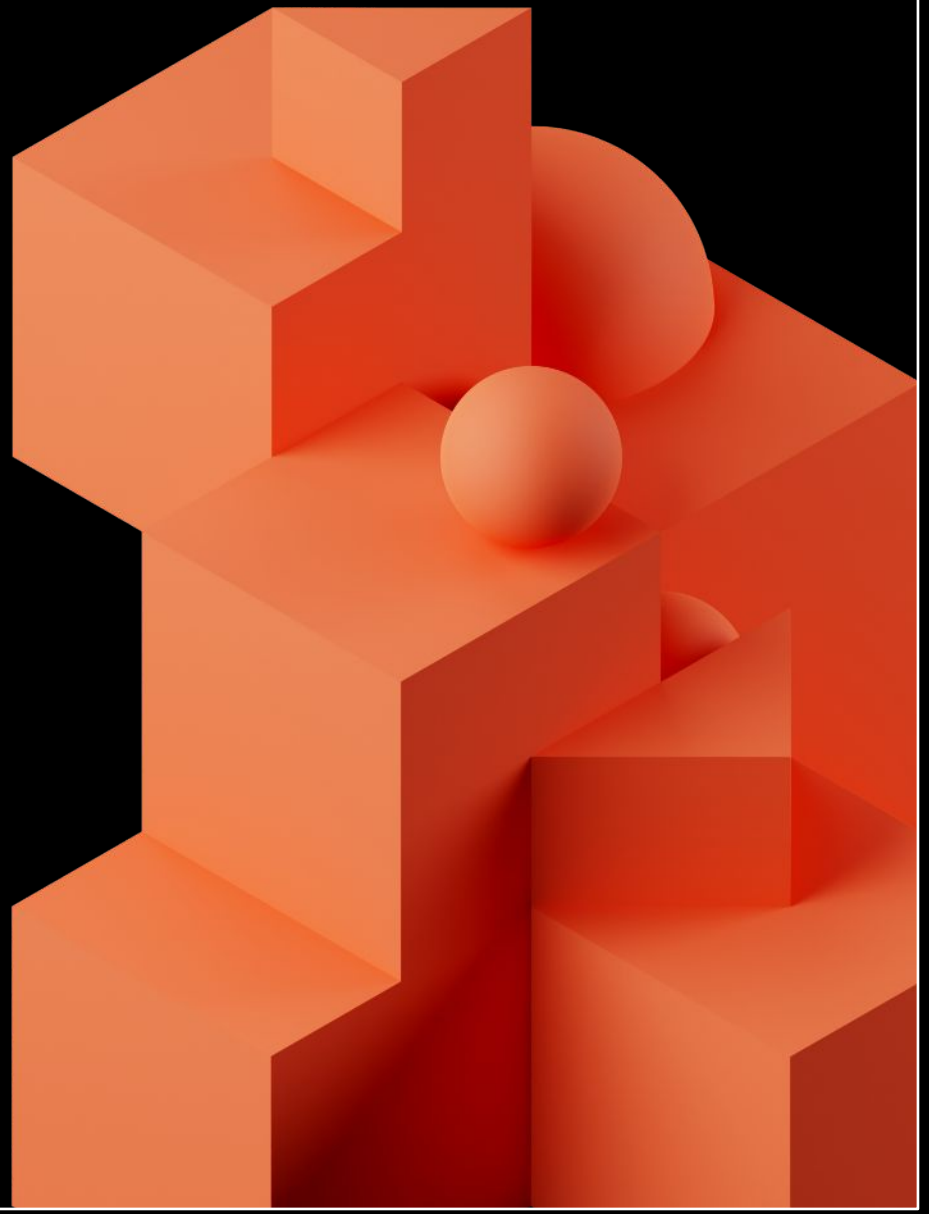
DATA+AI SUMMIT
BY databricks

Building LLMs on Your Data

... and on a budget



Sean Owen

Principal DS/ML Specialist @ Databricks

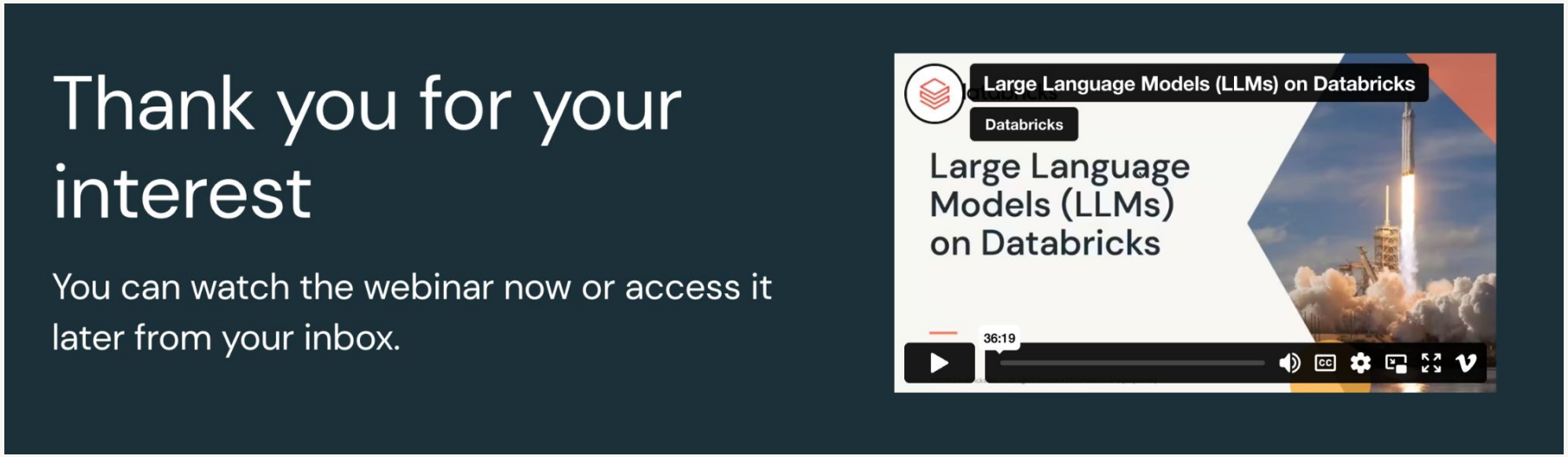
Product safe harbor statement

This information is provided to outline Databricks' general product direction and is **for informational purposes only.** Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all.

Remember Me?

From that Webinar?

- Continuation of Apr 2023 Webinar
  Build Your Own Large Language Model Like Dolly
  www.databricks.com/resources/webinar/build-your-own-large-language-model-dolly



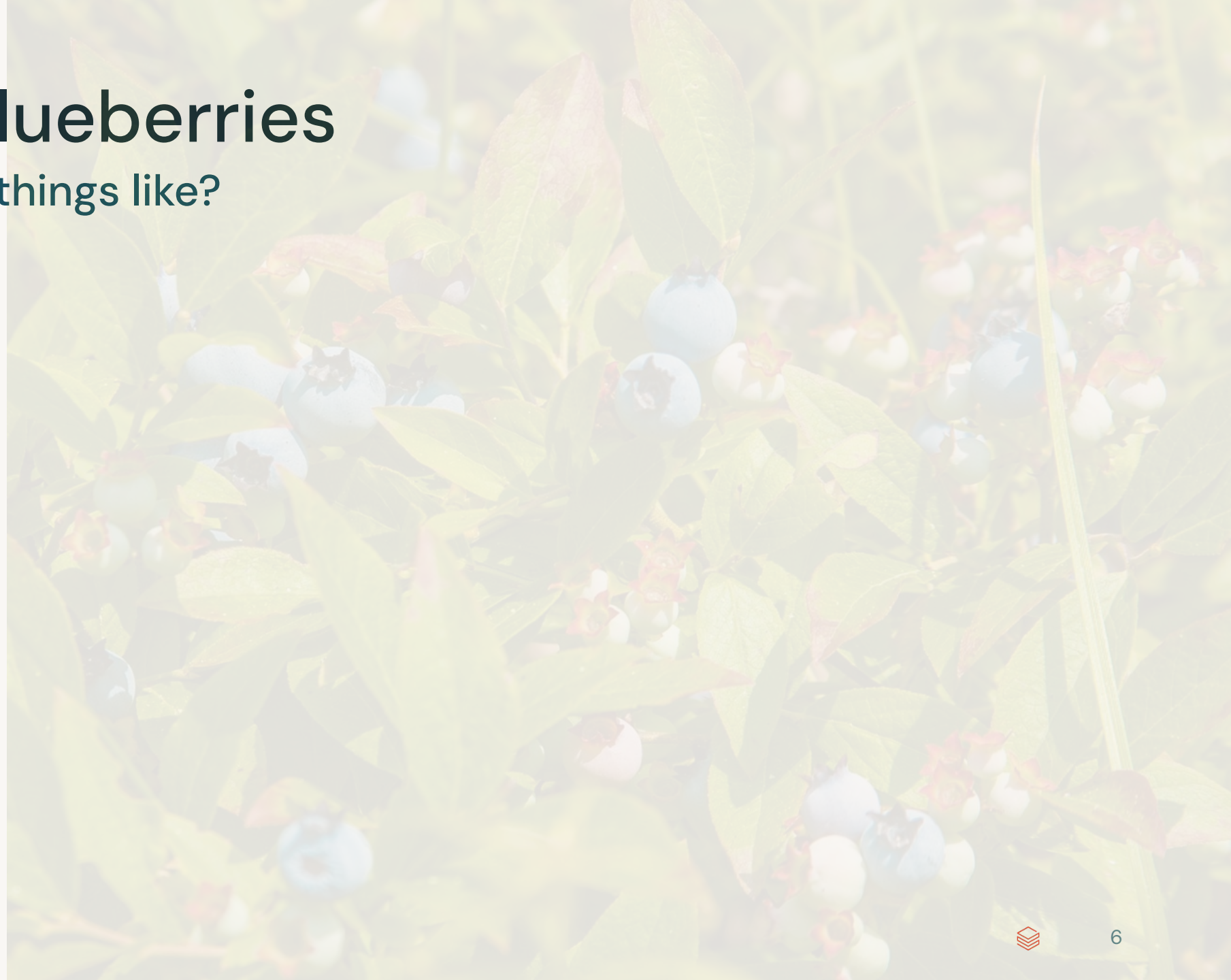
Agenda

- Use Case: Blueberries
- Don't Fine Tune!
  - feat. Chroma, langchain
  - … and Lessons Learned
- Fine Tune!
  - feat. deepspeed
  - … and Lessons Learned
- Next Steps and Recap

Use Case: Blueberries

Use Case: Blueberries

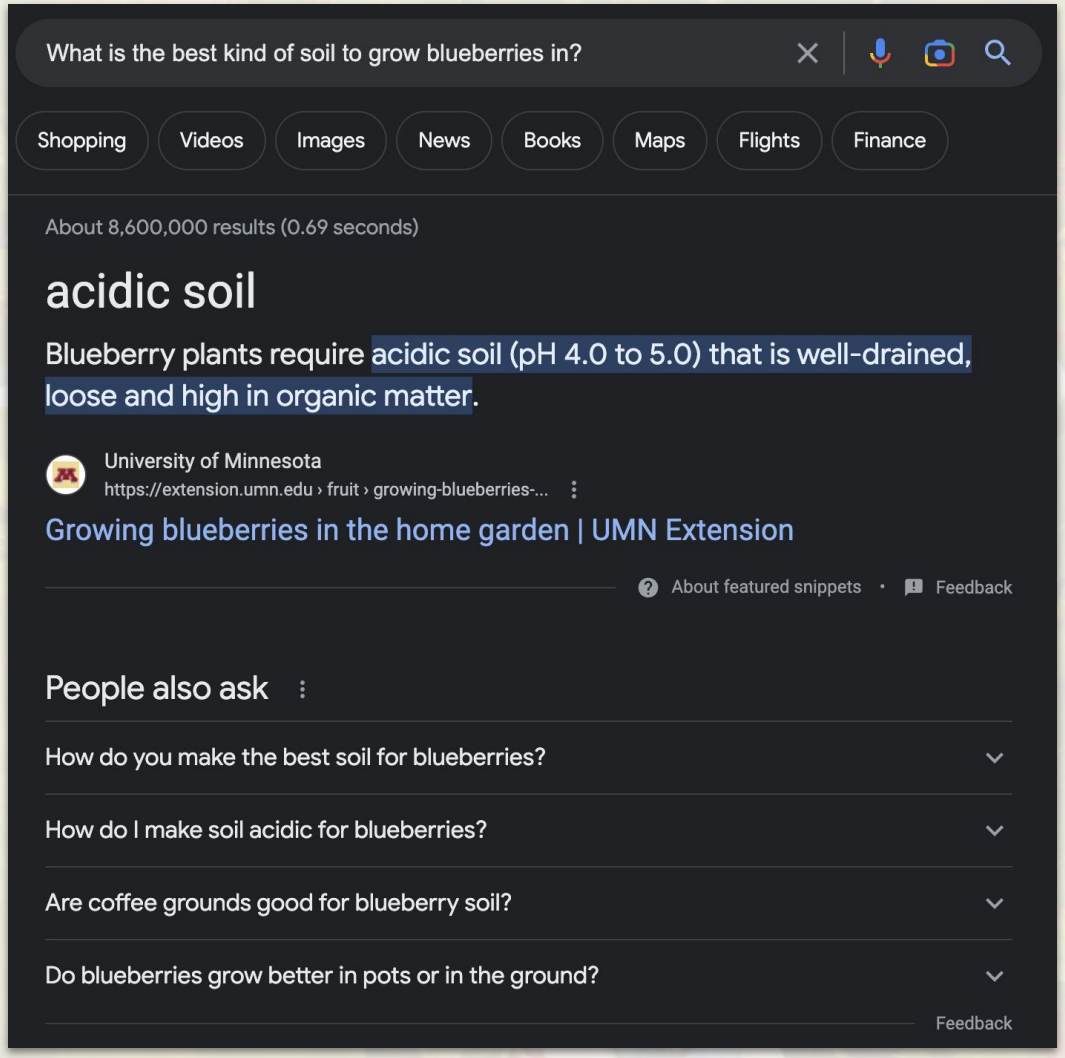
What soil do these things like?



Use Case: Blueberries

What soil do these things like?

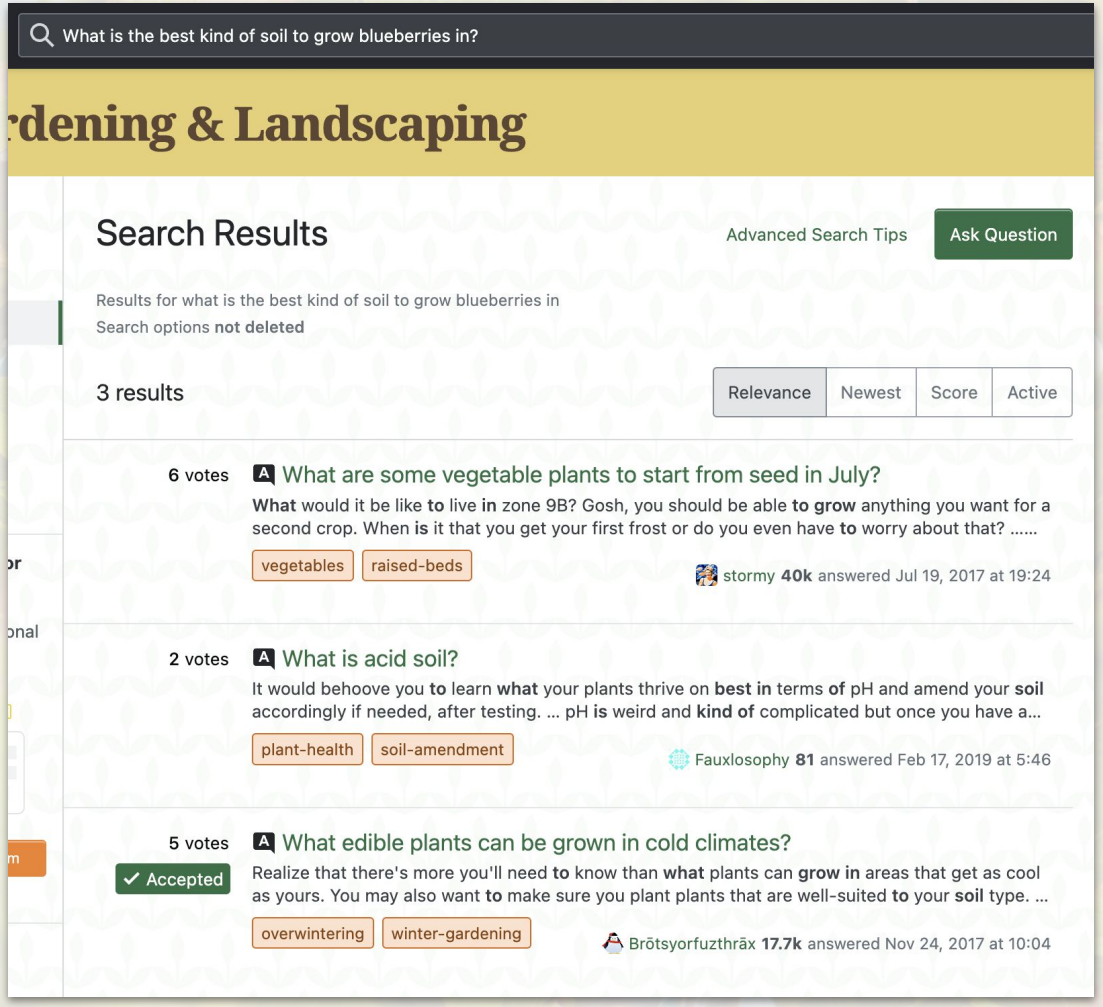
- Ask a search engine?
  - Useful snippet, but need to dig through results



Use Case: Blueberries

What soil do these things like?

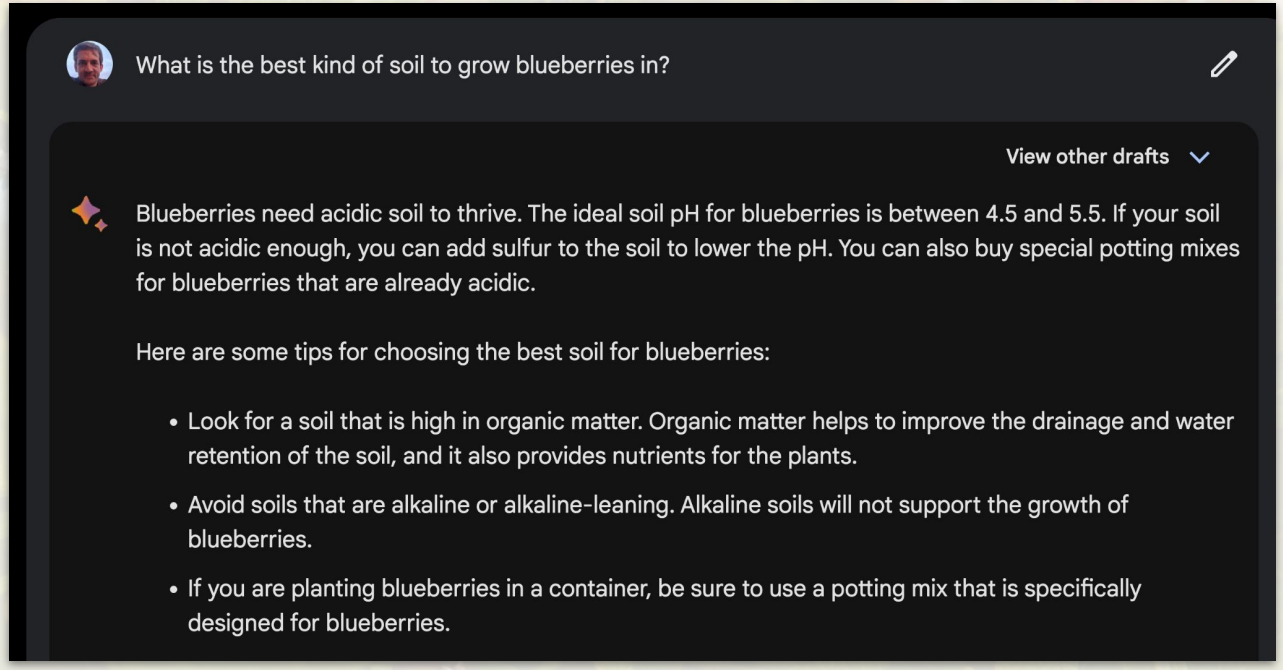
- Ask a search engine?
  - Useful snippet, but need to dig through results
- Ask a domain Q&A site?
  - Doesn't find relevant answer



Use Case: Blueberries

What soil do these things like?

- Ask a search engine?
  - Useful snippet, but need to dig through results
- Ask a domain Q&A site?
  - Doesn't find relevant answer
- Ask an AI?
  - Nice! Can I have my own?



Use Case: Blueberries

What soil do these things like?

- Ask a search engine?
  - Useful snippet, but need to dig through results
- Ask a domain Q&A site?
  - Doesn't find relevant answer
- Ask an AI?
  - Nice! Can I have my own?
- Ask an open LLM?

```
1  from transformers import pipeline
2
3  dolly = pipeline(model="databricks/dolly-v2-12b", torch_dtype="bfloat16", \
4    trust_remote_code=True, device_map="auto", model_kwargs={'load_in_8bit':
     True})
5
6  dolly("What is the best kind of soil to grow blueberries in?")[0]
   ['generated_text']
```

```
Out[4]: 'Blueberries grow best in well-drained soil with a pH of approximately 6.0 - 6.5.
Sandy soil is ideal, but you can also grow blueberries in heavy clay soil if it is amende
d with considerable amounts of organic material such as manure.'
```

Problem: Answering from Private Data

How can I answer questions based on my own text corpus?

- GardeningCo has deep proprietary expertise in gardening
  - Thousands of documents with expert knowledge, questions and answers
  - *Here: use Gardening Stack Exchange data set*
  - *Imagine your knowledge base, PDFs, contracts, documentation here*
- Want:
  - AI-like info retrieval and synthesis, over superior private data
  - Flexibility to use and customize AI
- Do not want:
  - Sending sensitive information to 3rd party
  - Dependence on 3rd party
  - … to spend a lot of money!

Two Paths

Start with an instruction-following text-gen model, and...

Fine Tune It

- Continue training with question-answer prompts
- ... or just the answers

- No extra runtime moving parts
- Efficient at runtime as model has already learned to answer

Don't Fine Tune It

- Retrieve text related to question at runtime
- Feed relevant text with question

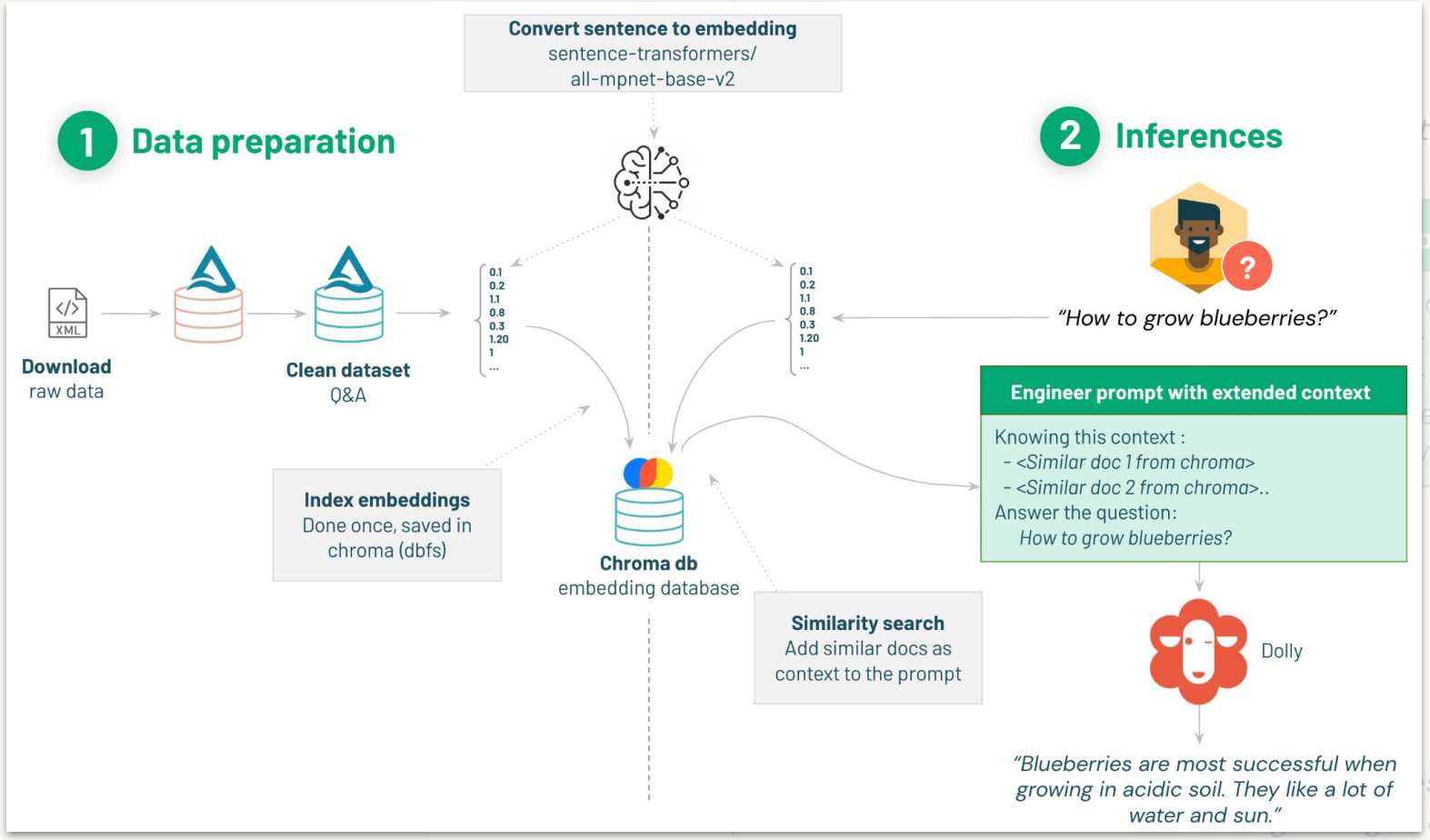
- Easy: no fine-tuning time/cost
- No re-fine-tuning necessary to add new information

Don't Fine Tune

feat. Chroma, langchain

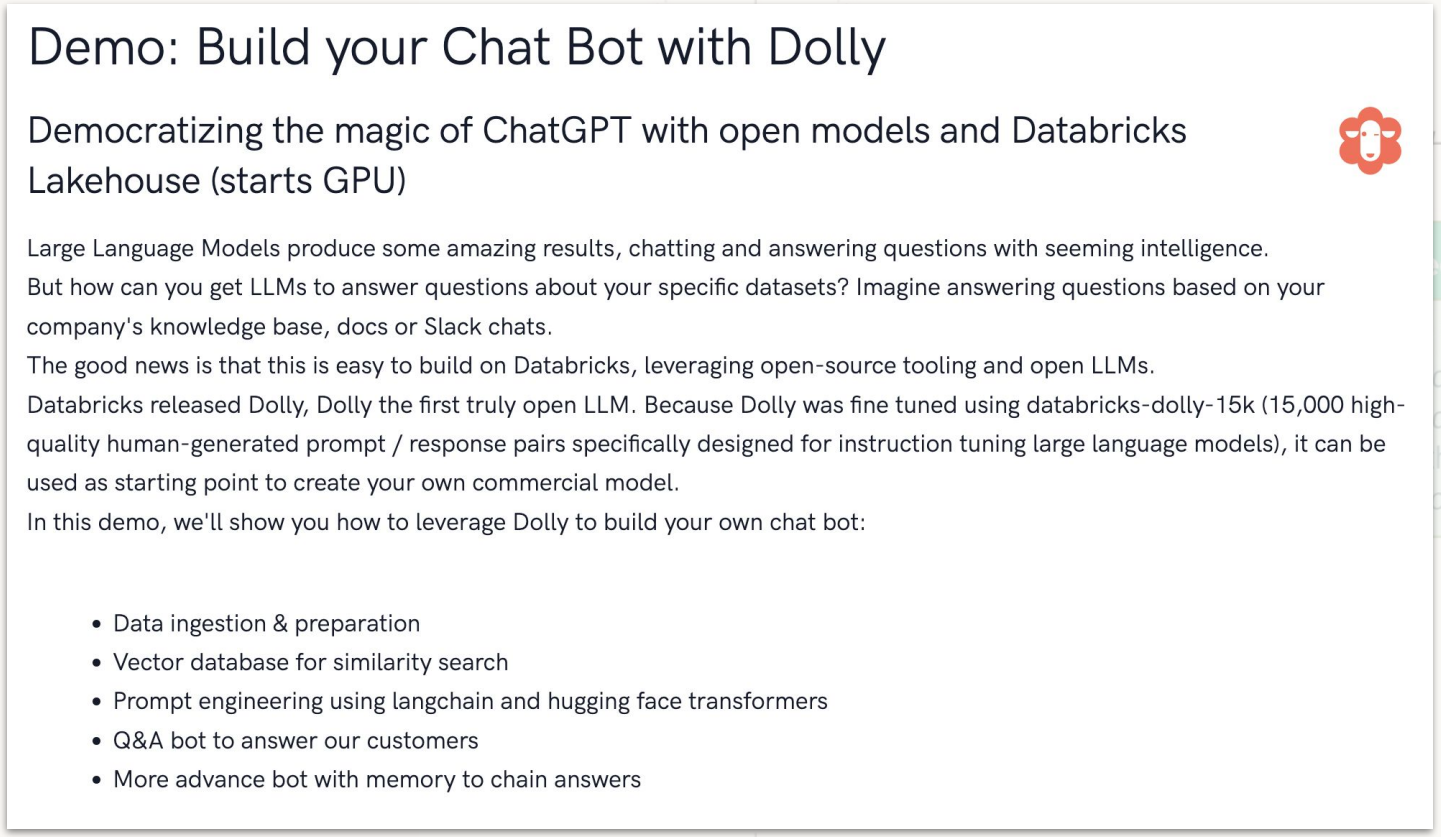
Don't Fine Tune

Index, then Search and LLM



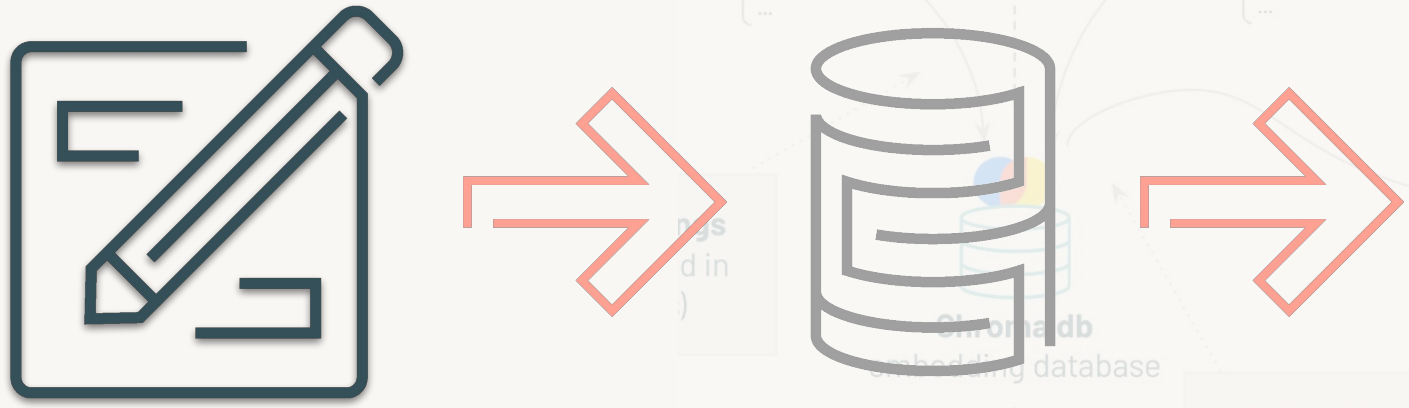
Ad: dbdemos.ai

Installable Databricks Demos



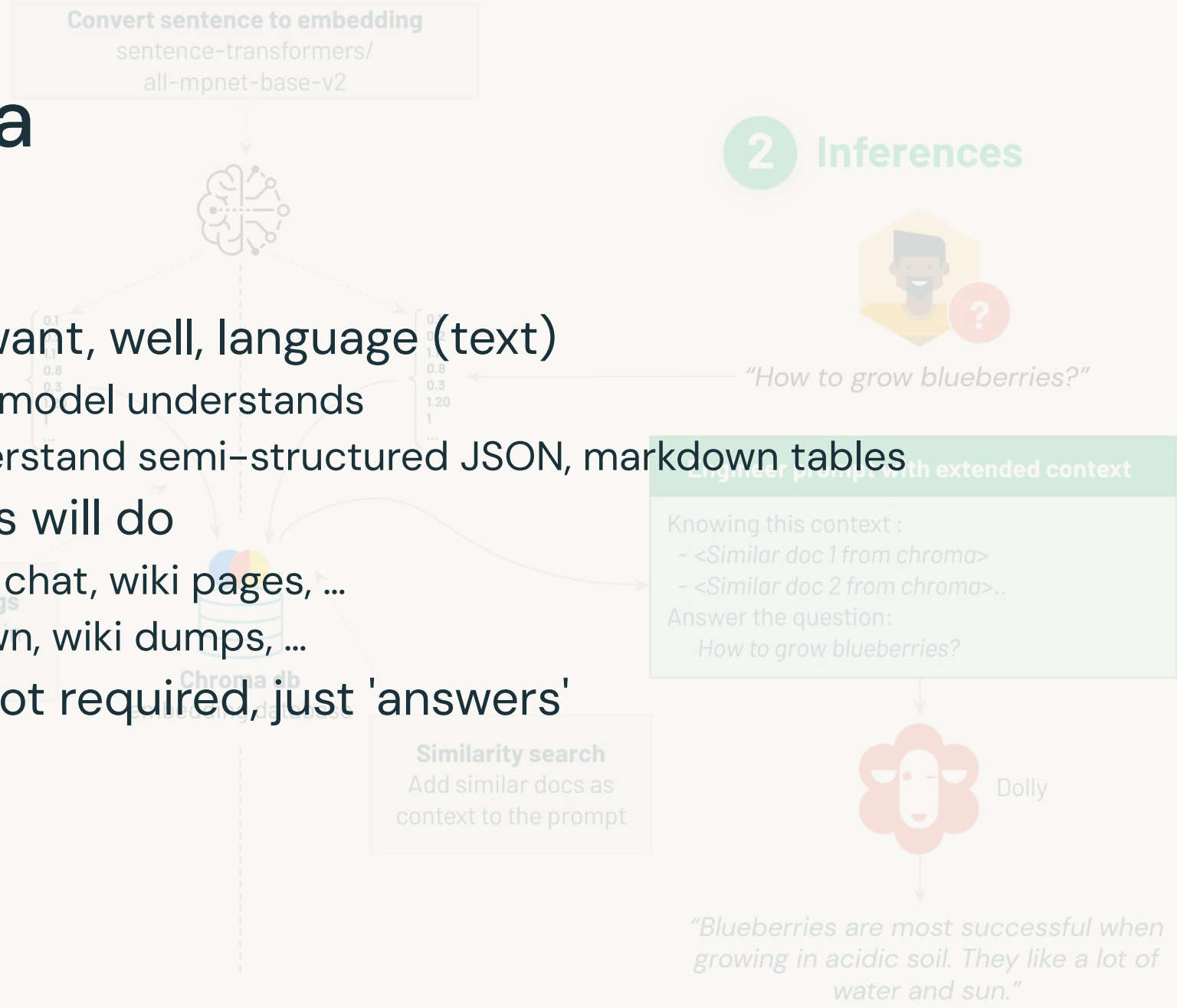
www.dbdemos.ai/demo.html?demoName=llm-dolly-chatbot

Prepare Text Data

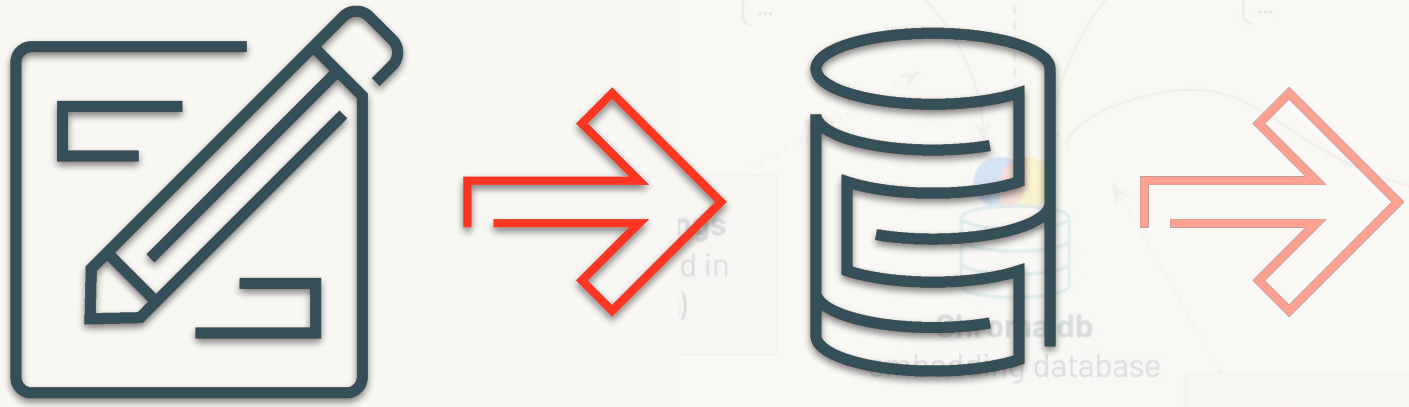


Prepare Text Data

- Large Language Models want, well, language (text)
  - Needs to be language the model understands
  - Models might kind of understand semi-structured JSON, markdown tables
- Any source of text chunks will do
  - Knowledge base, manuals, chat, wiki pages, ...
  - HTML, PDF, Word, markdown, wiki dumps, ...
- *Questions* and answers not required, just 'answers'



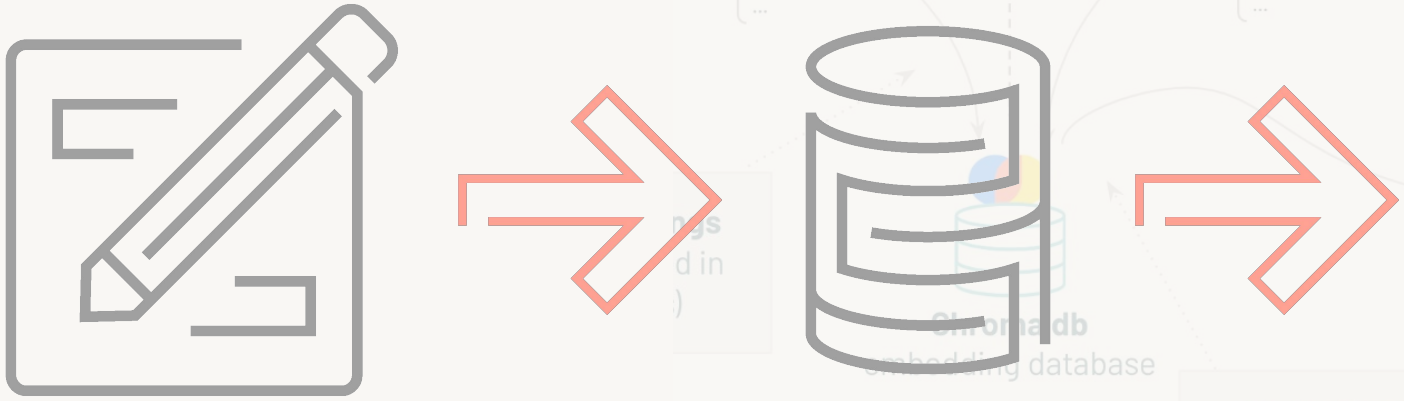
Index Text Data



Index Text Data

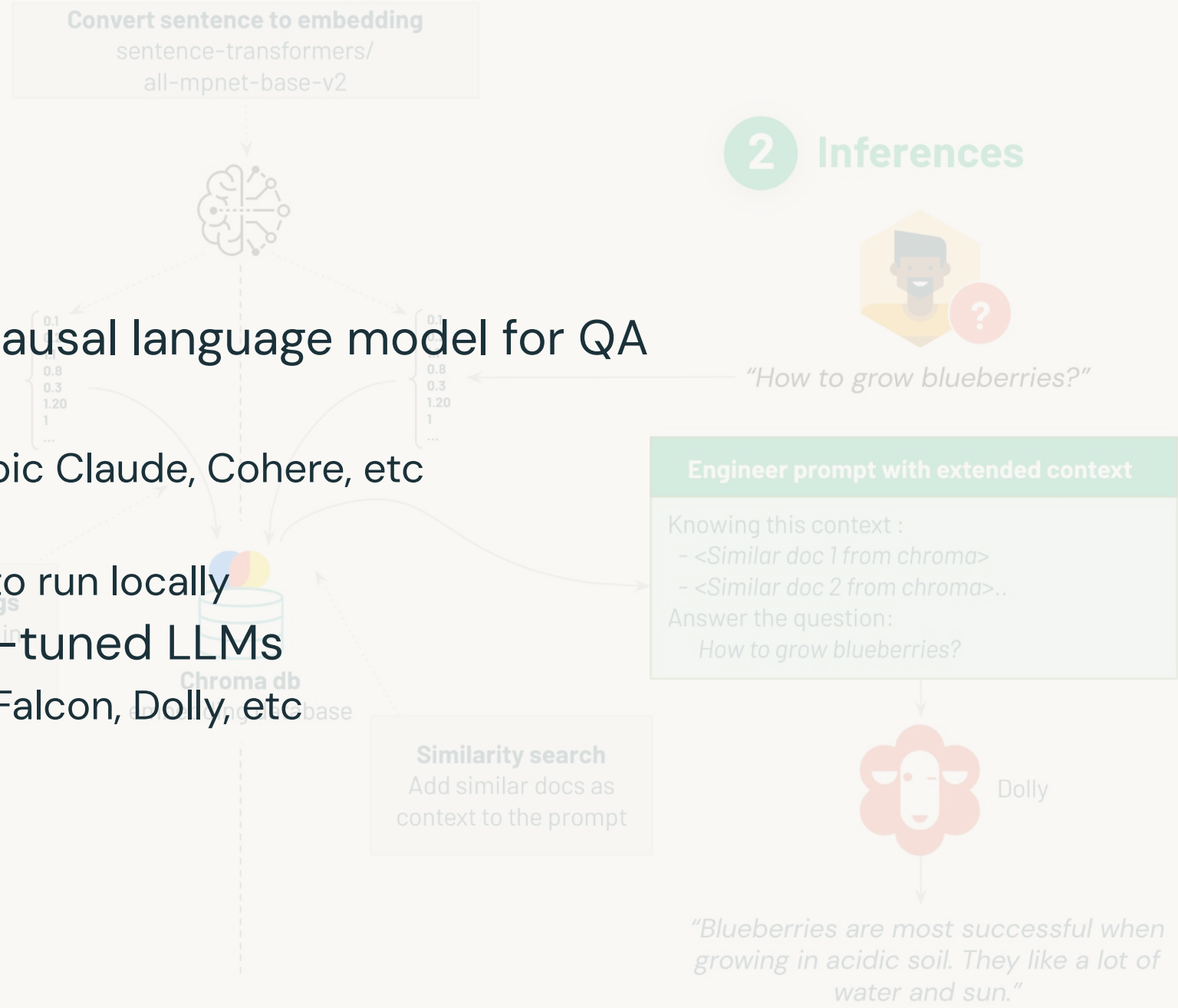
- From question, find relevant text
  - Good use case for vector DB; some use cases don't need them
  - Could be a search engine too
- Vector DB requires choice of "good" embedding
  - Off-the-shelf from `sentence-transformers`
- Which vector DB?
  - Chroma: OSS, embedded, simple; the "sqlite" of vector DBs
  - Pinecone, Milvus, Weaviate, etc as standalone servers
  - *Coming Now-ish: Databricks Vector Index*

Load an LLM



Load an LLM

- Need a text-generation causal language model for QA
- Proprietary LLM Services
  - Ex: OpenAI GPT-4, Anthropic Claude, Cohere, etc
  - State-of-the-art
  - Very easy to use, nothing to run locally
- Open-source instruction-tuned LLMs
  - Ex: FLAN-T5, Mosaic MPT, Falcon, Dolly, etc
  - Varied and customizable
  - Run and manage privately



Load an LLM on GPUs

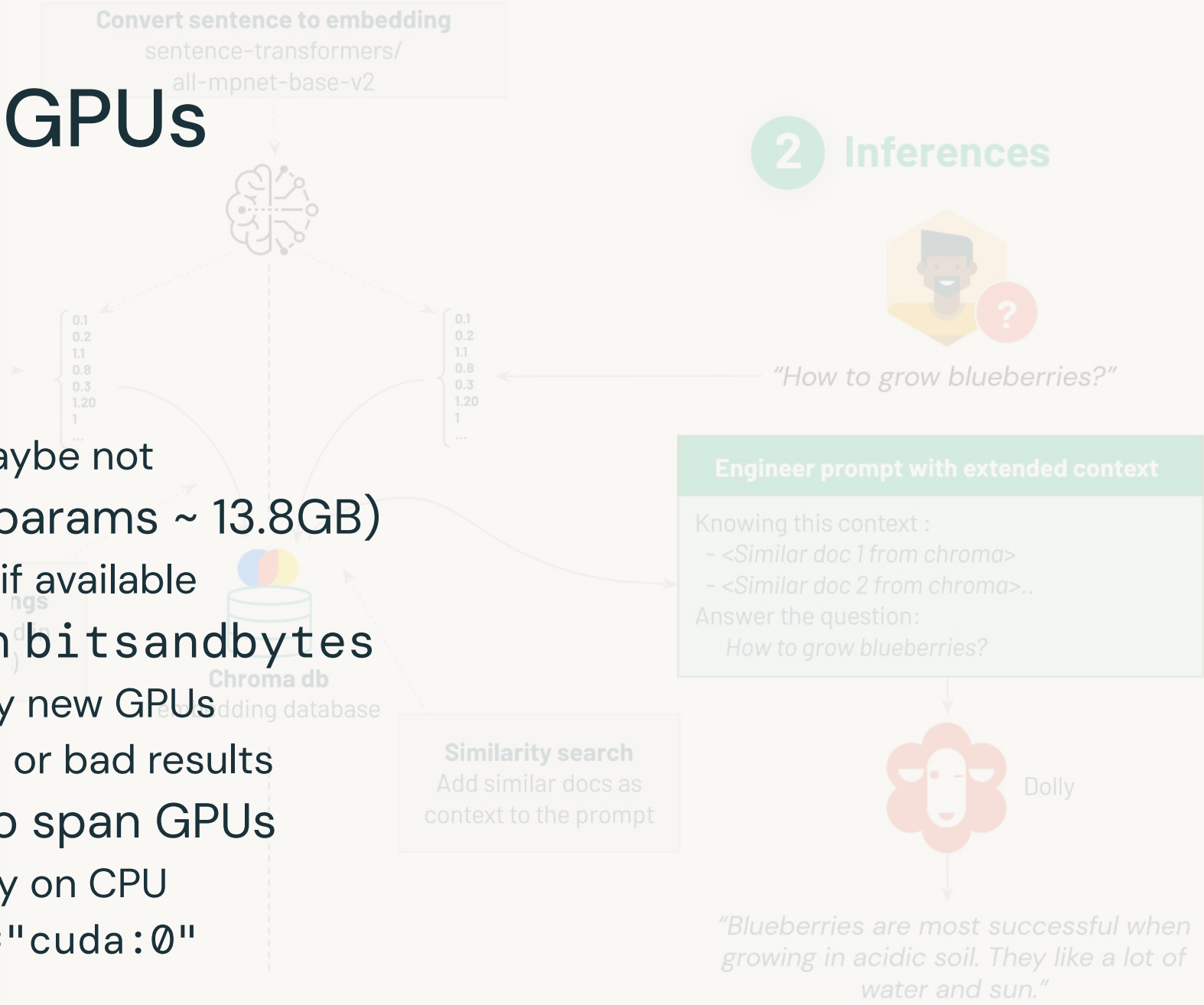
- Need GPUs for speed
- Bottleneck? GPU **memory**
- Common NVIDIA GPUs:

	Released	Memory	Half-Precision TFLOPS	Cost / Scarcity	Notes
V100	2017	16GB (32GB)	112	●●●	Training
T4	2018	16GB	65	●	Inference
A100	2020	40GB (80GB)	312	●●●●●	Training, bfloat16
A10	2021	24GB	125	●●	Inference, bfloat16

en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units#Tesla

Load an LLM on GPUs

- How big?
  - >1B params? yes
  - >10B params? maybe, maybe not
- Work in 16-bit (ex: 6.9B params ~ 13.8GB)
  - Use bfloat16 over float16 if available
- Use 8-bit if needed with `bitsandbytes`
  - Not faster, except on very new GPUs
  - Careful: can cause errors or bad results
- `device_map="auto"` to span GPUs
  - Careful: easy to run partly on CPU
  - Force GPU with `device="cuda:0"`



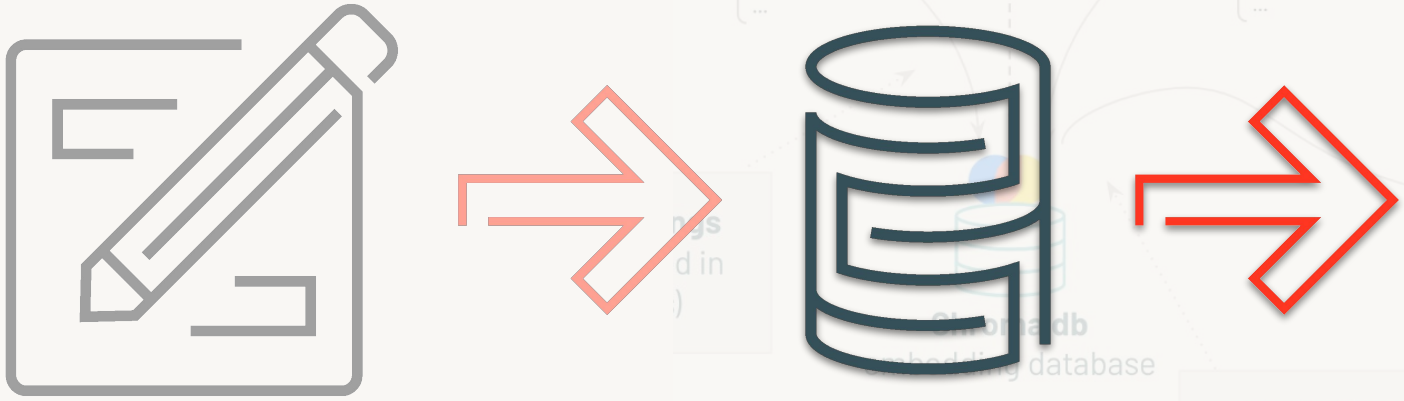
Load an LLM on GPUs

- Example: Dolly 7B

Hardware	Param Size	Time	Response
48-core CPU	16-bit	19 sec	Soil that is rich in organic matter, full of minerals, and has a pH of 6.0 to 6.8 is best for growing blueberries. Blueberries like sandy loam soil that is full of organic matter, has a pH of 6.2 to 6.6, and is low in clay.
T4 (16GB)	8-bit	7 sec	Soil that is rich in organic matter, has a pH of 6.0 to 6.8, and is well-drained is best for blueberry growing. Sandy soil is not good for blueberries, as they require a lot of soil moisture.
A10 (24GB)	16-bit (bfloat16)	3 sec	Soil that is neutral to slightly acidic with good drainage is ideal for blueberry plants. In general, soil that is rich in organic matter, has a pH of 6.5-7.0, and is sandy or rocky will provide the best growing conditions.

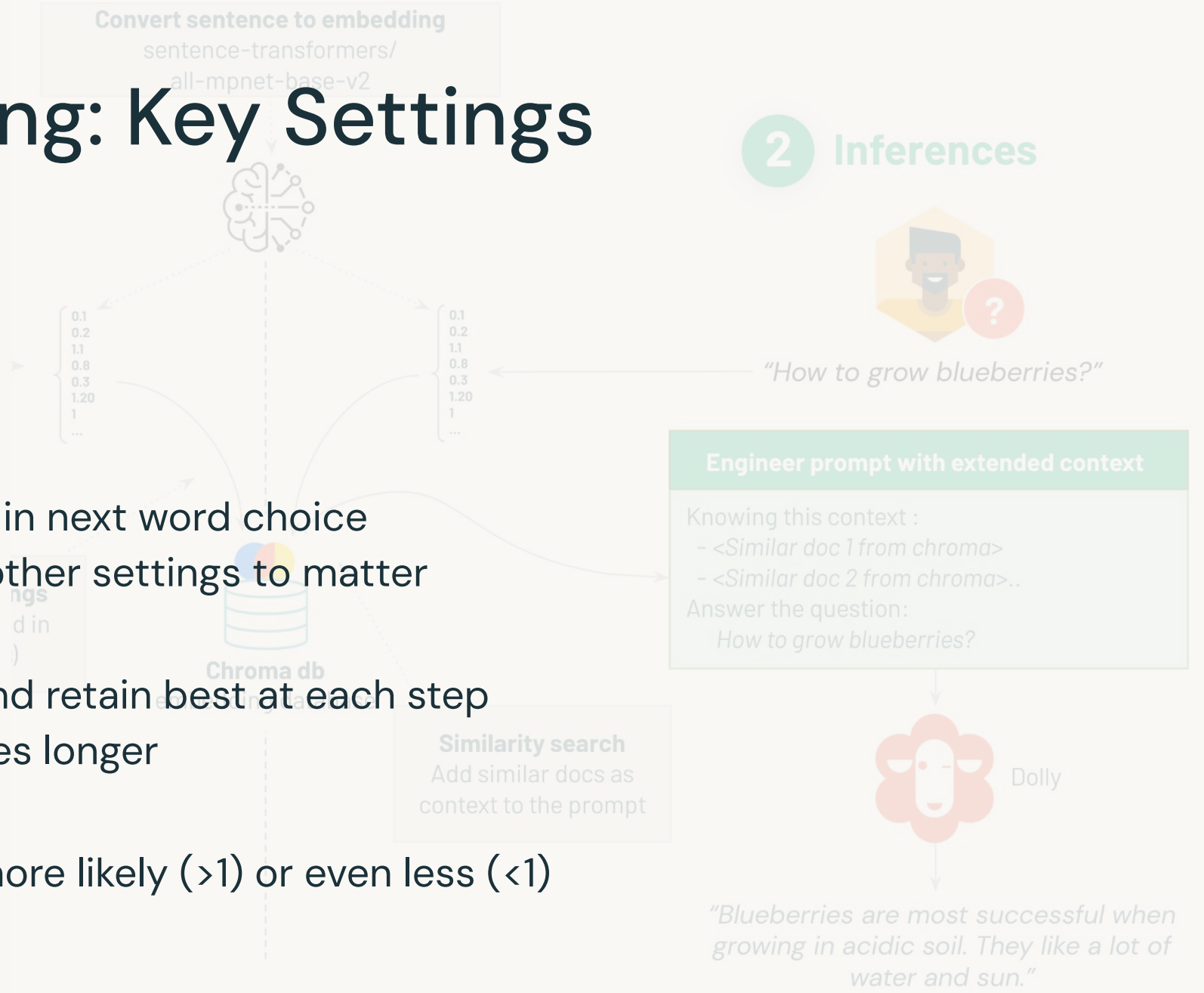
- Default recommendation: use A10, <10B params, bfloat16

Search and LLM



Generation Tuning: Key Settings

- `max_new_tokens`
  - limit generation length
- `do_sample`
  - Allow some randomness in next word choice
  - `True` required for most other settings to matter
- `num_beams`
  - Generate N responses and retain best at each step
  - Generation takes ~N times longer
- `temperature`
  - make less-likely words more likely (>1) or even less (<1)



Generation Tuning: Other Settings

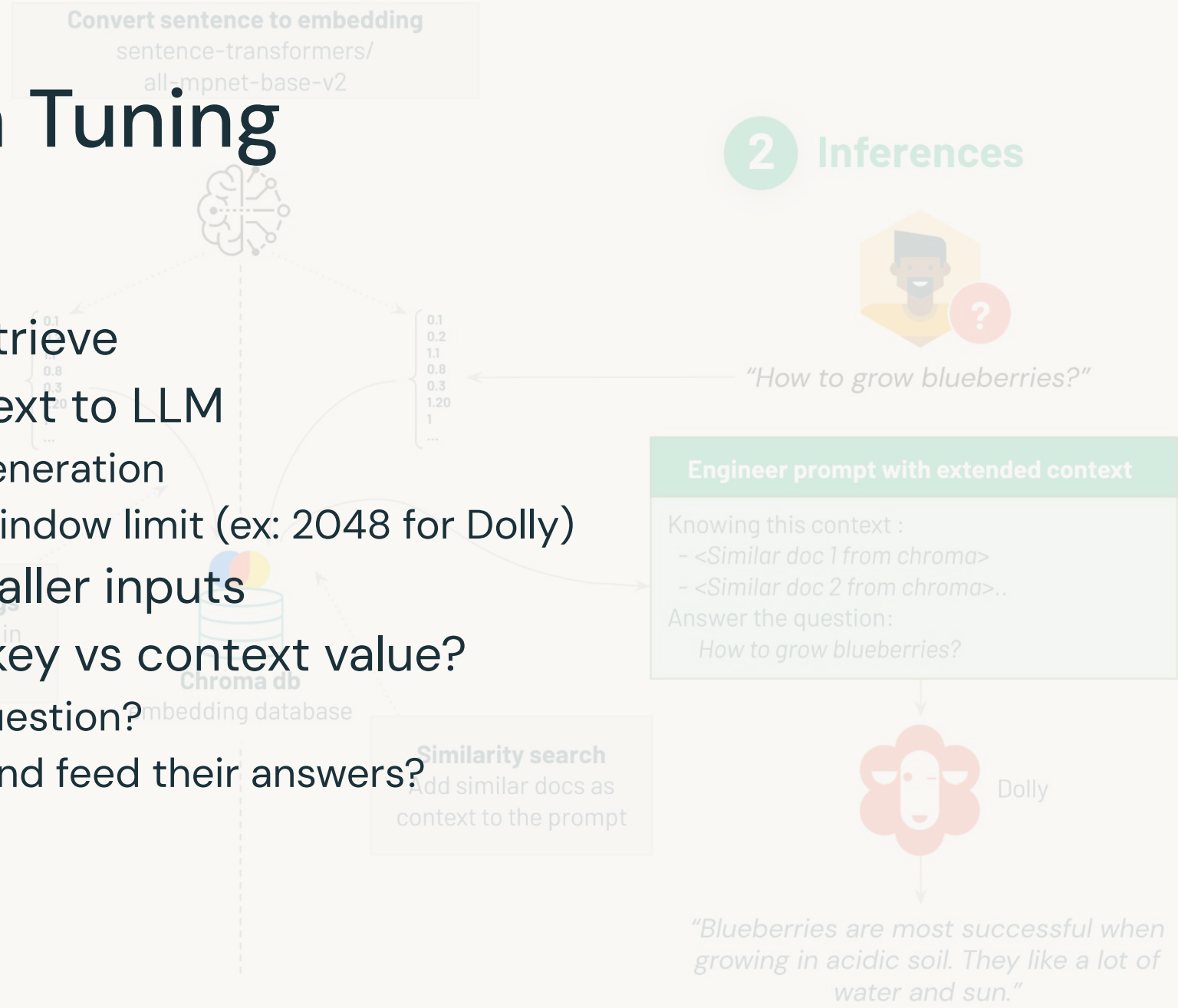
- `num_return_sequences`
  - return N responses
- Nucleus sampling
  - `top_p`: excludes long-tail next tokens by cumulative probability
  - `top_k`: same, by count
- `repetition_penalty`
  - penalizes duplicated passages in response
- Recommended reading:
  - How to generate text: using different decoding methods for language generation with Transformers huggingface.co/blog/how-to-generate

Generation Tuning

Setting	Change	Response
(default)		If you're growing blueberries in your garden, consider using a raised bed. Unlike a flat bed, the soil in a raised bed will naturally retain moisture, which reduces the need to water frequently. Additionally, the sloping sides help to prevent weeds from growing.
`num_beams`	1 → 4	Soil that is well-drained, high in organic matter, and full of minerals is best for growing blueberries.
`temperature`	1 → 1.7	If you are growing blueberries in a raised bed, then soil that is well prepared with organic matter will improve the soil's productivity. Many folks combine rotting leaves with straw for added organic matter and moisture. Water well throughout the year, and do not overwater! Integrated Pest Management is critical …
`temperature`	1 → 0.3	Soil with a pH of 6.0 to 6.8 is best for blueberry growth.
`top_p`	1 → 0.6	Blueberries like sandy soil that is well drained.
`top_k`	50 → 10	The best soil for growing blueberries is a sandy loam with a pH of 5.5 to 6.5.

Database Search Tuning

- k = number of docs to retrieve
- Higher means more context to LLM
  - More to process, slower generation
  - Runs up against context window limit (ex: 2048 for Dolly)
- Chunk more finely for smaller inputs
- What is the right search key vs context value?
  - Feed answers similar to question?
  - Or find similar questions and feed their answers?



Database Search Tuning

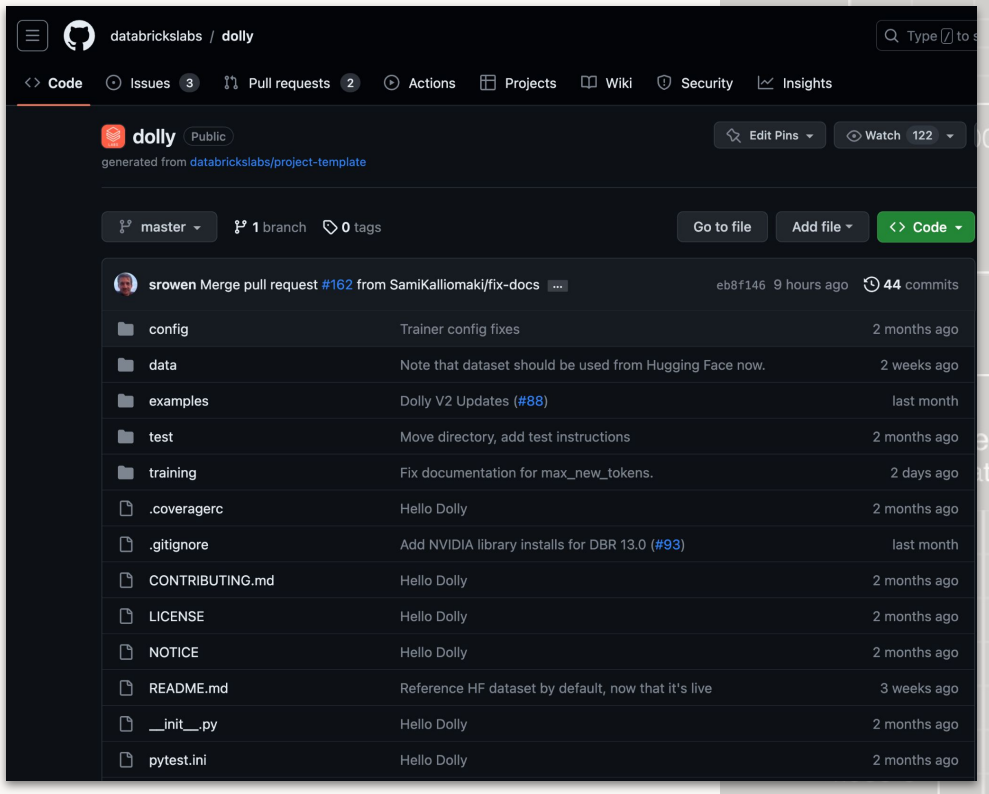
Q + A ?	k	Response
Answer	2	The best soil mix for blueberries is a sandy, acidic mix high in organic matter. A good mix could be made by mixing these materials by weight: 5/8 topsoil 2/8 peat moss 1/16 vermiculite 1/16 coarse sand Plus some slow release plant food. To help with the acidity, mix 1/4 pound of aluminum sulfate into the top 6" of soil. There are several mulches that fit the needs of the blueberry bushes...
Question + Answer	2	Blueberries like soil that is acidic. The best soil mix for blueberries is a porous, acidic mix high in organic matter. A good mix could be made by mixing these materials by weight: 5/8 topsoil, 2/8 peat moss, 1/16 vermiculite, 1/16 coarse sand.
Answer	4	The best soil mix for growing blueberries is a porous, acidic mix high in organic matter. A good mix could be made by mixing these materials by weight: 5/8 topsoil 2/8 peat moss 1/16 vermiculite 1/16 coarse sand Plus some slow release plant food. To help with the acidity, mix 1/4 pound of aluminum sulfate into the top 6" of soil. There are several mulches that fit the needs of the blueberry bushes...
Question + Answer	4	The best soil for growing blueberries is one high in organic matter, low in pH (acidity) and high in phosphorus. A soil mix for blueberry plants could be made by mixing these materials by weight: 5/8 topsoil 2/8 peat moss 1/16 vermiculite 1/16 coarse sand Plus some slow release plant food. Coffee grounds are only very slightly acidic, so they will not be useful in maintaining ph...

Fine Tune

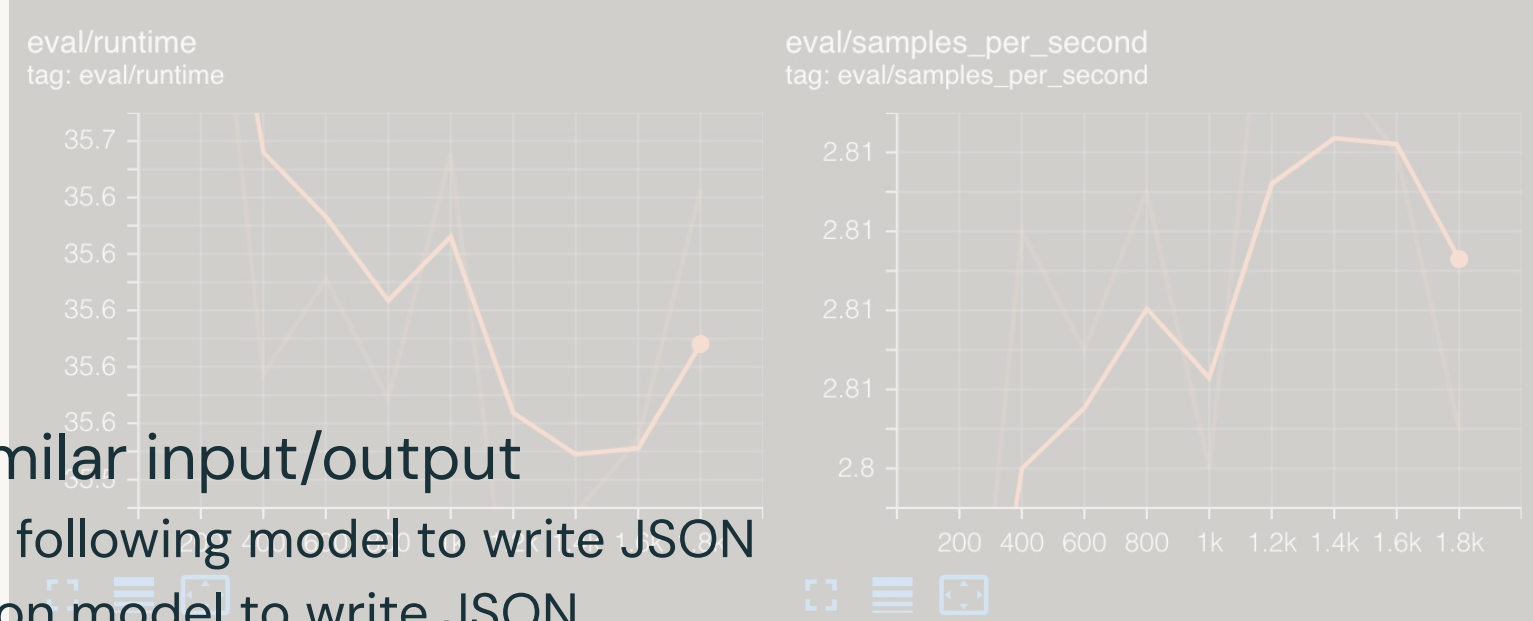
feat. deepspeed

Fine Tune

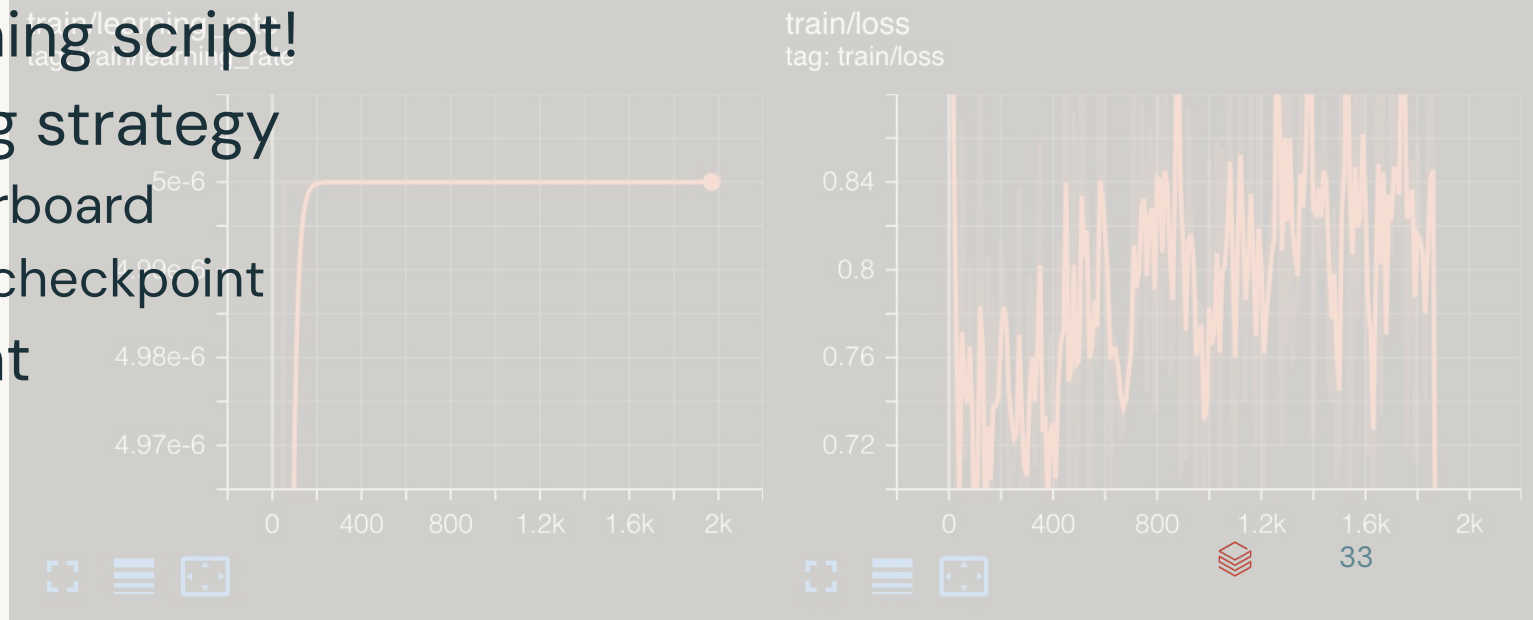
Fine-Tune, then LLM



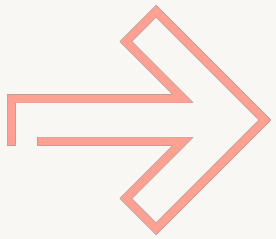
Preliminaries

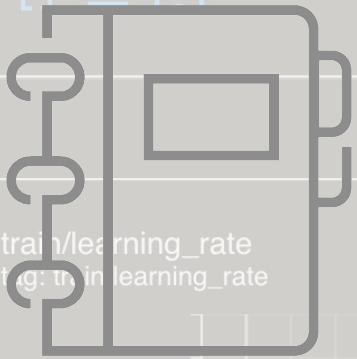


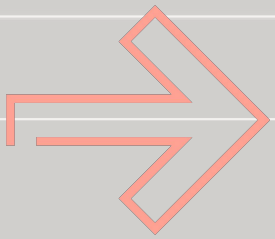
- Pick a model with most similar input/output
  - Good: fine-tune instruction following model to write JSON
  - Bad: fine-tune summarization model to write JSON
  - Does your task extend/specialize, or conflict with, base behavior?
- Understand format of input
- Find an off-the-shelf training script!
- Decide training monitoring strategy
  - Ex: Monitor loss with Tensorboard
  - Periodically generate from checkpoint
- Estimate time/cost upfront

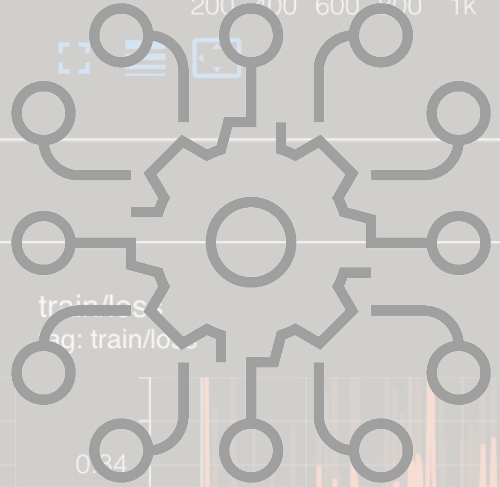


Prepare Text Data









Prepare Text Data

- Easiest if you have `Instruction,(Context,)Response`
- Could be: `Question,(Body,)Answer`
- Any strings will *work*
  - … with prompt modification
  - Just tuning any text does *not* induce instruction following
- Art more than Science

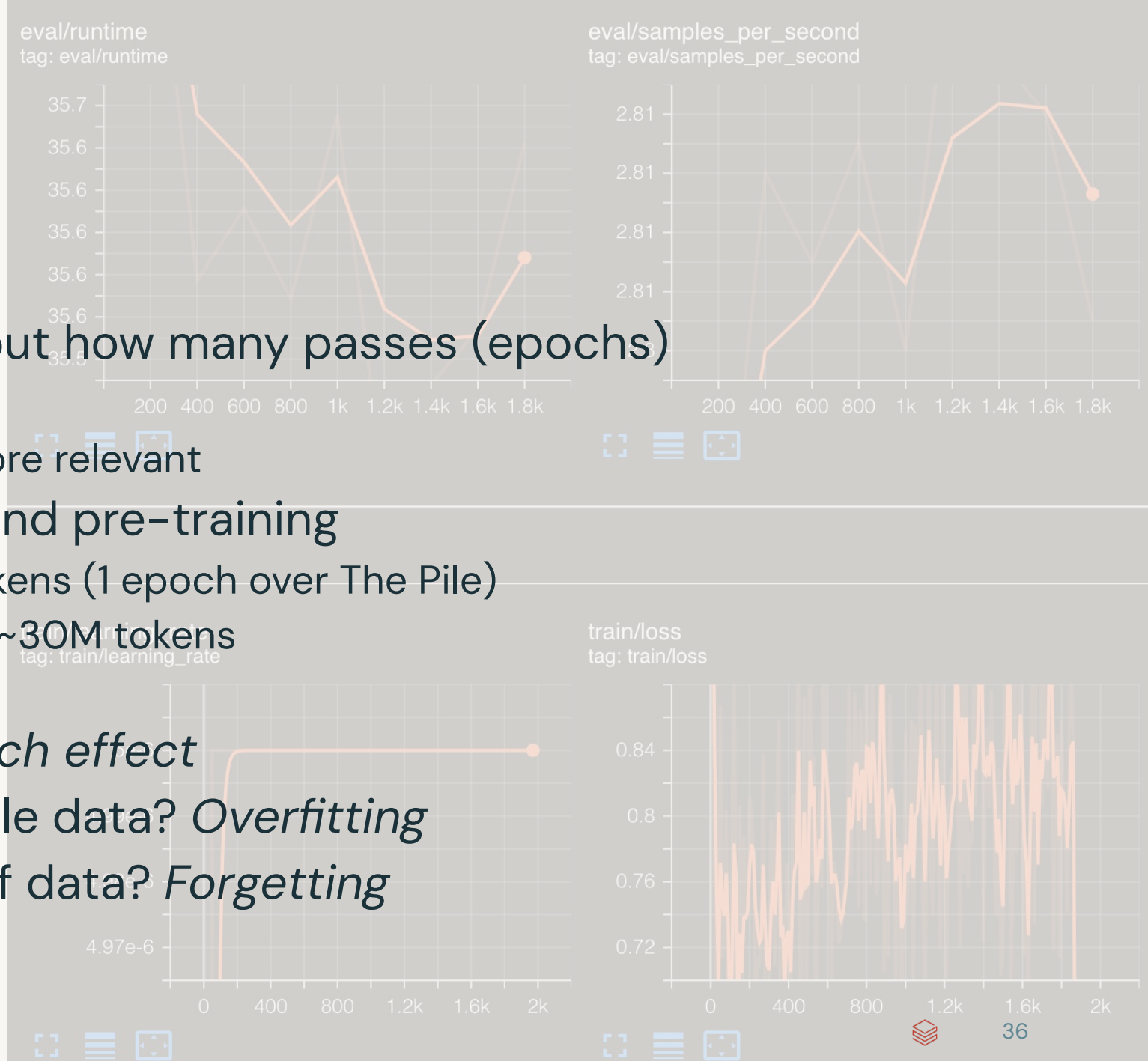

```
gardening_df = spark.read.format("xml").option("rowTag", "row").\
  load(f"{gardening_path}/Posts.xml").\
  filter("_Score >= 5").\
  filter(length("_Body") <= 1000).\
  withColumn("_Body", html_to_text("_Body")).\
  select("_Id", "_Title", "_Body", "_ParentId").\
  toDF("id", "title", "body", "parent_id")

# Self-join to assemble questions and answers
qa_df = gardening_df.alias("a").filter("parent_id IS NULL").\
  join(gardening_df.alias("b"), on=[col("a.id") == col("b.parent_id")]).\
  select("a.title", "a.body", "b.body").toDF("instruction", "context", "response").\
  select("instruction", "context", "response", lit("closed_qa").alias("category"))
```

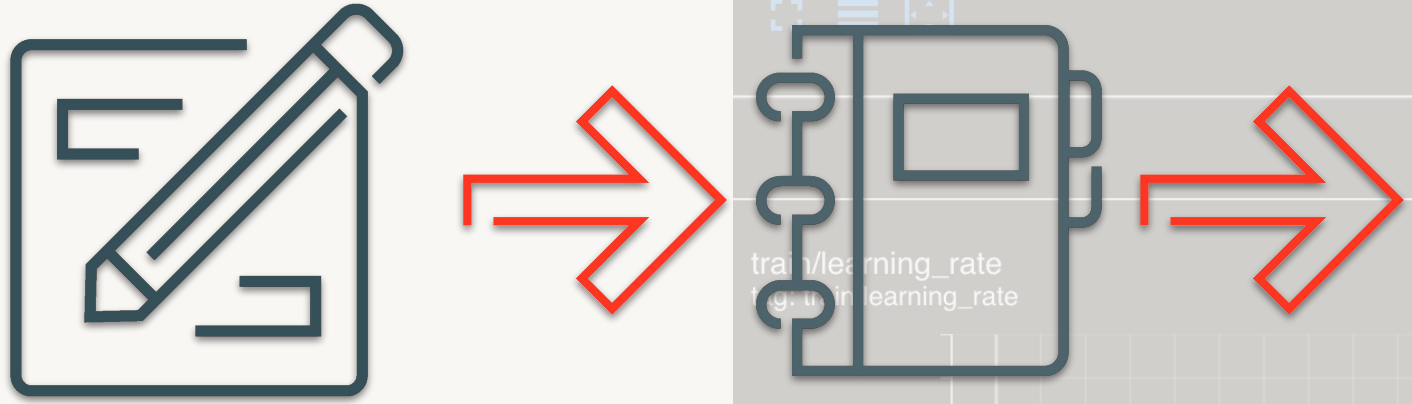

```
1  qa_df.toPandas().to_json(path_or_buf="/dbfs/.../gardening.jsonl", orient='records', lines=True)
```

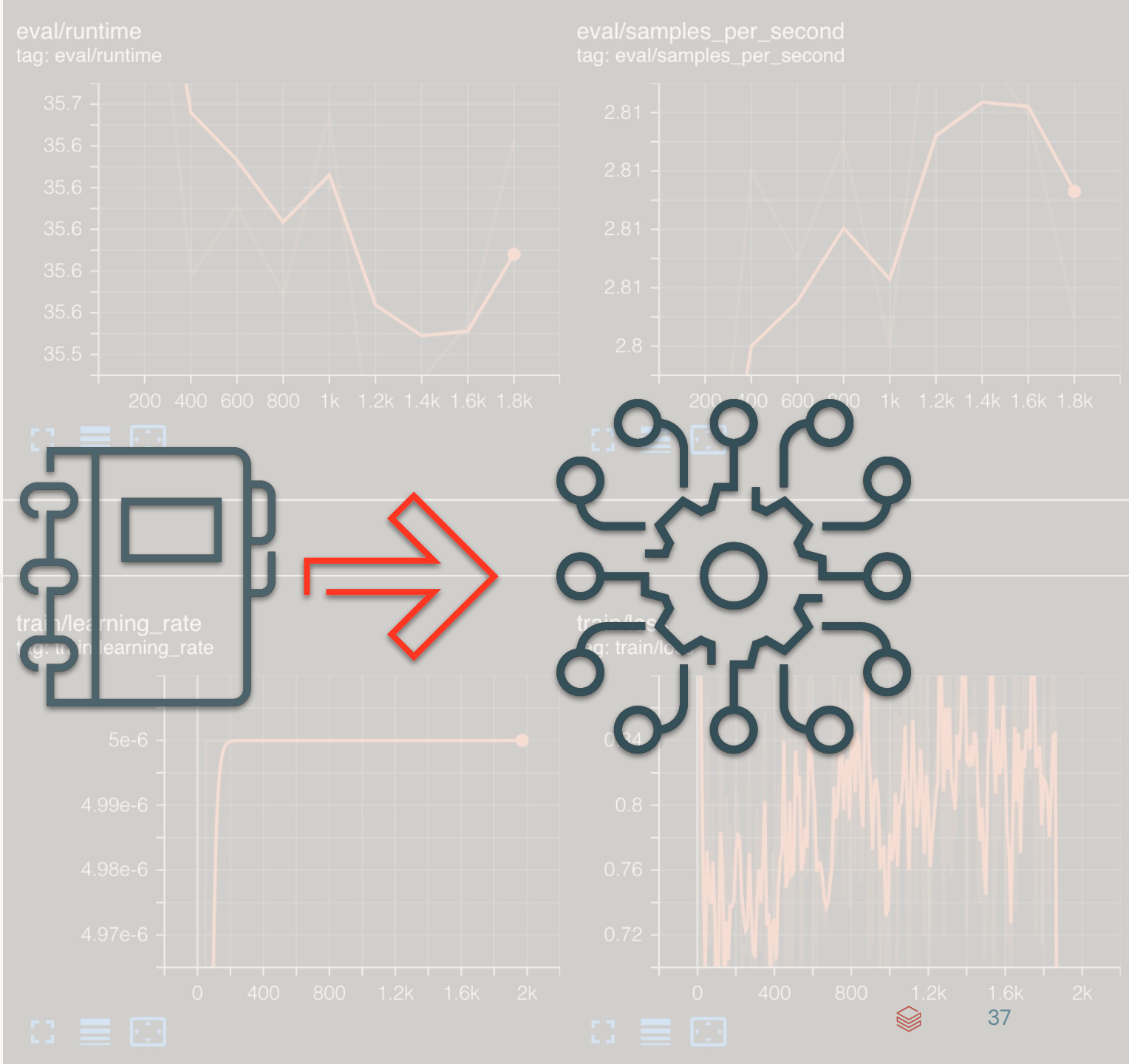
How Much Data?

- Not just how much data, but how many passes (epochs)
  - Tokens, not lines / bytes
  - Data tokens × epochs is more relevant
- Consider scale of model and pre-training
  - Ex: Pythia 12B saw 300B tokens (1 epoch over The Pile)
  - Ex: Dolly 12B fine-tuned on ~30M tokens
  - Here: ~1.5M tokens
- Too little training? *Not much effect*
- Too much training with little data? *Overfitting*
- Too much data with lots of data? *Forgetting*



Fine Tune





Fine Tune

- Training tuning tips at: `github.com/databrickslabs/dolly`
- Consider 3B/7B over 12B
- Prefer A100 (or A10)
- Will want multiple GPUs
- Memory is bottleneck
- Tradeoff memory and speed with deepspeed

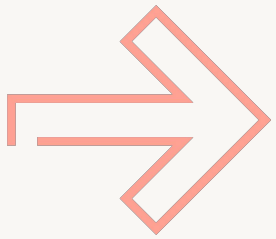

```
!deepspeed {num_gpus_flag} \
    --module training.trainer \
    --input-model {input_model} \
    --deepspeed {deepspeed_config} \
    --epochs 3 \
    --local-output-dir {local_output_dir} \
    --dbfs-output-dir {dbfs_output_dir} \
    --per-device-train-batch-size 3 \
    --per-device-eval-batch-size 3 \
    --logging-steps 50 \
    --save-steps 500 \
    --save-total-limit 4 \
    --eval-steps 100 \
    --warmup-steps 50 \
    --test-size 200 \
    --lr 5e-6
```

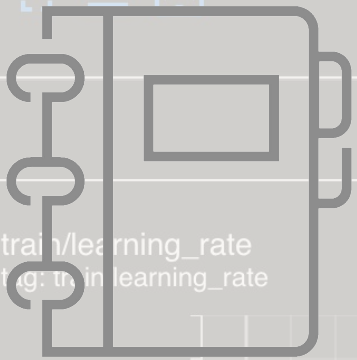
Tuning Fine Tuning

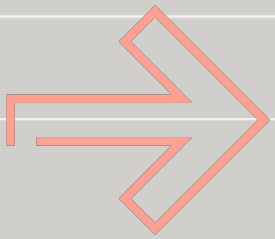
- Work in 16-bit: float16/bfloat16
  - `bf16` for Ampere: A10/A100
- Turn down batch size
- deepspeed
  - Enable gradient checkpointing
  - Enable optimizer offload
  - Enable param offload
- What about LoRA / PEFT?
  - Train smaller 'adapter'
  - Faster tuning, not inference

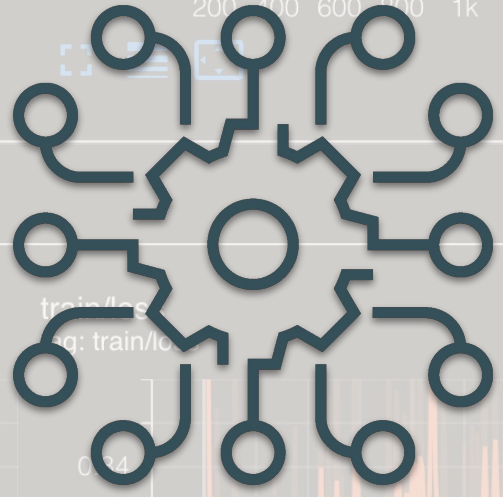

```
!deepspeed {num_gpus_flag} \
    --module training.trainer \
    --input-model {input_model} \
    --deepspeed {deepspeed_config} \
    --epochs 3 \
    --local-output-dir {local_output_dir} \
    --dbfs-output-dir {dbfs_output_dir} \
    --per-device-train-batch-size 3 \
    --per-device-eval-batch-size 3 \
    --logging-steps 50 \
    --save-steps 500 \
    --save-total-limit 4 \
    --eval-steps 100 \
    --warmup-steps 50 \
    --test-size 200 \
    --lr 5e-6
```

Generate

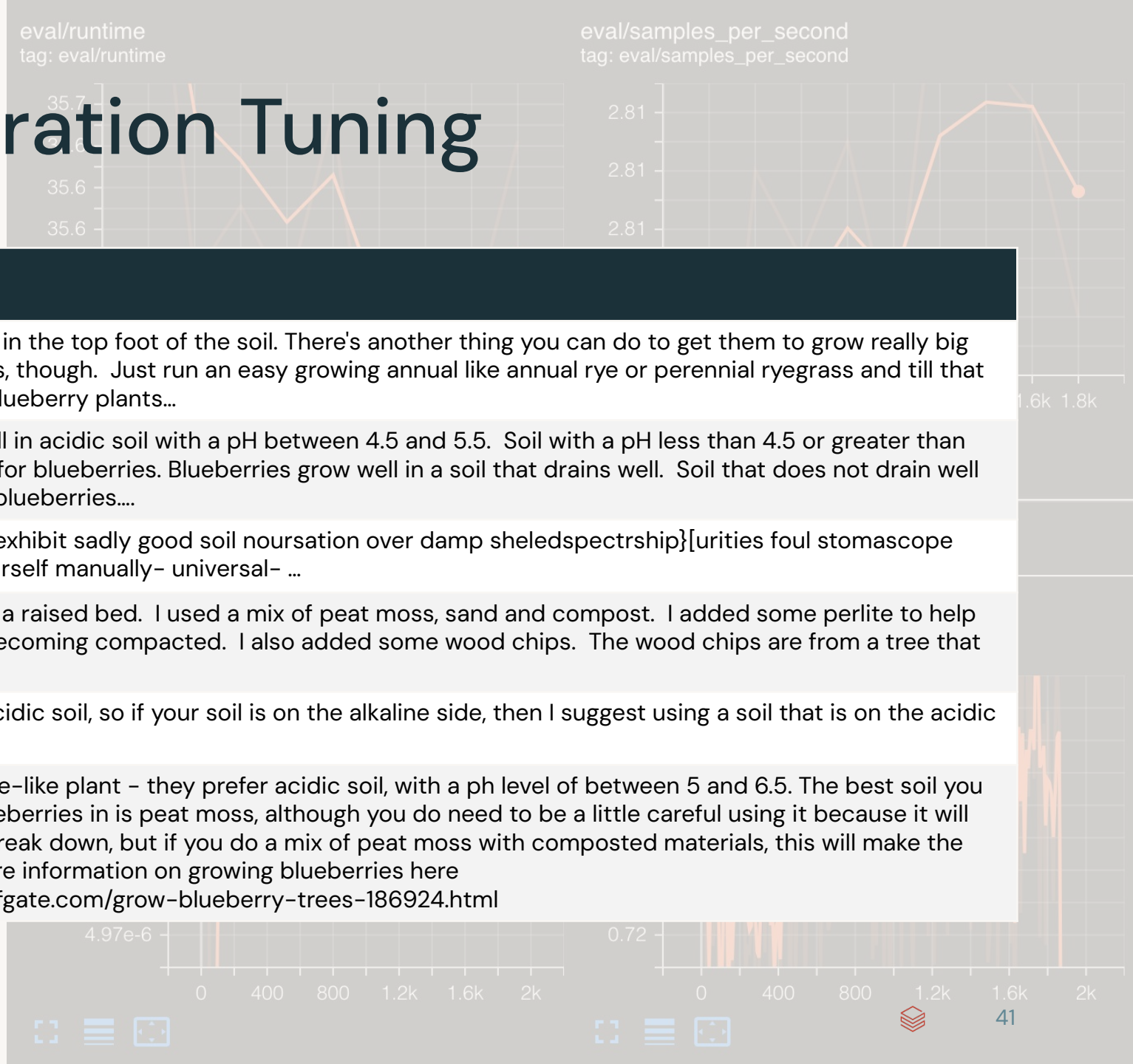








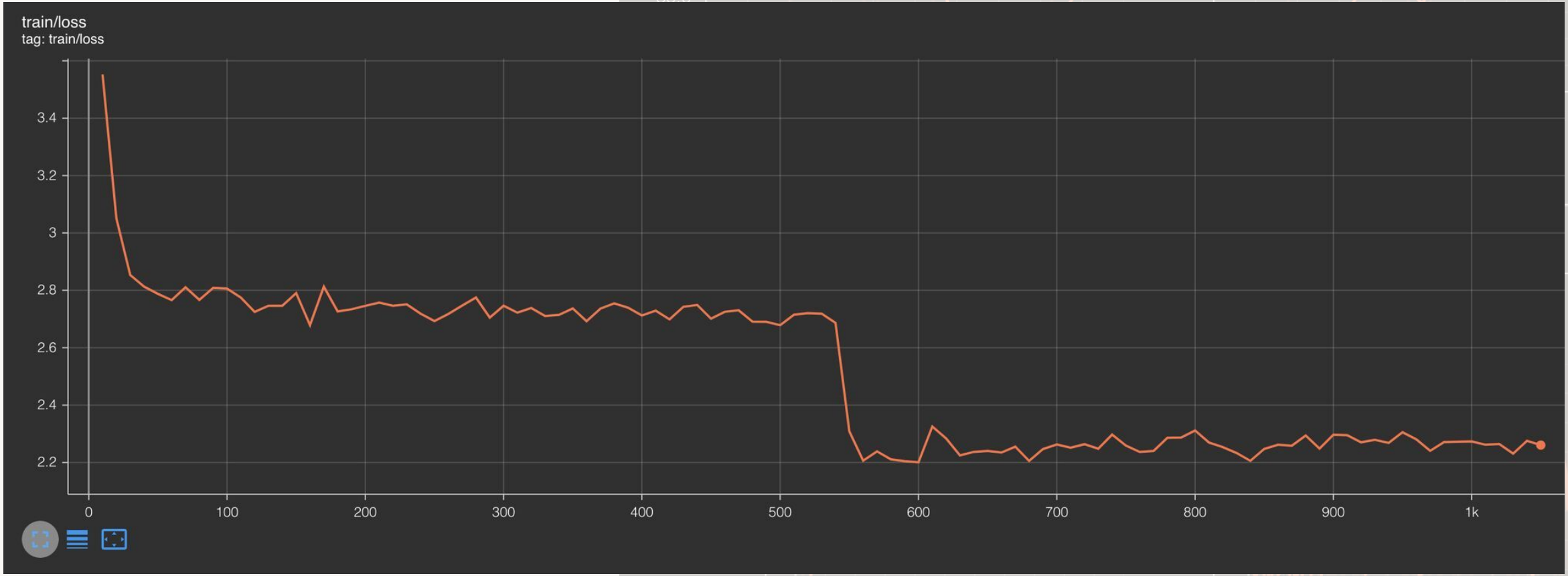
Fine-Tuned Generation Tuning



Setting	Change	Response
(default)		They'll grow just fine in the top foot of the soil. There's another thing you can do to get them to grow really big and delicious berries, though. Just run an easy growing annual like annual rye or perennial ryegrass and till that into the top of the blueberry plants...
`num_beams`	1 → 4	Blueberries grow well in acidic soil with a pH between 4.5 and 5.5. Soil with a pH less than 4.5 or greater than 5.5 will not be good for blueberries. Blueberries grow well in a soil that drains well. Soil that does not drain well will not be good for blueberries....
`temperature`	1 → 1.7	Watthemarn hedge exhibit sadly good soil noursation over damp sheledspectrship}[urities foul stomascope experiment bald yourself manually- universal- ...
`temperature`	1 → 0.3	I have blueberries in a raised bed. I used a mix of peat moss, sand and compost. I added some perlite to help keep the soil from becoming compacted. I also added some wood chips. The wood chips are from a tree that was cut down....
`top_p`	1 → 0.6	Blueberries prefer acidic soil, so if your soil is on the alkaline side, then I suggest using a soil that is on the acidic side.
`top_k`	50 → 10	Blueberries are a tree-like plant - they prefer acidic soil, with a ph level of between 5 and 6.5. The best soil you can buy to grow blueberries in is peat moss, although you do need to be a little careful using it because it will take some time to break down, but if you do a mix of peat moss with composted materials, this will make the soil more acidic. More information on growing blueberries here http://homeguides.sfgate.com/grow-blueberry-trees-186924.html

Serious Overfitting

Data Size << Parameters



Next Steps and Recap

Next Steps

- Try the demo! www.dbdemos.ai/demo.html?demoName=llm-dolly-chatbot
- See the accelerator! www.databricks.com/blog/driving-large-language-model-revolution-customer-service-and-support
- Check out a fine-tuning example! github.com/databrickslabs/dolly
- Try optimized GPU-enabled model serving! (*In private preview*)
- Check out Vector Index!

Recap

- Sometimes you want to customize LLM behavior
- Try off-the-shelf models first
  - Hugging Face `transformers` is your friend
  - Compose applications with tools like `langchain`
  - For QA: use vector DB for context retrieval
  - Explore and test generation settings
- Fine-tune if needed
  - Pick base model and data set carefully
  - Prepare to spend time optimizing training
  - Get to know `deepspeed`
- … but the most important takeaway of all is …

The best soil mix for blueberries is a porous, acidic mix high in organic matter. A good mix could be made by mixing these materials by weight: 5/8 topsoil, 2/8 peat moss, 1/16 vermiculite, 1/16 coarse sand.

Learn more at the summit!

Databricks Events App

Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.

What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!

Get trained and certified

- Visit the Learning Hub at the Databricks Zone!
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! databricks.com/learn

Thanks!

sean.owen@databricks.com