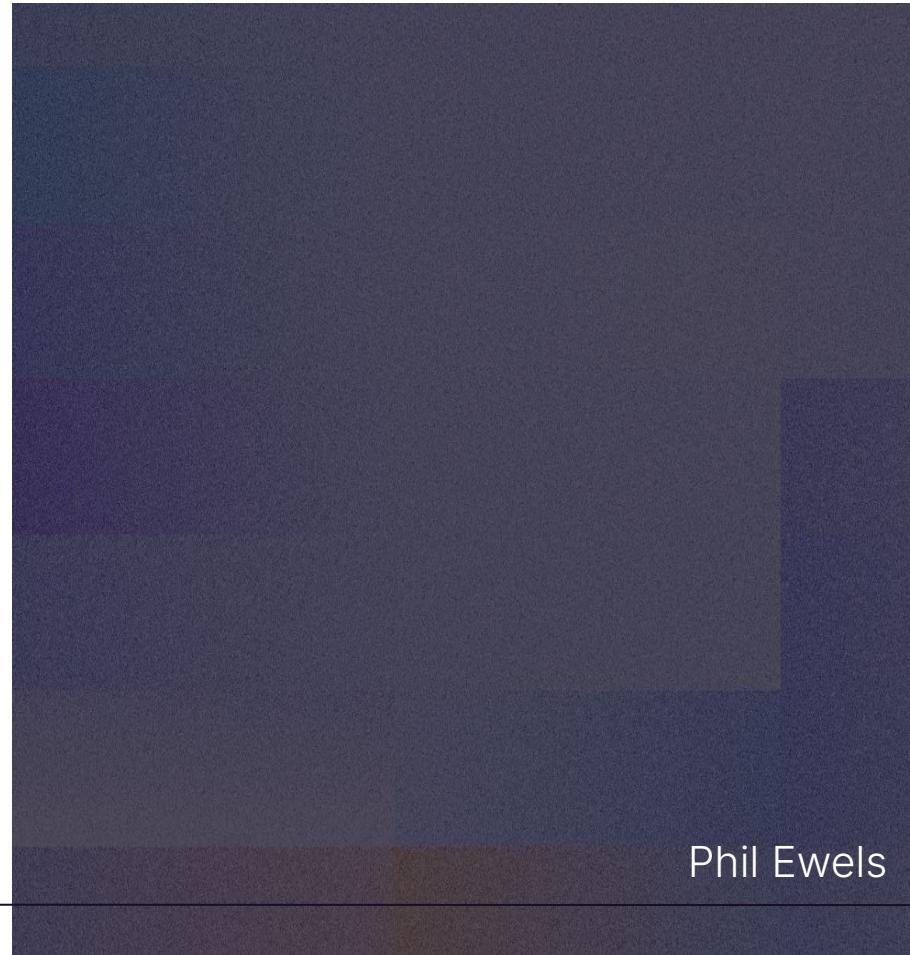




Building and testing scientific workflows with LLMs and AI agents

Creating systems that help scientists to build and run
complex workflows



Phil Ewels

Agenda

- 01.** Introduction
- 02.** LLMs to answer questions
- 03.** LLMs to write code
- 04.** AI agents
- 05.** Looking ahead



Introduction

01



About me

- Moved to Sweden in 2014, joined NGI at SciLifeLab
- Started building software to handle huge volumes of data
- Wrote MultiQC, adopted Nextflow, started nf-core
- Joined Seqera in 2022
- Set up the community team, now product manager for open source

The background is a dark blue gradient with several white and orange logos stacked vertically. From top to bottom, they are: Nextflow (green 'X' logo), MultiQC (orange 'Q' logo), nf-core (green 'nf' logo and a green apple icon), and Seqera (a stylized 'S' logo composed of red, orange, and blue segments). Below the logos, the name 'Phil Ewels' is written in white, followed by 'Senior Product Manager for OSS' and an email address 'phil.ewels@seqera.io'. To the right of this text is a portrait photo of a smiling man with short hair, wearing a yellow and blue jacket, set against a background of trees.

Phil Ewels

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

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



Scalable

Portable

Reproducible



LLMs that can help

02



ChatGPT struggles 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.. 

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

The screenshot shows the Seqera AI dashboard. At the top, there is a navigation bar with the Seqera logo, a search bar labeled "Seqera AI", and links for "Pipelines", "Containers", "Products", "Forum", "Docs", and a "Dashboard" button. Below the navigation bar is a sidebar with the "Seqera AI" logo and a "Start new chat" button. The main area is a chat window where a user has typed: "I got an error in my Nextflow pipeline. The error message command not found with an exit status of 127. Can you help please?". A timestamp "4:22 PM" is visible next to the message. The Seqera logo is also present at the bottom left of the page.



Seqera AI in MultiQC

Analysis help right where you need it

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

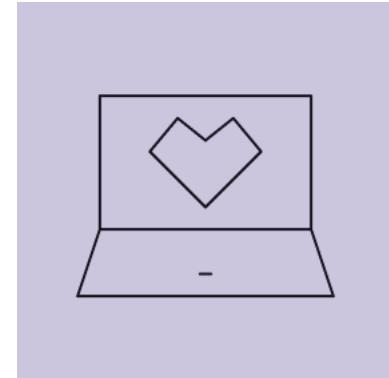
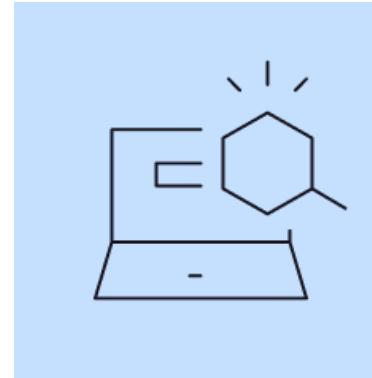
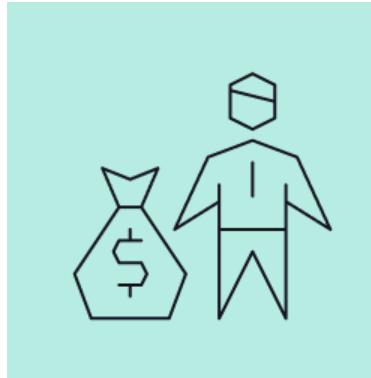
- Help from AI, in the context of a report
- Generation at report time, or on demand
- Flexible, can be used with most providers

The screenshot shows the MultiQC web interface. On the left, there's a sidebar with links: General Stats, STAR, Summary Statistics, and Alignment Scores. The main area features the MultiQC logo and a brief description: "A modular tool to aggregate results from bioinformatics analyses across many samples into a single report." Below this, text indicates the report was generated on 2025-01-21 at 23:53 CET based on data from /Users/ewels/GitHub/MultiQC/website/public/examples/rna-seq/data. A red box highlights the "Summarize Report" button, which has a small AI icon next to the text.



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.

LLMs that can write

03



Vanilla providers 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/>

The screenshot shows the Seqera AI web interface. At the top, there's a navigation bar with the Seqera logo, a search bar labeled "Seqera AI", and links for "Pipelines", "Containers", "Products", "Forum", "Docs", and "Dashboard". Below the navigation, the main area has a sidebar on the left with "Seqera AI" and "History" sections. The "History" section lists previous interactions: "Build an RNA-seq analysis pipeline for me" and "How do I edit modules.json ?". The main panel shows a conversation between the user and the AI. The user's message is: "Build an RNA-seq analysis pipeline for me". The AI's response is: "Certainly! I'd be happy to help you build an RNA-seq analysis pipeline. Before we start, I want to mention that there's an excellent, well-maintained pipeline available from nf-core called nf-core/rnaseq, which might be a great option for you. However, if you'd like to build a custom pipeline, I can help you with that as well." Below this, the AI continues: "Let's start by outlining the basic steps of an RNA-seq analysis pipeline and then we can dive into the Nextflow implementation. Here's a general workflow:".

Seqera AI

Pipelines Containers Products Forum Docs Dashboard

Start new chat

History

Today

Build an RNA-seq analysis pipeline for me

Ask community

Beta

Certainly! I'd be happy to help you build an RNA-seq analysis pipeline. Before we start, I want to mention that there's an excellent, well-maintained pipeline available from nf-core called nf-core/rnaseq, which might be a great option for you. However, if you'd like to build a custom pipeline, I can help you with that as well.

Let's start by outlining the basic steps of an RNA-seq analysis pipeline and then we can dive into the Nextflow implementation. Here's a general workflow:



Seqera AI

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

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



Start new chat

History

Today

Build an RNA-seq analysis pipeline for me

How do I edit modules.json ?

I got an error in my Nextflow pipeline.
 The error message 'command not found' with an exit status of

```
#!/usr/bin/env nextflow
nextflow.enable.dsl = 2

include { FASTQC } from './modules/fastqc'
include { TRIMMOMATIC } from './modules/trimmomatic'
include { STAR_ALIGN } from './modules/star_align'
include { FEATURECOUNTS } from './modules/featurecounts'
include { DESEQ2 } from './modules/deseq2'

params.reads = "path/to/reads/*_{1,2}.fastq.gz"
params.genome = "path/to/genome.fa"
params.gtf = "path/to/annotation.gtf"
params.outdir = "results"

workflow {
    read_pairs_ch = Channel.fromFilePairs(params.reads, checkIfExists: true)
    FASTQC(read_pairs_ch)
    TRIMMOMATIC(read_pairs_ch)
    STAR_ALIGN(TRIMMOMATIC.out, params.genome, params.gtf)
    FEATURECOUNTS(STAR_ALIGN.out, params.gtf)
    DESEQ2(FEATURECOUNTS.out.collect(), params.gtf)
}
```

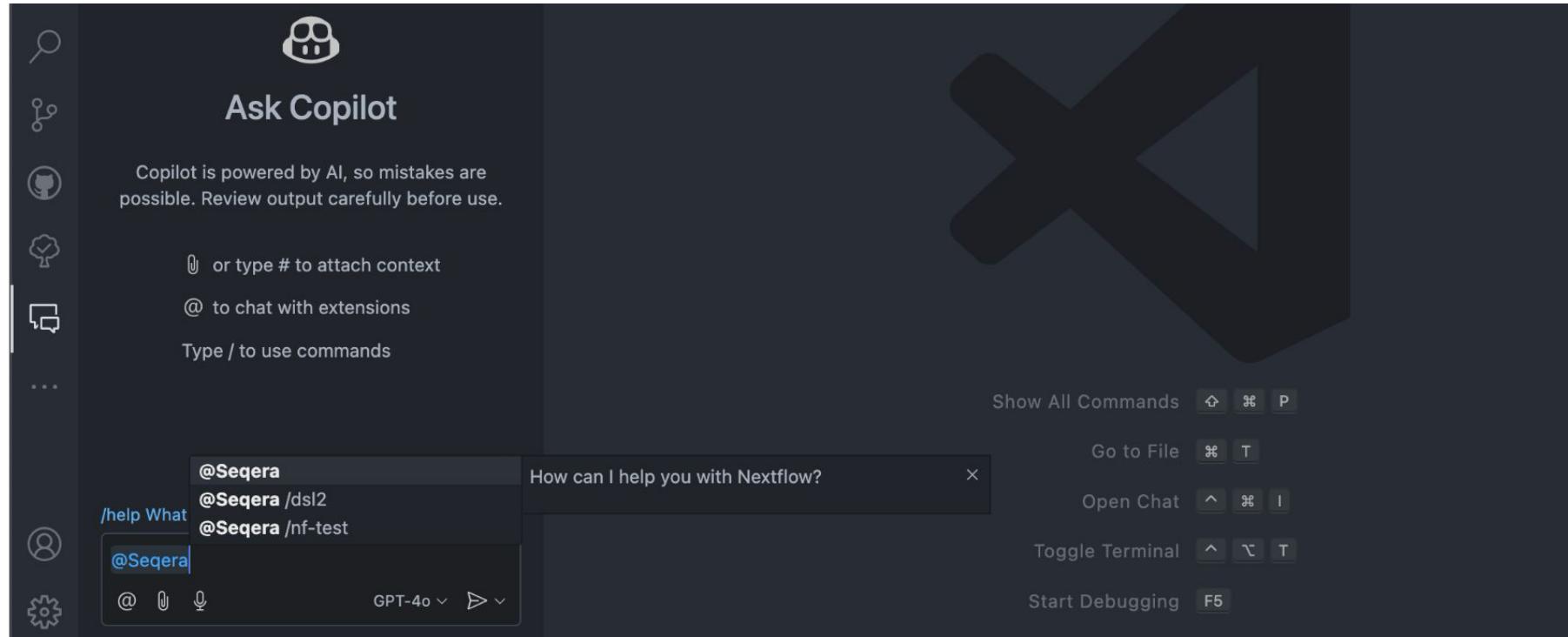


seqera.io

Seqera AI in VS Code

Nextflow help right where you need it

<https://nextflow.io/vscode>



seqera.io

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



modules/featurecounts.nf:

```
process FEATURECOUNTS {
    tag "$sample_id"
    publishDir "${params.outdir}/featurecounts", mode: 'copy'
    input:
        tuple val(sample_id), path(bam)
        path gtf
    output:
        tuple val(sample_id), path("${sample_id}.counts.txt"), emit: counts
    script:
        """
        featureCounts -a ${gtf} -o ${sample_id}.counts.txt -T ${task.cpus} ${bam}
        """
}
```

Script Testing

Generate tests that run on your local machine with lightweight nf-core sample datasets, or test this script in an AI sandbox environment.

AI sandbox test

Start test

Local testing guide

Generate guide

No test results yet. Click "Start test" to begin testing your script.



modules/deseq2.nf:

```
process DESEQ2 {
    publishDir "${params.outdir}/deseq2", mode: 'copy'
```

Send a message



Looking ahead

05



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

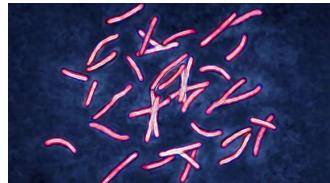


Looking ahead

AI cracks superbug problem in two days that took scientists years

20 February 2025

Tom Gerken
Technology reporter



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

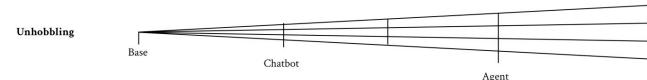
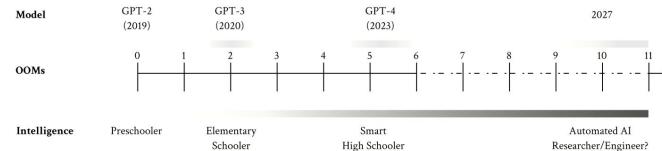
Share Save

SITUATIONAL AWARENESS

The Decade Ahead

<https://situational-awareness.ai/>

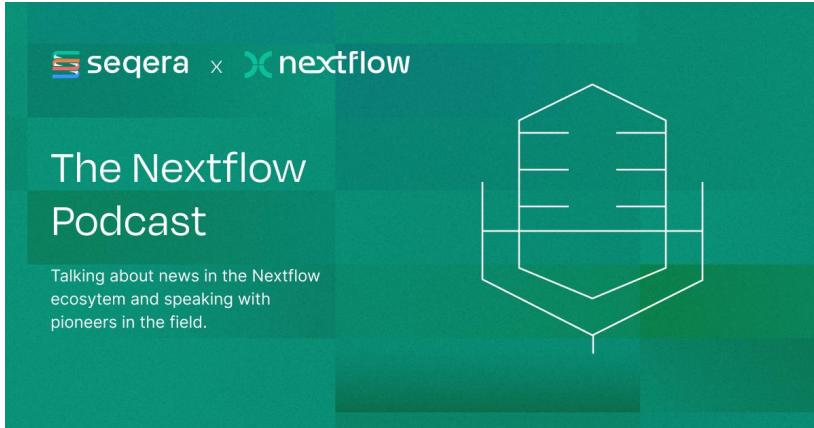
Counting the OOMs



SITUATIONAL AWARENESS | Leopold Aschenbrenner



Find out more



<https://seqera.io/podcasts/>



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



Thank you

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

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

<https://nextflow.io/vscode>



March 10

training.nextflow.io



March 24

<https://nf-co.re>



 May 12
 October 27

summit.nextflow.io