seqera

Building and testing scientific workflows with LLMs and AI agents

Creating systems that help scientists to build and run complex workflows

Phil Ewels
Senior Product Manager for OSS
phil.ewels@seqera.io

March 6, SciLifeLab Automation of research processes with LLMs and AI

Outline

01. Introduction

02. LLMs to answer questions

03. LLMs to write code

04. AI agents

05. Looking ahead

Introduction

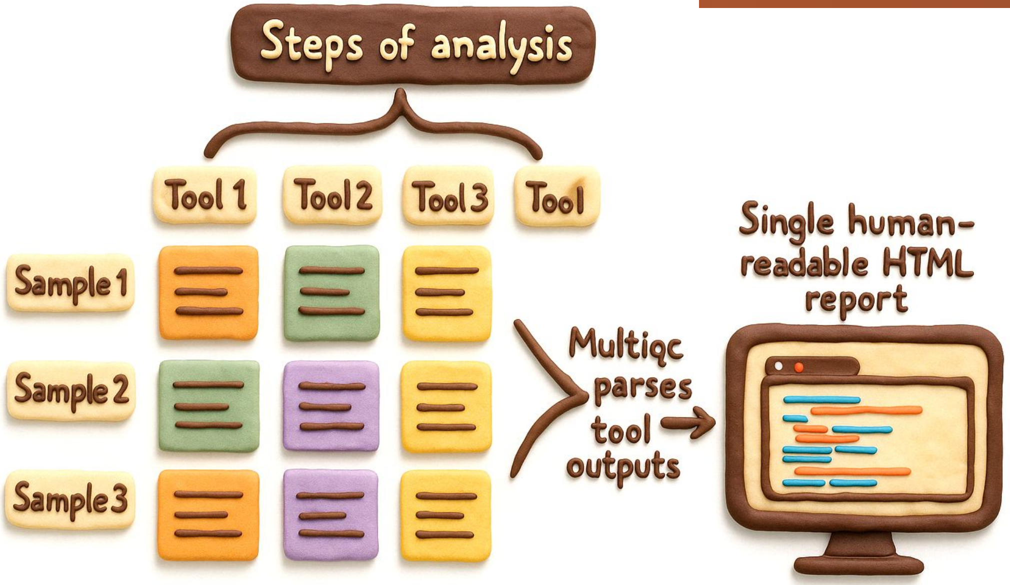
01

About me

- Scientist working in the lab in Cambridge (UK) in epigenetics
- Moved into bioinformatics (computational biology / data science)
- Moved to Sweden in 2014, SciLifeLab
- Started building software to handle huge volumes of data
- Joined Seqera in 2022
- Set up the community team, now product manager for open source

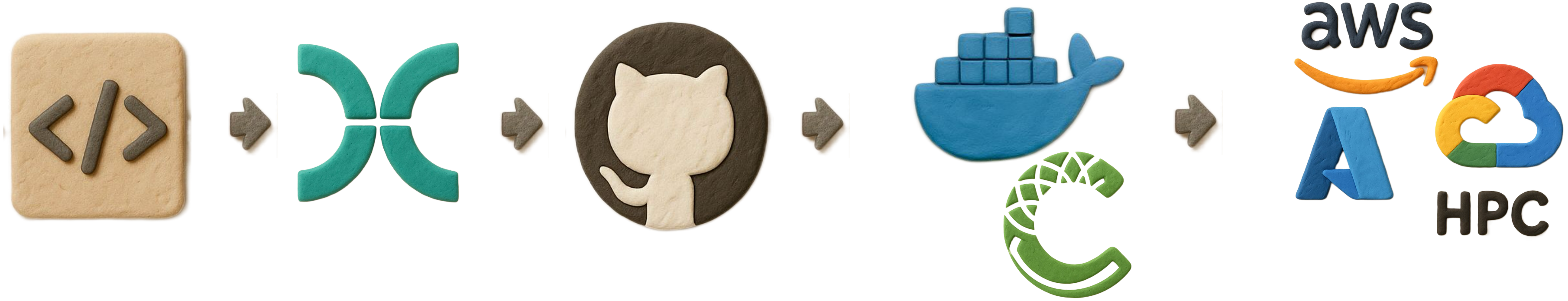
MultiQC

Open-source tool to aggregate bioinformatic analyses results.



Nextflow

Open-source orchestrator for writing & deploying workflows at scale.



Nextflow

Open-source orchestrator for writing & deploying workflows at scale.

- Programming language for bioinformatics workflows
- Orchestrator for executing workflows
- Runs on everything, with data from anywhere
- Reproducibility built-in, native support for software packaging and code version control

nextflow

Scalable

Portable

Reproducible

LLMs that can help

02

What’s the problem?

What do Nextflow users complain about the most?

Steep learning curve

It’s Groovy, baby

Syntax changed between versions

DSL1 → DSL2

Difficult / unusual errors

Generalists need not apply

Base models struggle with Nextflow

- Nextflow questions are a small subset of data science
- Answers are often very specific to Nextflow and not generally applicable
- Can't retrain the base model (too expensive)
- Can't include a long enough prompt (context window)

My Nextflow pipeline exited with an error in the FastQC process, saying exit code 127. What should I do?

Exit code 127 usually means a command not found error.

Check if it's installed:

```
fastqc --version
```

If it's not found, you need to install it, for example using conda:

```
conda install -c bioconda fastqc
```

Please don't tell people to do that.. 🤦🏻‍♂️

Base models struggle with writing Nextflow*

- Most public Nextflow code is old
- DSL1 → DSL2 syntax change in 2022+
- Hallucinates syntax from other (more common) languages
- Isn't as opinionated as we'd like it to be!

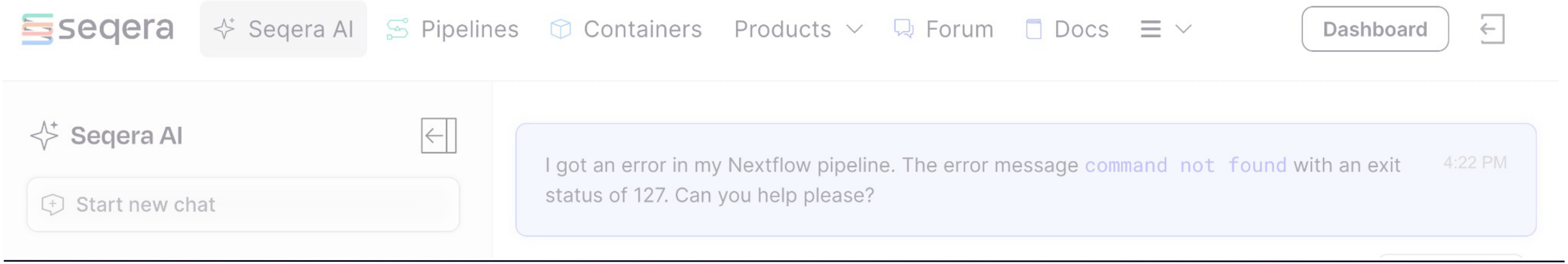
* Though they're getting better all the time!

Seqera AI

Bioinformatics AI that helps you get from 0 to 1 for all your omics

`https://seqera.io/ask-ai/`

- Latest Nextflow documentation is prioritised over everything else
- Also includes docs from nf-core, nf-test and other trusted sources
- Prioritises nf-core best practices and knows available pipelines



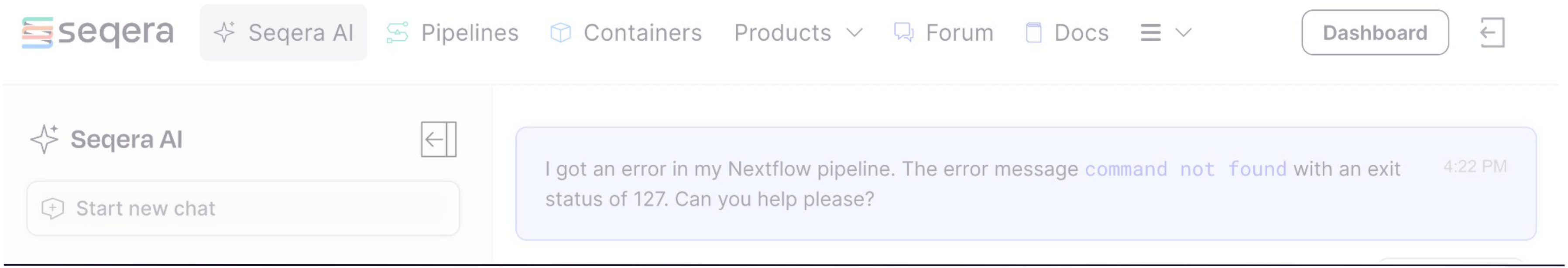
I got an error in my Nextflow pipeline. The error message `command not found` with an exit status of 127. Can you help please?

Seqera AI

Bioinformatics AI that helps you get from 0 to 1 for all your omics

`https://seqera.io/ask-ai/`

[demo]



I got an error in my Nextflow pipeline. The error message `command not found` with an exit status of 127. Can you help please?

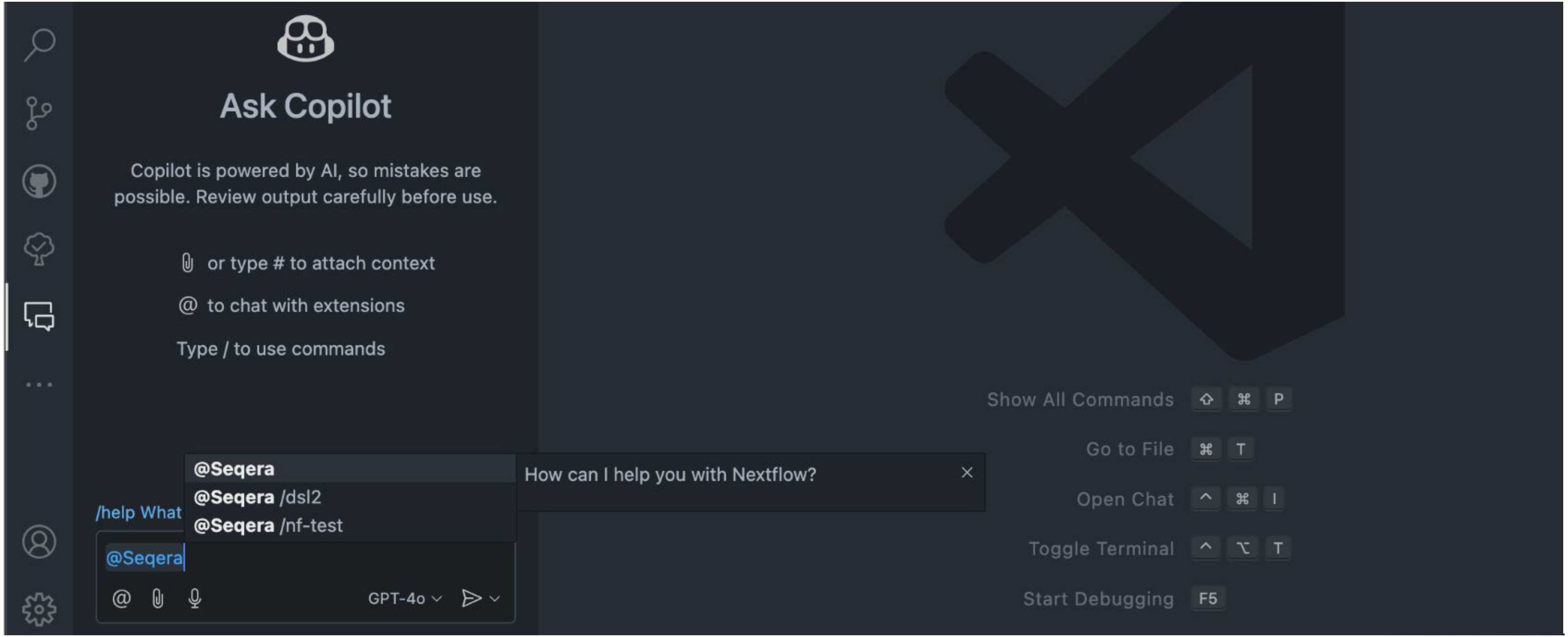
Tool integration

03

Seqera AI in VS Code

Nextflow help right where you need it

https://nextflow.io/vscode



Ask Copilot

Copilot is powered by AI, so mistakes are possible. Review output carefully before use.

or type # to attach context

@ to chat with extensions

Type / to use commands

/help What

@Seqera

@Seqera /dsl2

@Seqera /nf-test

How can I help you with Nextflow?

@Seqera

What's the problem?

What's most difficult about MultiQC reports?

Large sample numbers

Big data is..big

Understanding results

Is this wiggle normal?

Seeing the big picture

Detecting patterns

AI agents

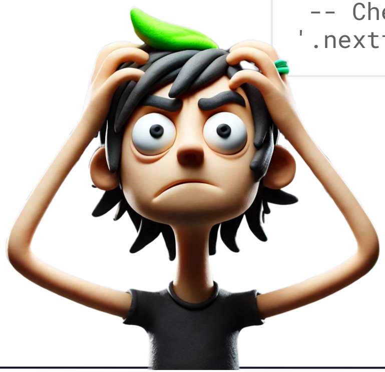
04

The problem with AI code generation

AI interns can create more work than they solve

- AI writes code, but doesn’t understand it
- Looks impressive until you try to run it
- Now you’re debugging an entire pipeline looking for the error, rather than iteratively writing + testing

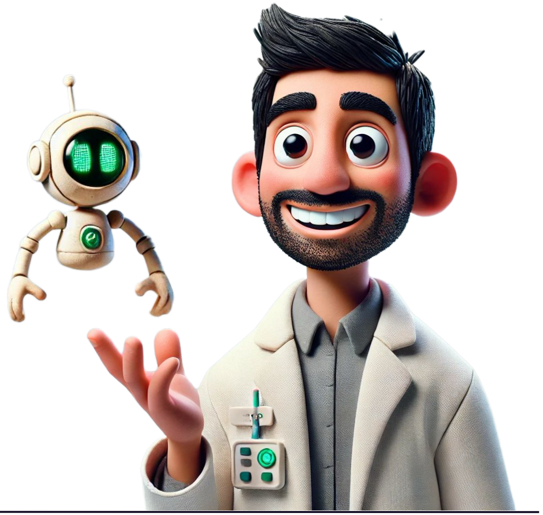
```
ERROR ~ No signature of method:
groovyx.gpars.dataflow.DataflowBroadcast.into() is
applicable for argument types:
(Script_85c4d82870d584dc$_runScript_closure1) values:
[Script_85c4d82870d584dc$_runScript_closure1@53aa2fc9]
Possible solutions: any(), find(), bind(java.lang.Object),
any(groovy.lang.Closure), find(groovy.lang.Closure),
is(java.lang.Object)

 -- Check script 'demo.nf' at line: 6 or see
'.nextflow.log' file for more details
```



Testing Nextflow code isn't trivial

- Proper modularity and code structure
- Finding example data
- Using the nf-test framework

- AI agents can help



AI agents

Going beyond prompts and responses

Agency Level	Description	What that's called	Example Pattern
☆☆☆	LLM output has no impact on program flow	Simple Processor	`process_llm_output(llm_response)`
★☆☆	LLM output determines an if/else switch	Router	`if llm_decision(): path_a() else: path_b()`
★★☆	LLM output determines function execution	Tool Call	`run_function(llm_chosen_tool, llm_chosen_args)`
★★★	LLM output controls iteration and program continuation	Multi-step Agent	`while llm_should_continue(): execute_next_step()`
★★★	One agentic workflow can start another agentic workflow	Multi-Agent	`if llm_trigger(): execute_agent()`

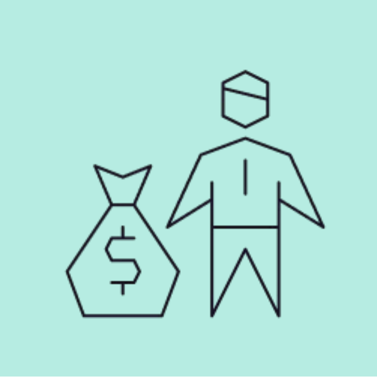
Source: `https://huggingface.co/docs/smolagents/en/conceptual_guides/intro_agents`

Looking ahead

06

Helpful LLMs

Approaches with Seqera AI



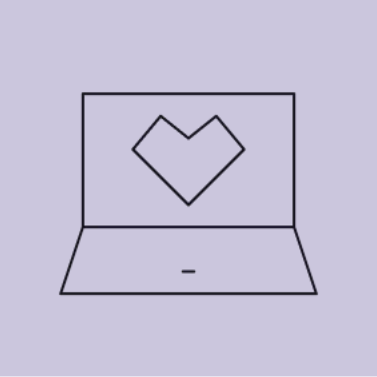
Use the big models

It’s impossible to win against the big model providers, better to use them: standardise and be flexible.

RAG and prompts

No-one else can do RAG (retrieval augmented generation) quite like you..

Fine-tuned LLMs can provide additional value over major providers.



Integration is king

Bringing LLMs out of a chat window and into applications helps to make them part of people’s workflows.

What’s next for Seqera AI

- Configure and launch pipelines
- Help when things go wrong
- Reduce the learning curve
- More integrations for AI agents
- Do all this in a way that is open, trusted and transparent

nf-core/rnaseq:
11 sections with 113 configurable fields

- Input/Output Options: 4
- Reference Genome Options: 21
- Read Trimming Options: 4
- Read Filtering Options: 5
- UMI Options: 9
- Alignment Options: 16
- Optional Outputs: 10
- Quality Control: 6
- Process Skipping Options: 20
- Institutional Config Options: 6
- Generic Options: 12

Looking ahead

- AI tooling is here to stay
- Make your content easy to find
- Build specialist tooling

- All scientists will become more and more like PIs, steering AI towards the most relevant work and approaches rather than doing the work directly

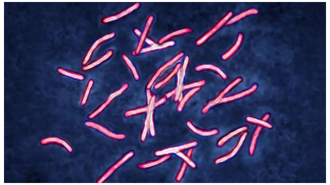
Looking ahead

AI cracks superbug problem in two days that took scientists years

20 February 2025

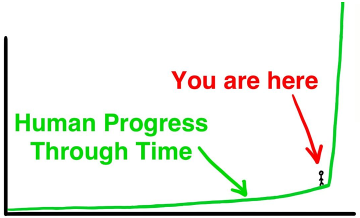
Share Save

Tom Gerken
Technology reporter



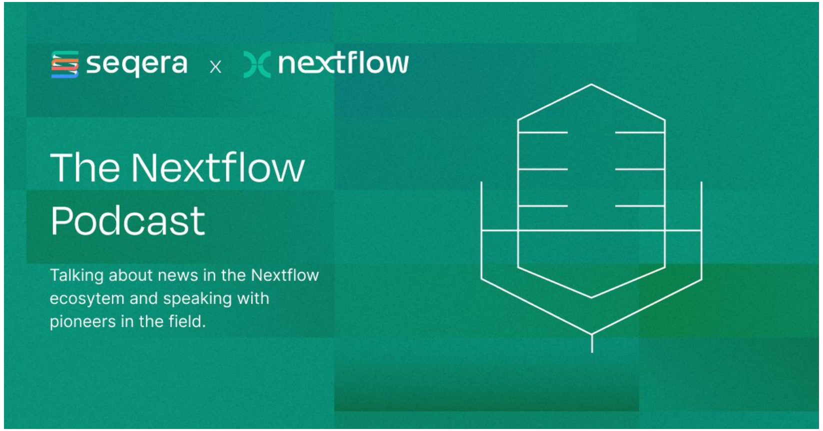
BBC

https://www.bbc.com/news/articles/clyz6e9edy3o



https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

Find out more



seqera x nextflow

The Nextflow Podcast

Talking about news in the Nextflow ecosytem and speaking with pioneers in the field.

nextflow SUMMIT

Barcelona 2024

OCT 28 - NOV 1, 2024

Join us for the latest developments and innovations from the Nextflow world.

https://seqera.io/podcasts/

https://summit.nextflow.io/2024/barcelona/

seqera

Phil Ewels
phil.ewels@seqera.io

Thank you

https://seqera.io/ask-ai/

https://docs.seqera.io/multiqc/ai

https://nextflow.io/vscode

nextflow

nextflow.io

nf-core

https://nf-co.re

multiqc

seqera.io/multiqc