Cognition



004 Decisions, Decisions

University of Toronto · Oct 17

devkick

Agenda

11am-11:30 - Meet and greet

11:30-12:15 - Kev Presentation

12:15 - Lunch and/or build

1:30 - Demos

Demo Site

https://yohann-60613--kev-demo.modal.run/

Click “Request a Key”

Devin Max Plan Claim

1) check in with yohann on luma
2) get $200 code from trydevin.ai
3) Fusion



```
> kev("Is this message spam?")
→ yes 0.97
```

40 ms to answer

0 words written

A four-week-old category

Sep 15

TypeSafe launches Jev, the first commercial "decision model."

Oct 1

AWS ships Strands Decider 2B, and Cloudflare ships Clef.

Oct 6

OpenAI opens a Decisions API in beta.

Oct 9

TypeSafe reportedly raises about $870M at a $7.5B valuation.

Kev

The open-source version, from Jared Palmer, Cognition's VP of Engineering.

github.com/jaredpalmer/kev

Generating vs deciding

Chat model

"Write me a paragraph"

Writes text, and your code has to parse it.

Decision model

"Pick A, B or C, with confidence"

Takes a document and typed questions, and returns probabilities.

Three question types

Kev reads a ticket, email, contract or record, then answers one of three ways.

Yes / no

```
"Is this message spam?"
```

Returns yes or no, with a confidence.

Pick one

```
"Which department?"
```

Returns one of your options, with a confidence.

Score

```
"Needs a human, 1-5?"
```

Returns a point on your scale, with a confidence.

Thresholds and fallbacks

Every answer gets a probability, so you decide what to automate.



What a decision model doesn't do

Explain itself

Retrieve facts

Come up with options

It only scores the options you give it.

How Kev works

Reads the document once

Every question branches off that cache, so five questions cost about the same as one.

A pointer head scores your options

Nothing gets generated.

Calibrated confidence

"0.9" means roughly 90% likely to be right.

Model sizes

0.8B	Any Mac
4B	32 GB Mac
9B	32 GB+ · tonight
27B	Big GPUs · tonight

Three examples to build

Jared's example

Support ticket router

Which department?	pick one
Escalate?	yes / no
Tone	score

Community

Community moderator

Spam?	yes / no
Needs a human?	1–5

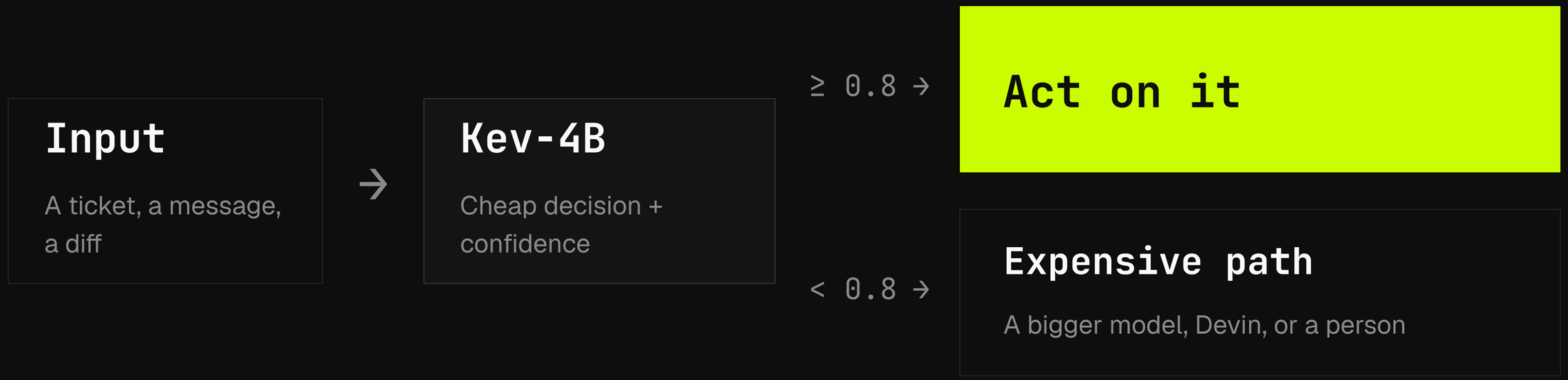
Code

PR triage

Does this diff need human review?	yes / no
Below threshold	→ Devin

The cheap-first pattern

A cheap decision first, an expensive one only when confidence is low.



06 · spin it up

Try Kev 1.0

yohann-60613--kev-demo.modal.run

01 Open your invite link. It signs you in.

No key? Press **Request a key** on the demo site.

02 Your mission is already loaded. Press **Run**.

The first run can take 3 to 4 minutes while the GPU wakes up.

03 Read the answers. Every option gets a probability.

04 Change a word or a question and run again.

Turn on **Live** to see the answers update as you type.

06 · spin it up

Try Kev 1.0, continued

05 Pick **Kev-9B**, **Kev-27B** or **Both** to compare the models.

06 Open **Run on your data**. Press **Load sample rows**, then **Run batch**.

Or paste your own list or CSV.

07 Move the **confidence slider** to see how much Kev handles on its own and how much goes to a person.

08 Press **Publish** to share it in the gallery.

Open someone else's and remix it.

Call the demo from code

Optional

```
$ curl -sL https://yohann-60613--kev-demo.modal.run/v1/systemone \
    -H "authorization: Bearer $KEV_DEMO_KEY" \
    -H "content-type: application/json" \
    -d '{"model": "kev-9b",
         "state": "I was charged twice",
         "questions": {"team": {"type": "choice",
           "criteria": {"billing": "Payments", "shipping": "Delivery"}}}}'
```

Set `KEV_DEMO_KEY` to your `kevdemo_` key. Use `"kev-9b"` or `"kev-27b"` as the model.

The playground's **`Call it from code`** panel gives you the exact curl and Python for what you built.

06 · spin it up

Run Kev on your own machine

Python 3.12 or 3.13 and uv. Works on a Mac (Apple Silicon) or an NVIDIA GPU.

```
$ git clone https://github.com/jaredpalmer/kev.git && cd kev
$ uv sync --extra serve
$ uv run --extra serve python -m kev.serve --run jaredpalmer/kev-4b --port 8009
```

The first run downloads the model.

Send requests to `http://localhost:8009/v1/systemone`. No key needed.

Same request format as the demo, with `"model": "kev-latest"`.

Change the name after `--run` to switch sizes.

On Apple Silicon, Kev-4B answers five questions on a short text in about 0.7s, or 0.14s cached.

No install: `huggingface.co/spaces/jaredpalmer/kev` (Kev-4B and Kev-0.8B).

Demo keys and the playground only work with the hosted demo. Your own server has no demo quota.

06 · spin it up

Your own Kev in the cloud

Sign up free at modal.com. The Starter plan includes $30 of compute credit every month.

```
$ pip install modal && modal setup
$ curl -LO \
    https://raw.githubusercontent.com/jaredpalmer/kev/main/skills/kev-deploy/scripts/kev_serve.py
$ KEV_API_KEY=$(openssl rand -hex 24) modal deploy kev_serve.py
```

Your own HTTPS endpoint, Kev-4B by default. For Kev-9B, add `KEV_MODEL=jaredpalmer/kev-9b`.

Send `Authorization: Bearer <your KEV_API_KEY>` .

It costs nothing when idle. GPU time comes out of your monthly credit first.

06 · spin it up

Or let Devin deploy it

```
$ npx skills add
jaredpalmer/kev@kev-deploy
```

Then tell Devin to deploy it.



06 · spin it up

Jev, the hosted option

Commercial, from TypeSafe.

Model	Per million input tokens	Status
jev-latest	$0.042	Early access

Already using Jev? Kev uses the same API. Keep your code and change the base URL to your Kev server.

```
from typesafe_sdk import TypeSafeClient
client = TypeSafeClient(api_key="local",
                        base_url="http://127.0.0.1:8009",
                        model="kev-latest")
```

06 · spin it up

Run JEV locally

TypeSafe's hosted Jev is closed and can't run locally. The open copy is `autotrust/JEV-9B`.

```
$ hf download autotrust/JEV-9B --local-dir JEV-9B
$ vllm serve JEV-9B --served-model-name autotrust/JEV-9B \
    --enable-lora --max-lora-rank 32 \
    --lora-modules jev-decision=JEV-9B/adapter_vllm \
    --logprobs-mode processed_logprobs --max-model-len 4096
```

About an 18 GB download. Needs vLLM on an NVIDIA GPU.

Also writes normal text through an OpenAI-style `/v1/chat/completions`.

It does not use the `/v1/systemone` API. For decisions, copy the `decide()` helper from the "Quickstart with vLLM" section of its Hugging Face card.

06 · spin it up

Four gotchas

~35s	A scaled-to-zero endpoint takes about 35s to cold start. Keep one container warm.
Batch	Batch your questions per document.
Mock	Mock Kev in your tests.
429	Back off and retry.

07 · make it yours

Fine-tuning Kev

```
$ npx skills add jaredpalmer/kev@kev-finetune
```

~$1

of GPU time for a small Kev-4B run

80% → 90%

accuracy on consumer complaints, after one pass

Limits of the evidence

The benchmarks are the author's own.

There's no head-to-head against a prompted chat model yet.

Calibration can drift, so test on your own data.

The brief

01 Ship an app where a model makes one real decision.

02 At demo time, show the decision process, what happens if it’s wrong

---

Six selected demos at 1:30. each demo gets another $200 code.