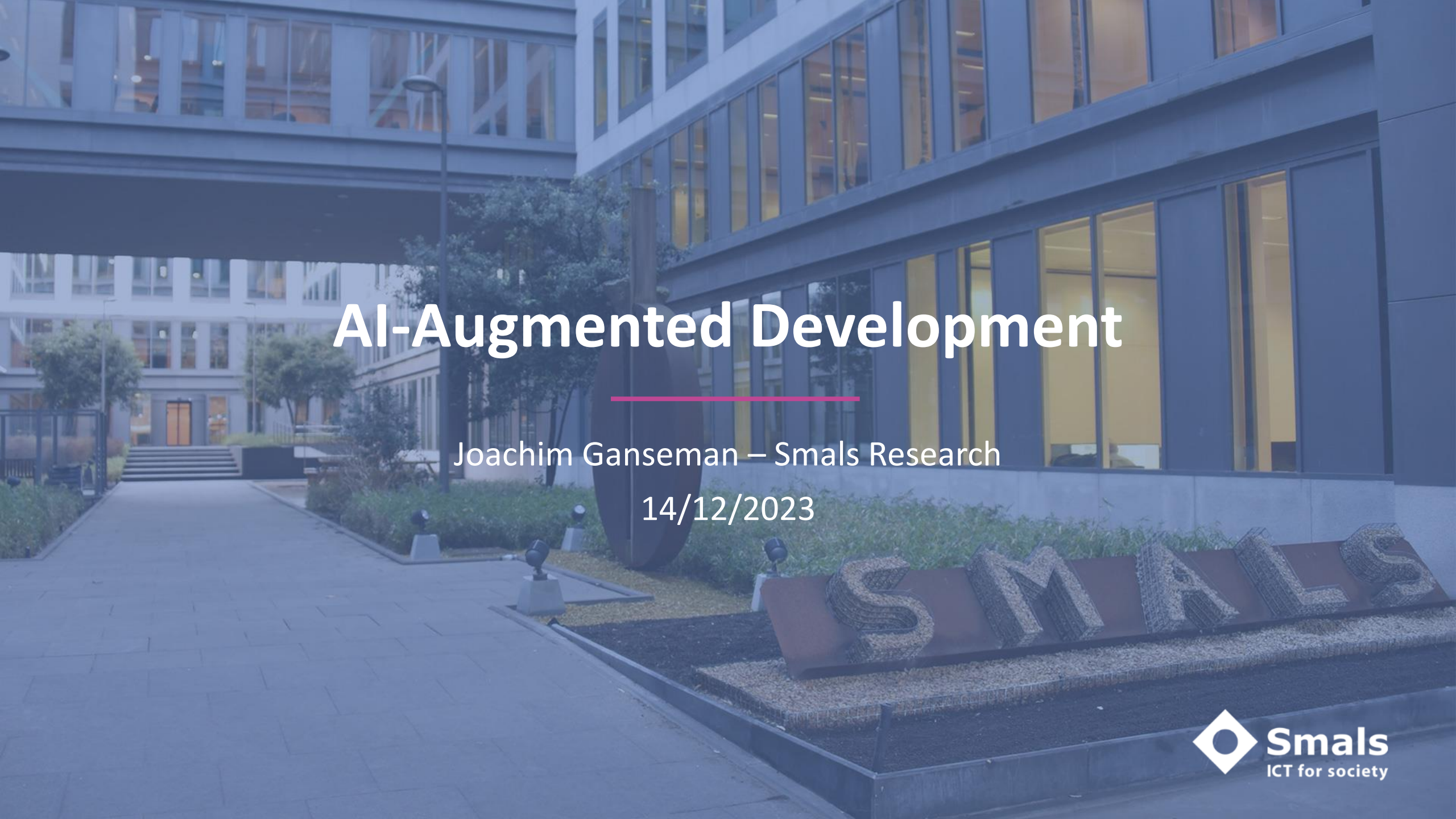


AI-Augmented Development

Joachim Ganseman – Smals Research

14/12/2023

Smals
ICT for society

Innovation with new technologies

Consultancy & expertise

Internal & external knowledge transfer

Support for going live

AI-Augmented Software Development

Workshops AI

Language Model for social security

Data protection in Legacy

Pseudonymization Service

Customer Service Automation

Recommender System Portal

Smals Research 2023

GIS Analytics

Confidential Computing

Custom Assistant

Causal AI

Event Driven Architecture

Graph Machine Learning

Synthetic Data in practice

On-prem Voice2voice Translation

Diffusion Models

Digital EU Wallet

Linkurious in practice

Best Practices in Cryptography

www.smalsresearch.be

Agenda

- Introduction – Live Demo
- Generative AI in Development Tools
- How does this work? Let’s find out!
- Caveats and considerations
- Conclusion

Smals ICT for society

About me

- Master (Licentiaat) Computer Science
- Career: mostly (academic) research
- Programming background: mostly C++ / Python
- Definitely not a super-experienced developer

  → AI to the rescue !

Smals ICT for society

Introduction

Smals
ICT for society

Live Demo

(Image by Bing Image Creator: *“At a webinar, the presenter demonstrates AI-powered software capable of generating code.”*)

Smals ICT for society

- Quickly make relevant or even novel suggestions
- Helps generate code, and tests, and docs
- Works OK for small tasks with enough context
- Works OK with commonly used libraries, file formats, programming languages, tools, …
- Good tool to “get started”, write boilerplate,
- Fairly easy to learn to use
- …

- Suggestions are “hallucinated”, not logically deduced
- No guarantees on correctess or fitness for purpose
- Same question can yield various responses
- May suggest deprecated functions, suboptimal code, bad practices or incomprehensible code
- Mimicks previously written code rather than suggesting the best or cleanest solution
- Strong oversight & testing/validation processes needed
- …

Smals ICT for society

Generative AI in Development Tooling

Smals
ICT for society

Generative models for coding assistance

- Code completion / infilling
  - Text continuation model
  - GPT-3 / BERT like
  - Trained on masked text (prose)

… for code:

  - Fill in a function
  - Continue this incomplete piece of code
  - Write a comment to follow this code
  - Write some code to follow this comment

- Dialog / Conversation
  - Responds to questions
  - ChatGPT-like
  - Trained on “Instructions” (= Question-Response) incorporating human feedback (“RLHF”)

… for code:

  - Reply to statements formatted as question
  - “How could this code be improved ?”
  - “Write code / config file that does … “

Smals ICT for society

Github CoPilot : the first mover advantage

- Initially based on OpenAI Codex, a modified production version of GPT-3 (exclusively licensed to Microsoft)
- March 2023: deprecation of Codex models, move towards “Chat Completion” model
- Functionality:
  - Turn comments into code
  - Complete your next line or function in context
  - Offer new knowledge, such as finding a useful library or API call for an application
  - Add comments
  - Rewrite code for efficiency
- Examples: see
  - https://platform.openai.com/docs/guides/code
  - https://cookbook.openai.com/

Smals
ICT for society

Reception and benefits

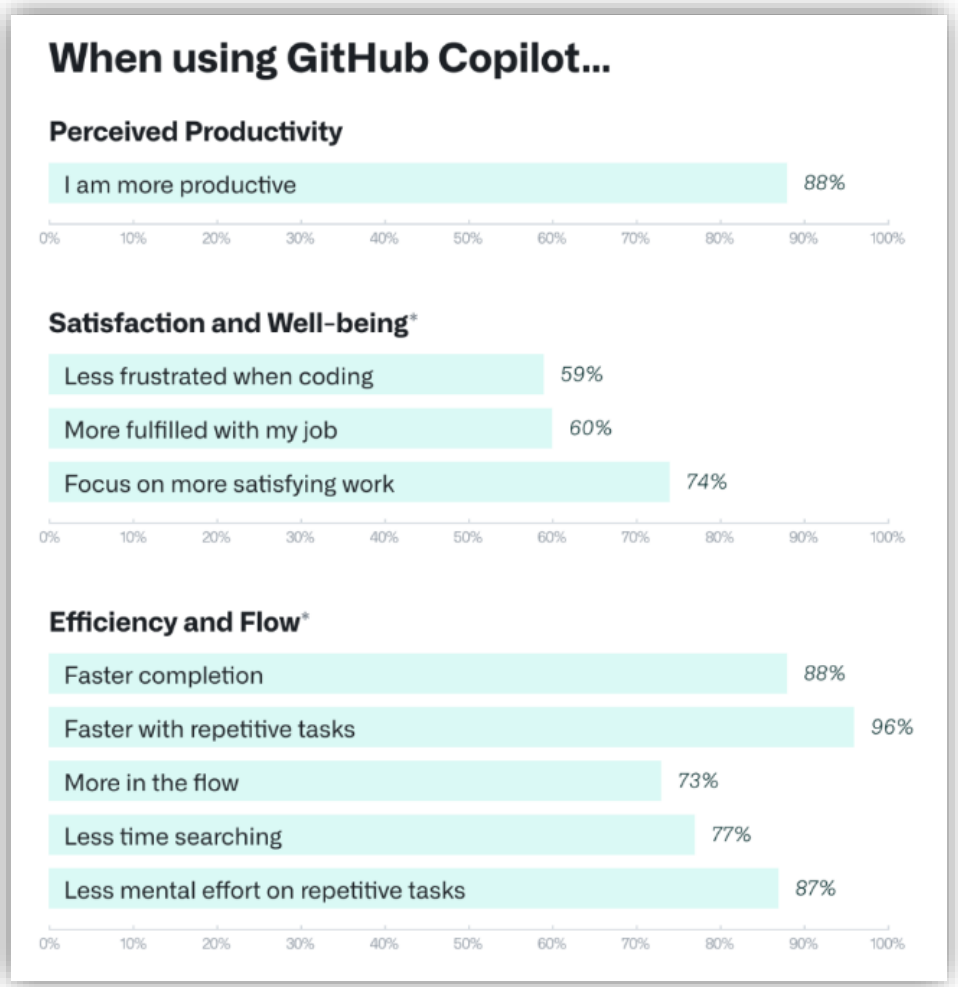
How favorable is your stance on using AI tools as part of your development workflow?

	Favorable	Indifferent	Unfavorable
Learning to Code	76.1%	20.7%	3.2%
Professional Developers	77.3%	19.7%	3.0%

For the AI tools you use as part of your development workflow, what are the MOST important benefits you are hoping to achieve?

	Greater efficiency	Improve accuracy in coding	Improve collaboration	Increase productivity	Speed up learning
Learning to Code	33.7%	23.8%	6.8%	41.5%	42.4%
Professional Developers	27.9%	14.1%	4.2%	37.4%	27.4%

(Source: StackOverflow developer survey, june 2023)



(Source: Github blog)

Smals ICT for society

Developer sentiment around AI

Which parts of your development workflow are you currently using AI tools for and which are you interested in using AI tools for over the next year?

Learning about a codebase

	Currently Using	Interested in Using	Not interested in Using
Learning to Code	50.0%	48.6%	15.2%
Professional Developers	32.8%	58.2%	15.2%

Project planning

	Currently Using	Interested in Using	Not interested in Using
Learning to Code	29.0%	52.5%	25.9%
Professional Developers	15.1%	48.9%	39.2%

Writing code

	Currently Using	Interested in Using	Not interested in Using
Learning to Code	82.0%	27.8%	6.2%
Professional Developers	86.2%	24.4%	4.4%

Documenting code

	Currently Using	Interested in Using	Not interested in Using
Learning to Code	38.2%	8.5%	11.9%
Professional Developers	40.0%	58.2%	58.6%

Debugging and getting help

	Currently Using	Interested in Using	Not interested in Using
Learning to Code	53.9%	37.9%	6.3%
Professional Developers	53.9%	48.6%	7.5%

Testing code

	Currently Using	Interested in Using	Not interested in Using
Learning to Code	27.1%	62.1%	18.0%
Professional Developers	29.1%	65.2%	11.6%

(Source: StackOverflow developer survey, june 2023)

Smals ICT for society

A few other coding assistant tools

Product	Trained on	Self hosting option	Fine tuned to company codebase	Status	Open source	Pricing
Tabnine	Open source code with permissive licenses.	Yes (enterprise plan)	Yes (enterprise plan)	Production	No	Free + Paid
GitHub Copilot	Publicly available code (regardless of license)	No	No	Production	No	Paid
Replit Ghostwriter	Publicly available code (regardless of license)	No	No	Production	No	Paid
Amazon CodeWhisperer	Amazon and open-source code: license not specified	No	No	Preview	No	Free
Codeium	Own model. Includes training on code with non-permissive licenses.	Yes (enterprise plan)	Yes	Production	Some plugins	Free + Paid
SourceGraph Cody	Users can select: OpenAI or Anthropic currently. Future plans for other models.	Yes	No	Experimental	Yes	N/A
CodeComplete	Open source code with permissive licenses	Yes	Yes	Private beta	No	N/A
FauxPilot	Salesforce CodeGen: code with permissive licenses	Yes	No	Production	Yes	Free
Tabby	Unclear	Yes	Yes	Production	Yes	Free

pragmaticengineer.com

Smals
ICT for society

(Source: Gergely Orosz, "Github CoPilot Alternatives", 16 May 2023)

Some open-source alternatives

- FauxPilot: an alternative backend to Github CoPilot *[note: possibly abandoned]*
  - Deploy a Salesforce CodeGen model yourself with nvidia-docker
  - Configure the official VSCode Copilot plugin to use your local server
- Tabby: “self-hosted, open-source and on-premises alternative to GitHub Copilot”
  - Completion models: StarCoder, CodeLlama, DeepSeekCoder
  - Chat models: WizardCoder, Mistral

- llm-vscode / llm-intellij: developed by Huggingface
  - Originally for StarCoder, now also supports CodeLlama, Phind and WizardCoder
  - Includes handy (accidental) plagiarism checker

- Continue.dev : generic plugin for multiple open source and commercial models
  - Work locally with ollama, LM Studio, Llama.cpp, LocalAI etc.
  - Use larger or self-finetuned models through Replicate.ai, or Huggingface on AWS, GCP, Azure, …

Smals ICT for society

Benchmarks

For Open-source models, leaderboards exist: *[Note: today, closed models (OpenAI family) still perform more than 2x better]*

T	Models	Win Rate	humaneval-python	java	javascript	Throughput (tokens/s)
🔶	Phind-CodeLlama-34B-v2	35.96	71.95	54.06	65.34	15.1
🔶	Phind-CodeLlama-34B-v1	35.42	65.85	49.47	64.45	15.1
🔴	DeepSeek-Coder-33b-base	34.92	52.45	43.77	51.28	
🔶	Phind-CodeLlama-34B-Python-v1	34.65	70.22	48.72	66.24	15.1
🔶	WizardCoder-Python-34B-V1.0	33.5	70.73	44.94	55.28	15.1
🔴	DeepSeek-Coder-7b-base	31.67	45.83	37.72	45.9	
🔶	CodeLlama-34b-Instruct	30.81	50.79	41.53	45.85	15.1
🔶	WizardCoder-Python-13B-V1.0	30.35	62.19	41.77	48.45	25.3
🟢	CodeLlama-34b	30.19	45.11	40.19	41.66	15.1
🟢	CodeLlama-34b-Python	29.5	53.29	39.46	44.72	15.1
🔶	WizardCoder-15B-V1.0	28.77	58.12	35.77	41.91	43.7
🔶	CodeLlama-13b-Instruct	27.73	50.6	33.99	40.92	25.3

Benchmarks

Commercial providers report their own benchmarks. Best consider them **selective** and **often not reproducible**:

Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

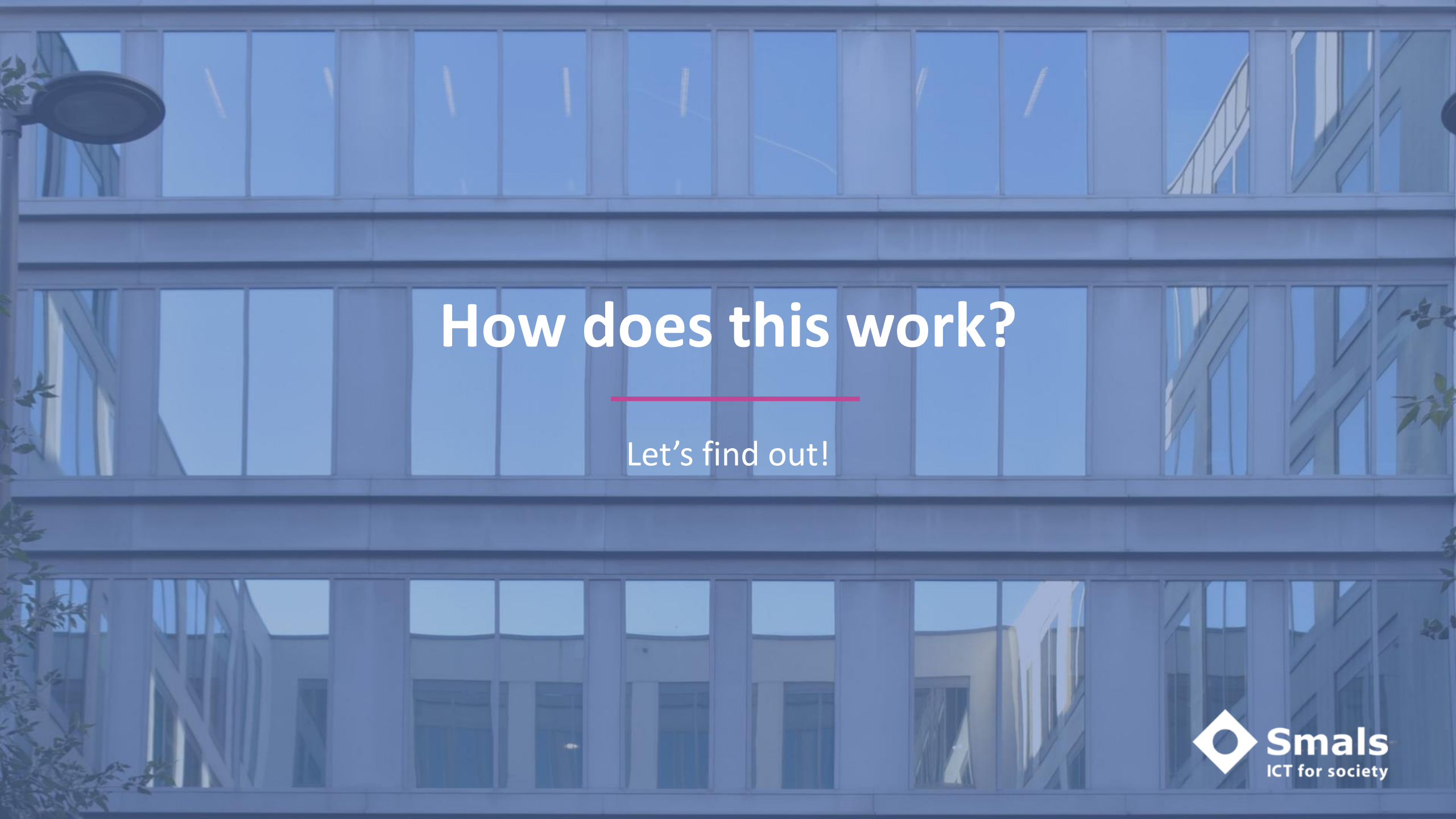
Table 1. Averaged performance on grouped benchmarks compared to popular open-source SLMs.

Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Table 2. Comparison between Phi-2 and Gemini Nano 2 Model on Gemini's reported benchmarks.

Smals ICT for society

(Source: Microsoft Research Blog on the release of Phi-2, 12/12/2023)



How does this work?

Let’s find out!

Smals
ICT for society

(Large) Language Models

- **A generative model predicts plausible next word(s) = tokens :**

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

Out[ ]= {The best thing about AI is its ability to,
The best thing about AI is its ability to learn,
The best thing about AI is its ability to learn from,
The best thing about AI is its ability to learn from experience,
The best thing about AI is its ability to learn from experience.,
The best thing about AI is its ability to learn from experience. It,
The best thing about AI is its ability to learn from experience. It's,
The best thing about AI is its ability to learn from experience. It's not}

- **… and allows for some variation (= "temperature") :**

The best thing about AI is its ability to learn. I've always liked the

The best thing about AI is its ability to really come into your world and just

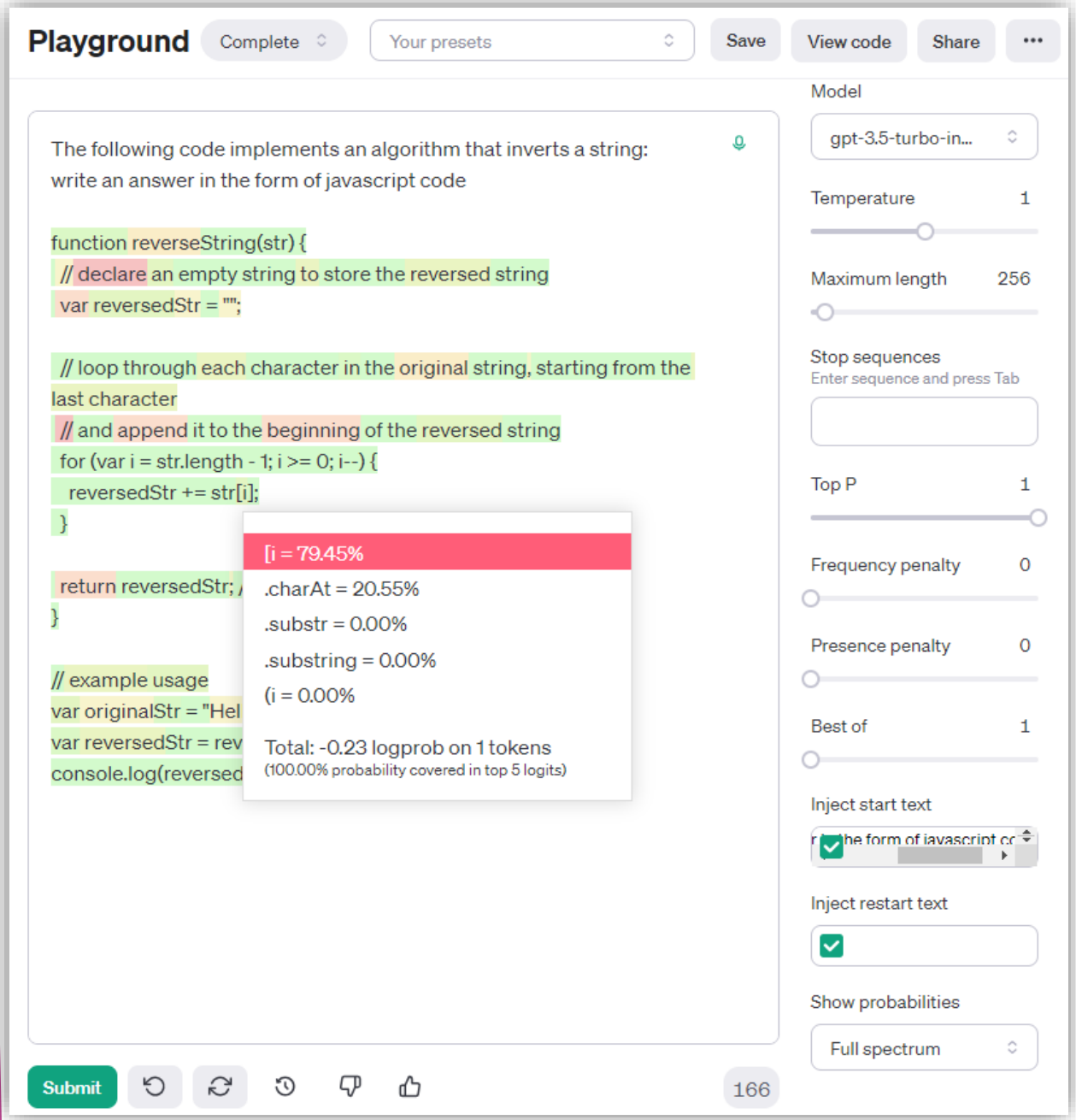
The best thing about AI is its ability to examine human behavior and the way it

The best thing about AI is its ability to do a great job of teaching us

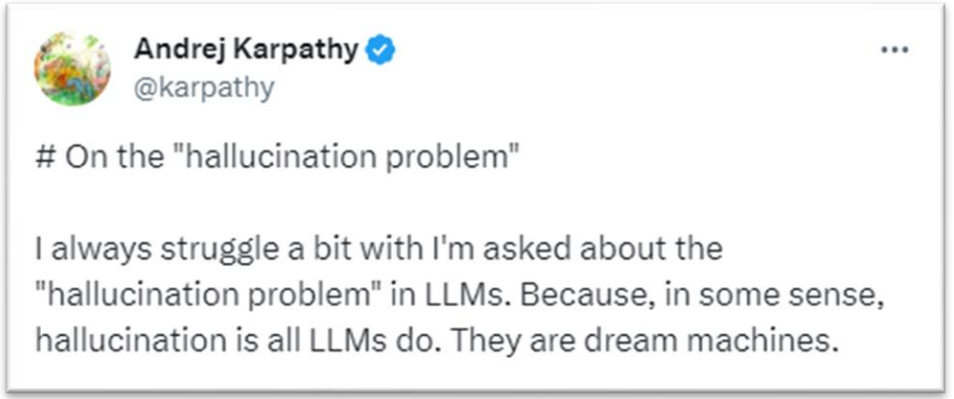
The best thing about AI is its ability to create real tasks, but you can

Smals ICT for society

(Images © Stephen Wolfram, "What is chatGPT doing and why does it work", 14/02/2023, MIT license)



Hallucation



- The LLMs at the core of these tools, generate suggestions probabilistically.
- Any logic, determinism, or reasoning, is added around the LLM, i.e. through pre/post-processing routines.

Smals ICT for society

Multi-Language performance differences

- Challenge: Given a description and function header, fill in the function
- Success: passes all tests and assertions at first try (mean **pass@1**) – scores are %
- Programming puzzles datasets - not necessarily representative for actual code
- Note: results for other languages obtained by transpiling Python to other languages

	CPP	C#	Java	JavaScript	Python	Shell	TypeScript
CodeGen-16B-Multi	21	8.24	22.2	19.15	19.26	0.61	20.07
CodeGeeX	16.87	8.49	19.14	16.92	21.62	2.75	10.11
Github CoPilot v1	30.59	22.06	31.9	31.27	30.71	11.74	31.26
StarCoderBase (generic)	30.56	20.56	28.53	31.7	30.35	11.2	32.15
StarCoder (Python)	31.55	21.01	30.22	30.79	33.57	10.46	32.29
Github CoPilot v2 *	48.44	27.47	40.12	48.99	46.68	23.24	48.87

* CoPilot v2 = code-davinci-002 model, accessed through API.

Source: StarCoder paper - data from 9 May 2023 or before.

Code is different from text

- To improve robustness, training datasets for coding LLMs may be **annotated** with additional signaling tokens.
- A corresponding IDE plugin may need to do similar **preprocessing** on the request before sending it out to the model.
- Changing models may require changing **tokenizers** in the IDE plugin.
- The plugin or interface needs to process these tokens if they appear in the model's response.

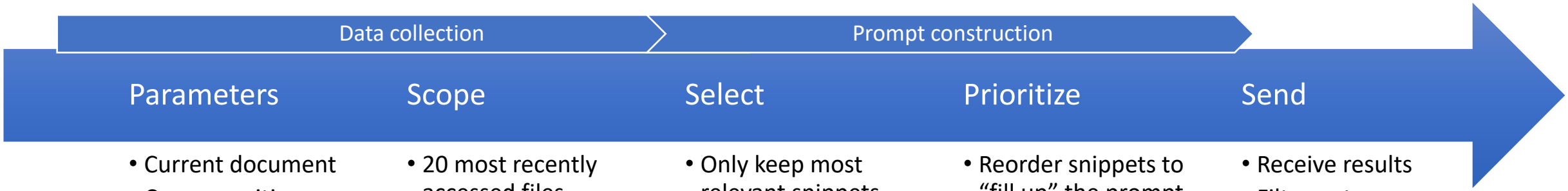
Token	Description
<|endoftext|>	end of text/sequence
<fim_prefix>	FIM prefix
<fim_middle>	FIM middle
<fim_suffix>	FIM suffix
<fim_pad>	FIM pad
<reponame>	repository name
<filename>	file name
<gh_stars>	GitHub stars
<issue_start>	start of GitHub issue
<issue_comment>	start of GitHub issue comment
<issue_closed>	GitHub issue closed event
<jupyter_start>	start of Jupyter notebook
<jupyter_text>	start of Jupyter text cell
<jupyter_code>	start of Jupyter code cell
<jupyter_output>	start of Jupyter output cell
<empty_output>	output cell without content
<commit_before>	code snippet before commit
<commit_msg>	commit message
<commit_after>	code snippet after commit

Table 10: Overview of the sentinel tokens.

(Signaling tokens used in the StarCoder model.)

Behind the scenes

How a request for code suggestions works (based on a reverse-engineering of the Github CoPilot plugin) :

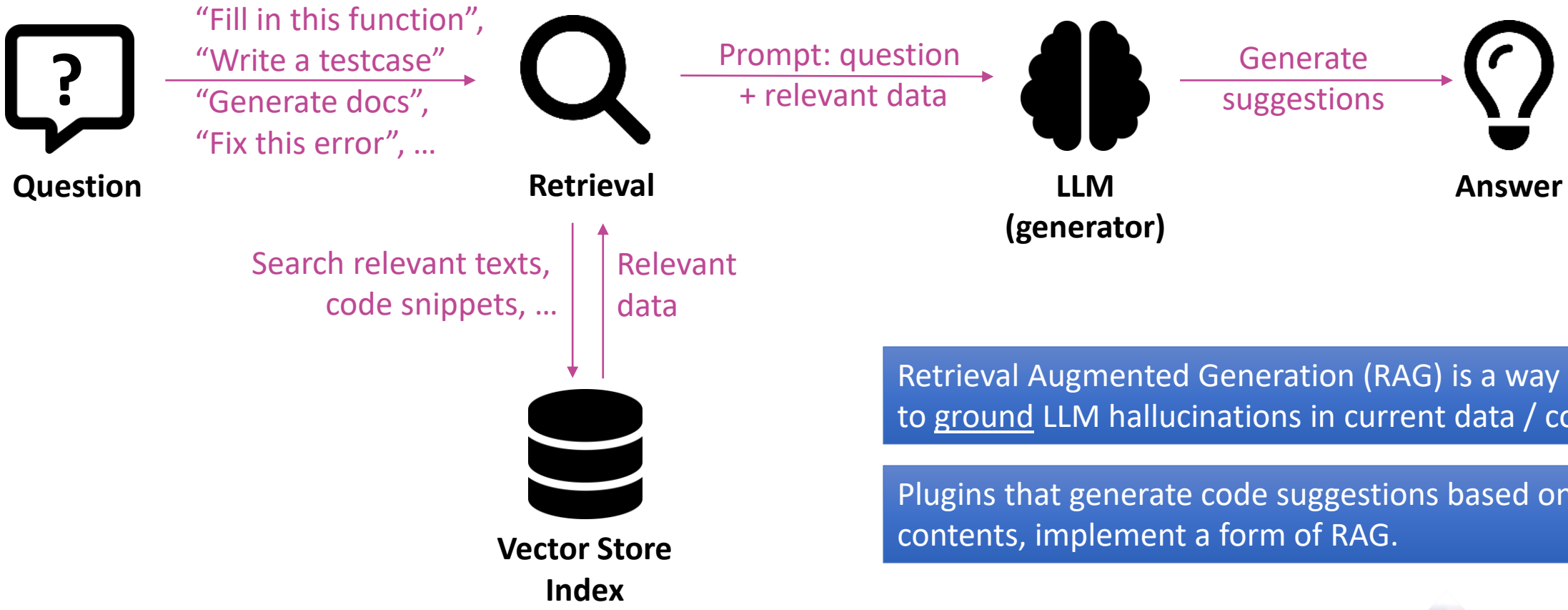


+ Telemetry : after acceptance, checks if inserted code is still present after 15sec, 30sec, 1min, 2min, 5min, ...

→ Used to collect data/feedback for future finetuning (OPT-OUT is possible.)

Smals ICT for society

Retrieval Augmented Generation (RAG)

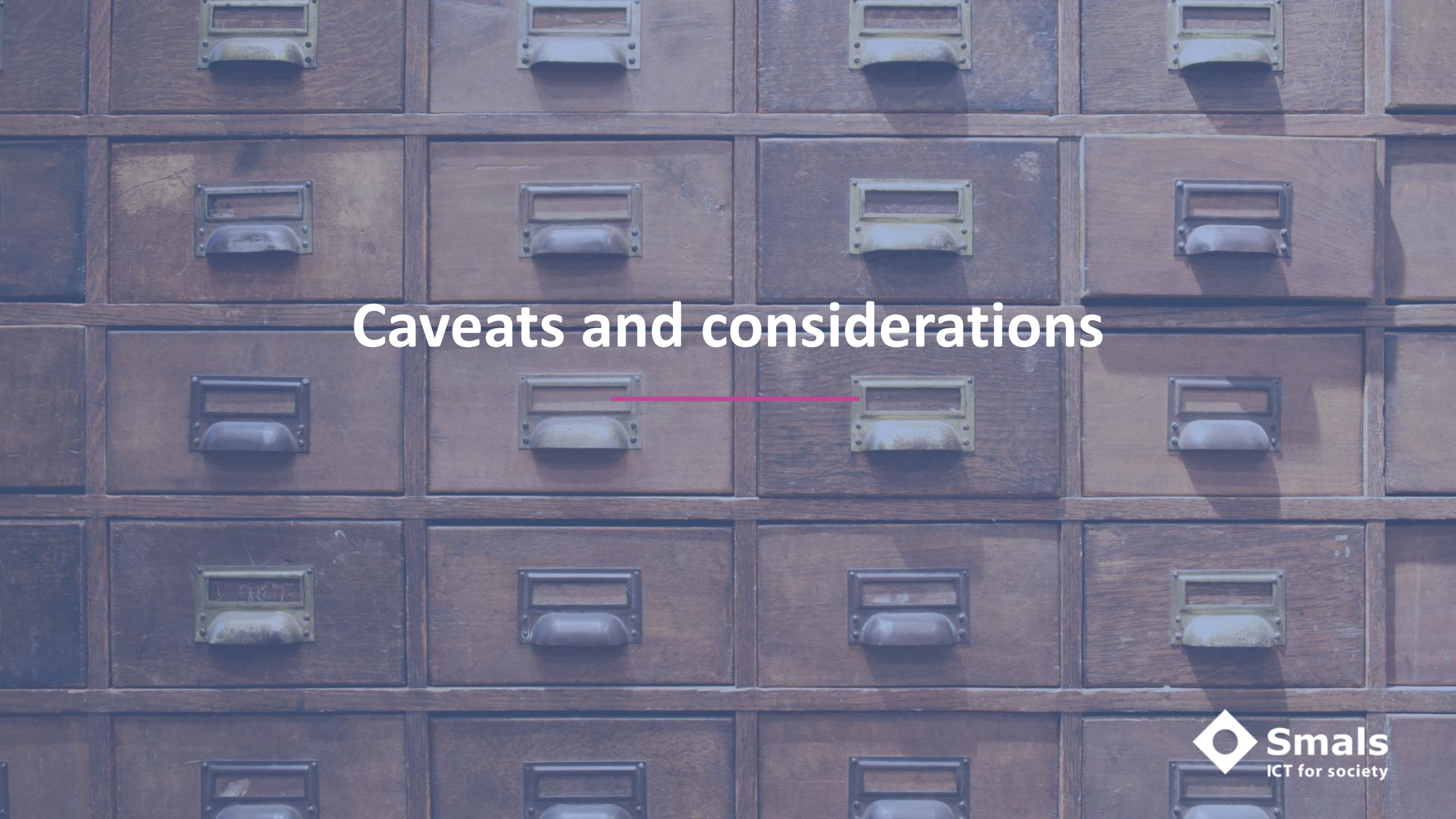


Retrieval Augmented Generation (RAG) is a way to ground LLM hallucinations in current data / context.

Plugins that generate code suggestions based on IDE contents, implement a form of RAG.

(Schematic design by Bert Vanhalst)

Smals ICT for society



Caveats and considerations

Smals
ICT for society

AI Coding assistant SWOT

Strenghts

- Faster project setup, scaffolding and boilerplate
- Faster onboarding of new developers
- Less time lost “tinkering”
- Follows existing codebase patterns/style
- Generate code from description
- Explain or generate comments/docs from code
- Ease of use

Weaknesses

- Tends to follow existing (non-optimal) patterns
- No guarantee on correctness/optimality
- May generate code beyond programmer’s understanding / obscure solutions
- Works better for popular programming languages
- Resource-hungry technology
- Self-hosting = specialized hardware + knowledge

Opportunities

- Offer modern tooling to developers
- Increased overall productivity and quality
- Increased developer flexibility/mobility
- Resolve issues/incidents/bugs faster
- Underlying technology will evolve even further

Threats

- Confidentiality: sending code to cloud
- Legal (copyright, plagiarism, liability, GDPR)
- Overreliance, leading to increased technical debt or long-term maintainability problems
- Businesses are very young and still taking shape.

Smals ICT for society

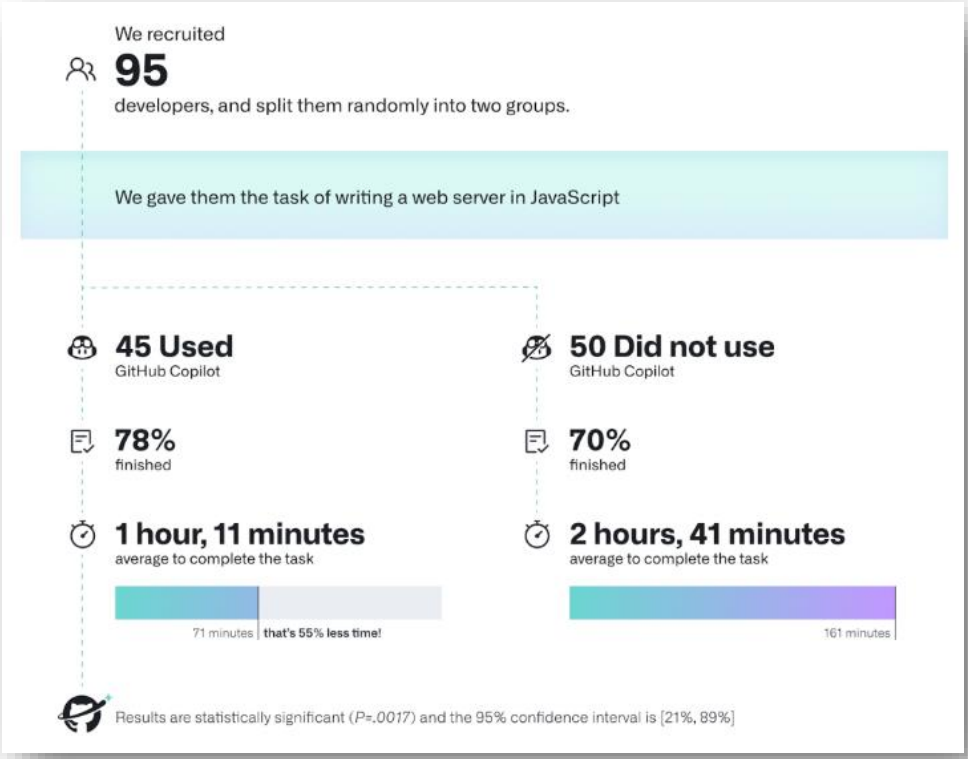
Correctness

Smals ICT for society

Productivity: ⚠ your mileage may vary ⚠

- Code writing is a relatively small part (10-20%) of the time spent on a project.
- Time gained in coding may be lost again in debugging, deciphering, understanding, explaining…
- Reported measurements so far are mostly **anecdotal** :

	Task 1 - Easy		Task 2 - Medium		Task 3 - Hard	
	Intellisense	Copilot	Intellisense	Copilot	Intellisense	Copilot
	9:35	1:46	7:48	12:53	13:41	11:08
	3:50	3:57	15:52	16:45	13:43	11:05
	4:49	4:55	16:28	7:26	22:42	4:04
	9:04	6:18	14:16	15:05	13:06	DNF
	5:18	1:18	7:35	13:24	23:13	19:54
	15:54	7:52	12:39	DNF	4:48	DNF
	5:27	3:12	10:47	6:02	DNF	DNF
	2:09	20:12	8:30	DNF	DNF	9:19
Average Time	7:01	6:11	11:44	11:56	13:36	11:06
Overall average time for all tasks combined					10:23	9:18



(Sources: P. Vaithilingam, CHI2022 , and Github Blog)

Intellectual property

TECH / MICROSOFT / COPYRIGHT

GitHub's automatic coding tool rests on untested legal ground

Photo Illustration by Pavlo Gonchar / SOPA Images / LightRocket via Getty Images

/ The Copilot tool has been trained on mountains of publicly available code

By DAVE GERSHGORN

Jul 7, 2021, 2:00 PM GMT+2 | 0 Comments / 0 New

See also https://hackernoon.com/doe-v-github-original-complaint-pertaining-to-copyright-infringement-open-source-licenses-and-more

Smals ICT for society

Copyright / ownership

- AI-specific legislation is coming slowly. In the meantime, **existing law is in full effect**!
- Current prevailing opinion is that only humans can claim patent rights or copyright.
  - Cfr. EPO decisions of 27 January 2020 on EP 18275163 and EP 18275174 , later confirmed by Board of Appeal.
  - Cfr. Congressional Research Service report LSB10922.
  - EU: copyright claims generally require “original contribution through intellectual creation”.
  - compare: Nikon can’t claim copyright on photos taken by its cameras.
- Most AI providers do not claim copyright on generated output - but don’t grant exclusive ownership either.
- **Check the terms and conditions**, they may:
  - Grant the AI maker a free license to reuse output and/or input , e.g. “for service improvements”.
  - Demand the user to have full rights to share anything they may submit.
- When in doubt, **contact your legal department!**

Example: Microsoft

8. **Ownership of Content.** Microsoft does not claim ownership of Captions, Prompts, Creations, or any other content you provide, post, input, or submit to, or receive from, the Online Services (including feedback and suggestions). However, by using the Online Services, posting, uploading, inputting, providing or submitting content you are granting Microsoft, its affiliated companies and third party partners permission to use the Captions, Prompts, Creations, and related content in connection with the operation of its businesses (including, without limitation, all Microsoft Services), including, without limitation, the license rights to: copy, distribute, transmit, publicly display, publicly perform, reproduce, edit, translate and reformat the Captions, Prompts, Creations, and other content you provide; and the right to sublicense such rights to any supplier of the Online Services.

   No compensation will be paid with respect to the use of your content, as provided herein. Microsoft is under no obligation to post or use any content you may provide, and Microsoft may remove any content at any time in its sole discretion.

   You warrant and represent that you own or otherwise control all of the rights to your content as described in these Terms of Use including, without limitation, all the rights necessary for you to provide, post, upload, input or submit the content.

Security

- Benchmark "Asleep at the keyboard", a code completion challenge
  - 89 code generation scenarios based on **MITRE top-25 vulnerabilities** list
  - Success 1 : generated code is Valid (= compiles) (**pass@1**) – scores are %
  - Success 2 : Valid code is Not Insecure - scores are % relative to Valid code
- Conclusion: better models generate more syntactically correct code, but not more secure code!
- Unsafe code still needs to be identified by the user, better alternatives needs to be asked for explicitly.

	Valid (higher is better)	Insecure (lower is better)
InCoder-6B	87.1	**35.48**
CodeGen-16B-Multi	95.5	43.25
Github CoPilot v1	96.4	42.32
StarCoderBase (generic)	85.5	39.77
Github CoPilot v2 *	**98.4**	42.99

* CoPilot v2 = code-davinci-002 model, accessed through API.

Source: StarCoder paper

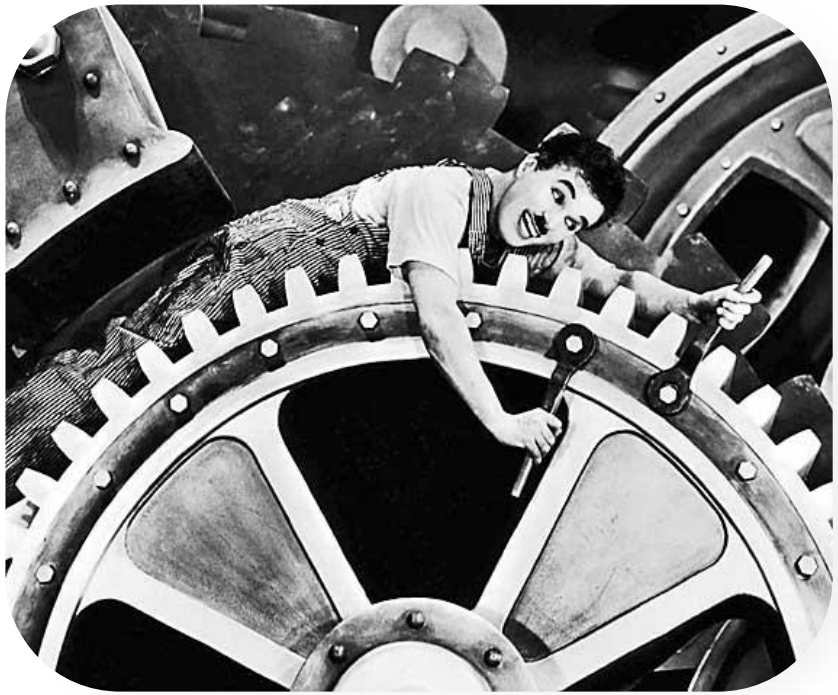
Smals ICT for society

AI in the workplace

- AI won't replace people – but people who use AI
  will replace people who don't.
  (IBM, "Augmented work for an automated, AI-driven world")

- AI assistants will become part of default office software
  (ref. Microsoft 365 Copilot announcement)

- The content of software development jobs may change too:
  - Less writing / creating from scratch
  - More reviewing



Charlie Chaplin, "Modern Times", 1936

Smals ICT for society

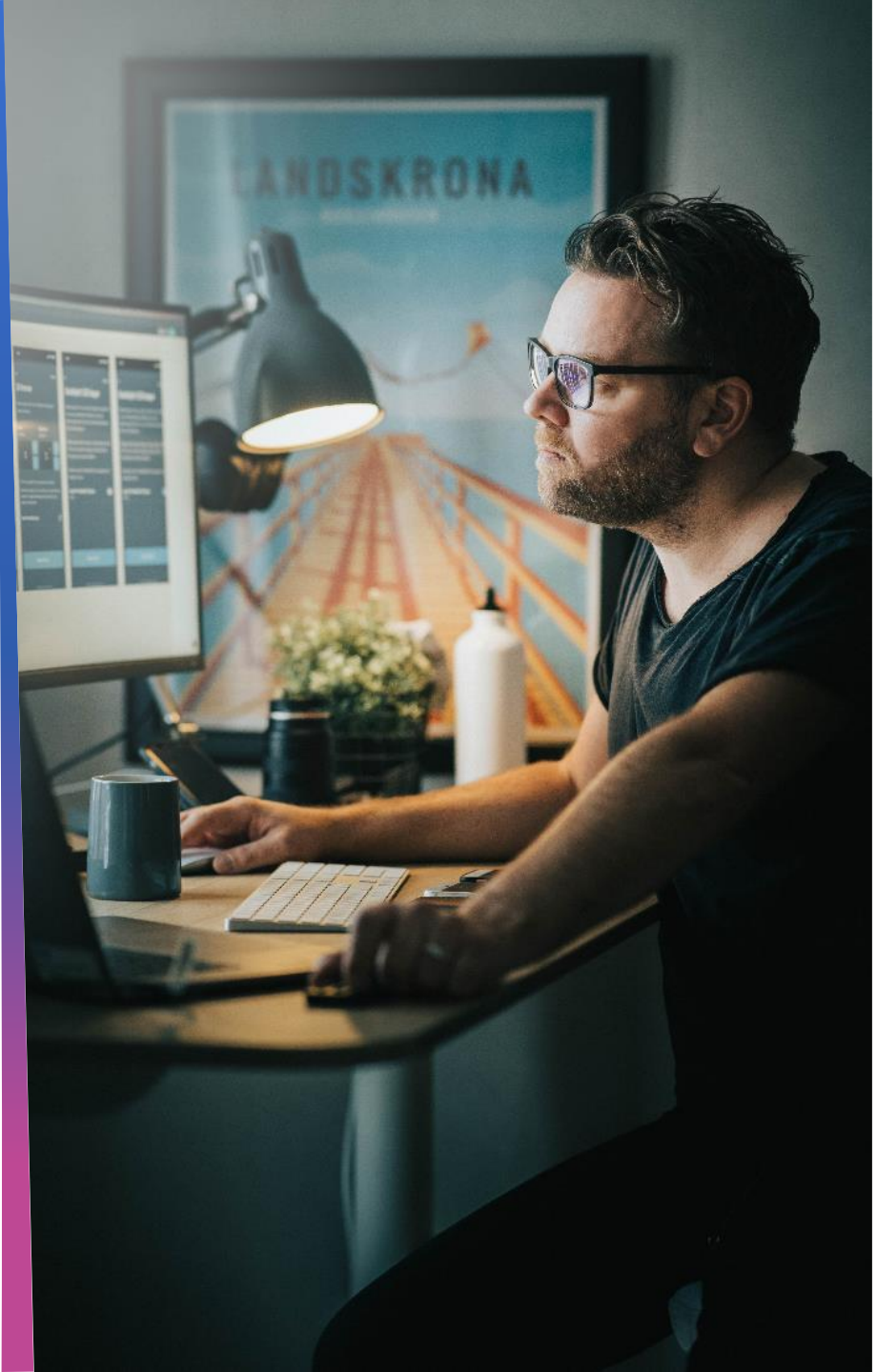
Conclusion

and what's next?

Smals
ICT for society

General observations

- Soon, **coding without assistant will feel like writing without spellchecker**.
- Closed-source cloud-based models currently offer best performance (extremely large models).
- Interesting developments continuing in the open source community:
  - Smaller models or specialized models,
  - Increased attention for dataset quality,
  - Efforts to develop tools to deploy on-premise or on-device (GPT4All, ollama, LM Studio, …)
- Leaderboards offer insight in latest performance.
- Keep in mind:
  - Avoid overreliance.
  - Your full codebase is shared with the coding assistant.
  - Best results when working incrementally in small steps.
  - Treat the results with caution.
  - No guarantees!



Good practices

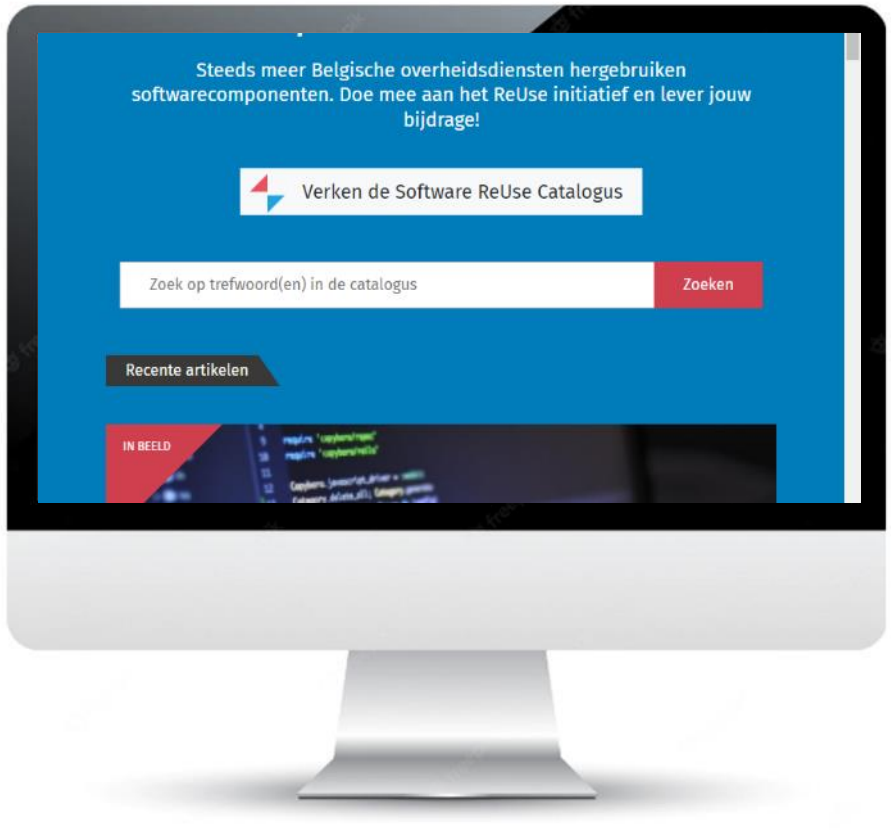
- Do not put sensitive information (personal info, keys, tokens, credentials, access codes, passwords, …) in any files that are read by a coding assistant.
- Mark generated code as such.
- Thoroughly review, document and test all generated code.
- Do not accept any code into a codebase that you don’t understand.
- Adhere to high standards. Here, too, **garbage in = garbage out**.

→ Do not use blindly! Be aware of safety, confidentiality, legal, … issues.

Smals ICT for society

@ Smals

- Main idea: **let's not stay behind, but let's do this wisely**
- First guidelines: see org.com communication on 20/11/2023
  - A more detailed policy is under construction
- A new Competence Center on (generative) AI
- Inventory of tools currently in use
- Pilot with test licenses for a coding assistant, to be used by one or more dev-teams
- POC with AI tools that can help project managers and analysts with their tasks
- Contact: GenerativeAIGovBoard@smals.be



Smals ICT for society

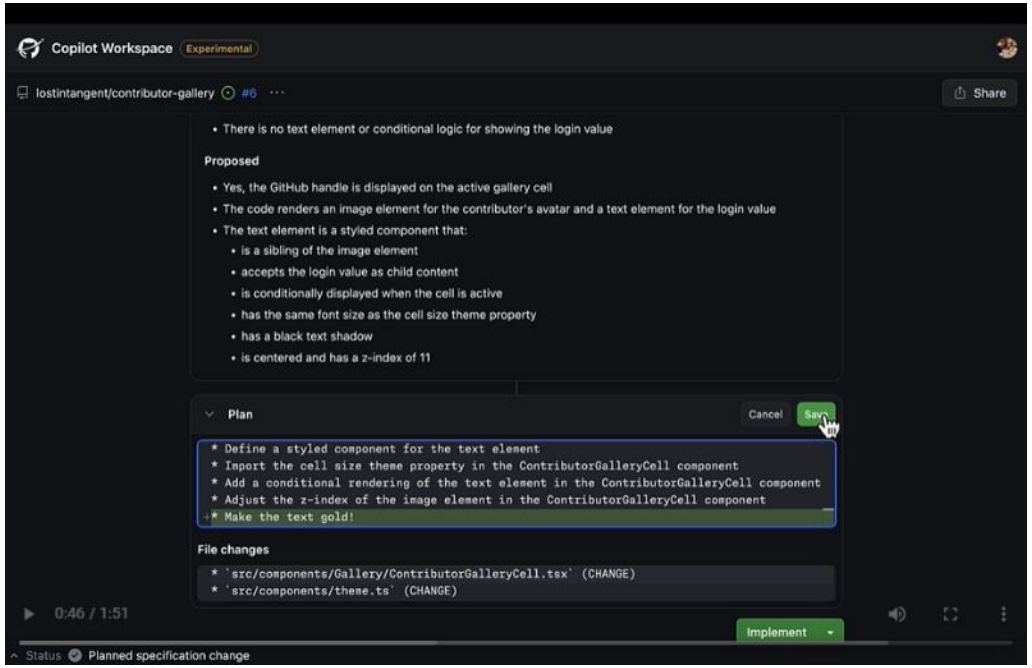
Developments in the last weeks:

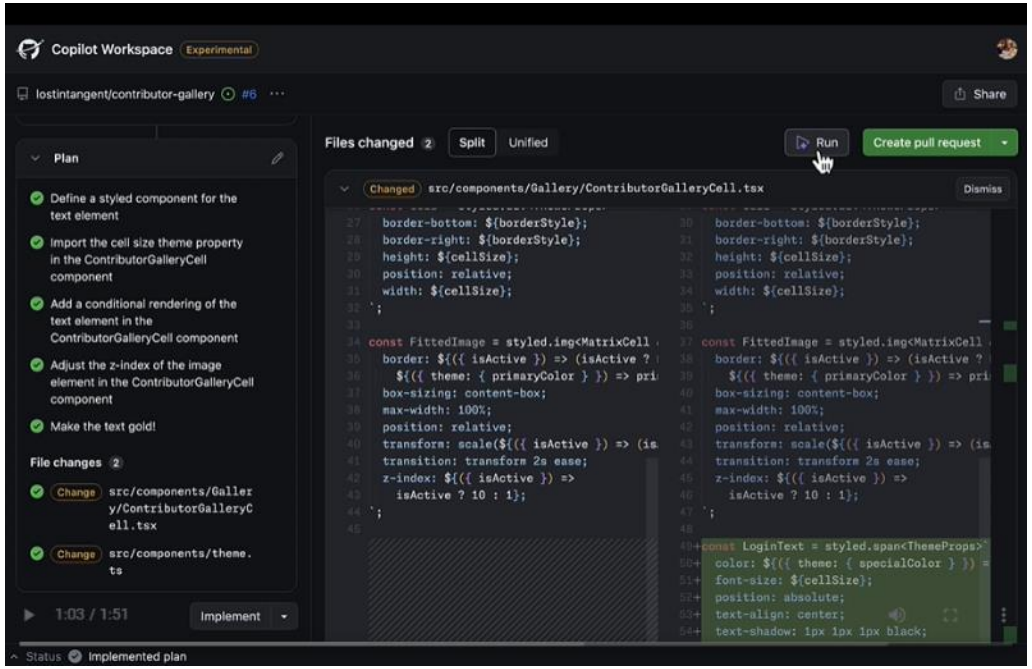
- Mistral 8x7B release , 32k context length, performs on par with LLama2-70B
- Smaller, condensed or specialized models,
  - Phi-2 for Python coding assistance (Microsoft): 2.7B parameters
  - Gemini Nano2 (Google): 3.2B parameters, included in Pixel 8 Pro smartphones

  [small is relative: *"The training for Phi-2 took 14 days on 96 A100 GPUs […] on 1.4T tokens."*]
- Further initiatives to improve reproducibility: LLM360 (UAE)
- Smarter prompt engineering: MedPrompt / Promptbase (Microsoft) , prompt compression, …
- Fine-tuning with generated data is further explored
- Function Calling in open source models: Gorilla OpenFunctions
- AI Act is slowly moving forward
- A new applied AI R&D initiative: Answer.ai and a new AI Alliance (IBM, Meta and 50 major vendors)
- …

2024 : Copilot-in-Everything?

- Integration in Microsoft 365 Teams, for which one can create plugins/bots with the Teams AI library,
- Github: CoPilot Workspace, providing an integrated flow: issue → generated solution plan → generated code





- Still many issues! Standardization, compatibility, communication between plugins / tools, …

Smals ICT for society

Thank you !

joachim.ganseman@smals.be
www.smalsresearch.be

https://www.linkedin.com/in/jganseman/

Smals
ICT for society

Further Reading

Articles:

- De AI Augmented Developer [NL]
- LLMs voor Code [NL] | LLMs pour code [FR] | LLMs for code [EN]
- 1 jaar ChatGPT [NL] | ChatGPT a 1 an [FR] | ChatGPT's 1st birthday [EN]

Smals ICT for society