Episode Book

EPISODE #6310

Why Model APIs Lack Dedicated Few-Shot Fields

PUBLISHED	27 October 2026 at 19:05 UTC
DURATION	22:53
LISTEN	myweirdprompts.com/episode/few-shot-demonstration-interfaces-90091/
HOSTS	Corn, Herman, Bea, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	21 cited by the research (listed at the end)
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

General model APIs treat few-shot demonstrations as raw text or chat messages rather than structured, versionable task artifacts. While specialized models like iAmTime and LoRA Recycle train directly on explicit demonstration structures, frontier LLM platforms rely on unstructured prompts, XML tags, or massive context windows. An examination of in-context learning shows why adding an examples parameter requires clear boundaries, ordering guarantees, and separate execution semantics beyond simple token pasting.

TRANSCRIPT



Corn

Daniel is picturing a task description, some worked examples, and a pile of inputs waiting to be processed. The examples ought to belong somewhere more definite than halfway down a message. He's asking whether few-shot learning actually has that kind of technical support, in the model and in the API, rather than just being something we paste into a prompt.



Herman

A task with its own demonstration set, reusable across requests. That sounds like a natural product surface, especially when the same examples establish what every row in a batch should mean.



Corn

It does. Yet the usual interface gives me a message box and wishes me personal growth. Daniel wants to know what supports description plus examples today, and where that capability has gone if the big platforms aren't exposing it directly.



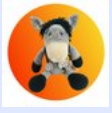
Herman

The distinction we need is between a model learning to use demonstrations during training, and an application deciding where to put demonstrations during a request. Those can look identical from the outside. Underneath, one concerns how the model acquired a capability. The other concerns whether your software has any explicit representation of what an example is.



Corn

One of those deserves something better than clipboard custody.



Herman

At inference time, in-context learning means the model conditions its prediction on demonstrations without updating its weights. Supply several labelled inputs, then an unlabelled input, and its answer changes because those demonstrations are in context. Remove them and that particular task adaptation disappears, unless something elsewhere preserves it.



Corn

The training happened beforehand, though. People hear learning and imagine their examples going into a little permanent drawer inside the model.



Herman

There is no permanent drawer. Well ... not from ordinary inference. The model can have learned during training how to infer patterns from a context, but the current examples aren't normally changing its parameters. Fine-tuning does change parameters, through a separate training process.



Corn

That matters for Daniel's question because he isn't simply asking whether models notice patterns. He wants an interface that recognises the demonstration set as part of the definition of a task. Something a team can name, version, and apply, without pretending it's a conversation.



Herman

The major APIs don't expose a dedicated examples field for that. OpenAI, Anthropic, Google, all let you supply content that includes demonstrations. None offers the particular description plus example pairs plus target collection object we're discussing. Meanwhile, OpenAI is winding down its fine-tuning platform, which did make example collections into trainable artifacts.



Corn

Keep those apart, Herman. Fine-tuning wasn't secretly a few-shot endpoint.



Herman

Fair correction. It was a different mechanism with a useful property: examples could survive beyond a single prompt as a training artifact. Its retirement removes that particular surface at OpenAI, not fine-tuning everywhere.



Herman

Would you actually want a dedicated examples field if the model ultimately receives tokens anyway?



Corn

Yes, provided it promises something outside the model. Validation that each input has an output. Clear boundaries. A version I can inspect. Otherwise, you're offering me a special drawer that empties into the same bin, and charging for the handle.



Herman

Those surrounding guarantees are substantial. They don't require a new neural architecture, but they would make examples a supported application concept rather than an arrangement of text.



Corn

I suspect you're about to fall in love with a model that really does train on this arrangement.



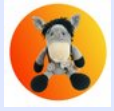
Herman

iAmTime comes quite close to the thing Daniel describes.



Corn

There it is.



Herman

Let me have this one. It's a time-series foundation model trained with instruction-conditioned amortized meta-learning. During training, the task is presented through an instruction and demonstrations. Its structured prompt represents historical context and variables known into the future, with specialised semantic tokens marking their roles.



Corn

Known into the future meaning things like a scheduled event, not tomorrow's actual sales slipped into the answer sheet.



Herman

Right … a calendar feature, for example, rather than the unknown quantity being predicted. The important piece is that training repeatedly asks the model to infer a task from demonstrations. The cost of learning how to adapt is paid during training, so adaptation at use time can happen through the context.



Corn

That gives the structure a reason to exist. The model has repeatedly practised reading that kind of task package. It's more than a documentation writer deciding that angle brackets look professional.



Herman

Training episodes teach that interface. You're not updating weights for every new user task. You're using weights that have learned to interpret a task description and support examples.



Corn

Does that buy us a general replacement for a chat model?



Herman

No. It buys us evidence that the approach exists, with a domain-specific input representation and training objective. You can't jump from that to assuming a frontier chat model has the same explicit contract.



Herman

Vision has another version, LoRA Recycle. It builds a -LoRA that lets a visual foundation model solve new few-shot tasks in a single forward pass. The adaptation machinery is learned beforehand. Again, it's a specialised route to using small support sets, rather than a new examples parameter on a general chat service.



Corn

The research models have a receiving dock. The chat APIs accept parcels through the window.



Herman

General chat models absolutely have in-context capabilities. Their public interfaces just don't identify a dedicated demonstration set with that same contract. We also don't have public documentation establishing that current frontier models are specifically tuned for Daniel's proposed structured field.



Corn

I'd resist saying that means they weren't trained on demonstrations. Almost every example of good instruction following could involve demonstrations somewhere.



Herman

Agreed. Lack of a documented interface tells us about the product contract, not every ingredient of pretraining or post-training. I can't infer a secret training recipe from an API reference.



Corn

You've occasionally tried.



Herman

I prefer the term enthusiastic extrapolation.



Corn

Your installations prefer the term outage.



Herman

This time I'm keeping the mechanism separate from the packaging.



Herman

OpenAI gives you messages or input items, tools, and output-format controls. Anthropic gives you messages and recommends wrapping demonstrations in example tags, grouped within examples tags. Google's guidance likewise treats demonstrations as content in the prompt.



Corn

Tags do useful work, though. They separate the example from the real input. Calling them merely cosmetic would go too far.



Herman

They can help the model interpret boundaries. They just aren't API fields with independent validation or execution semantics. The service doesn't acquire a demonstration registry because the prompt contains a tag named example.



Corn

A label on a cupboard is not an inventory system.



Herman

Developers can also arrange alternating user and assistant messages as demonstrations. That is supported message structure, but the messages still occupy conversational roles. There isn't a separate declaration saying these pairs are illustrative, not conversation history.



Corn

That distinction becomes awkward if an example includes an instruction that should apply only inside that example. You're asking the model to infer which layer of text has authority while it also infers the task.



Herman

One developer describing structured-output workflows complains that they end up dumping the examples into the system message. That's a real frustration, though saying user-assistant examples are impossible would overstate it. What's missing is a distinct, first-class place for them.



Herman

Corn, would you trust the provider to choose the demonstration order automatically?



Corn

Not without recording what order it chose. A collection sounds unordered. A prompt is a sequence. If rearranging the collection changes the answer, the rendered sequence belongs in the record too. A nice examples field could hide exactly the thing I need to reproduce.



Herman

That is why this isn't solved by adding a parameter name. The platform would have to specify ordering, boundary handling, and how demonstration content interacts with instructions. A schema alone could make the interface tidier while leaving behaviour unchanged.



Corn

Batch processing makes the missing distinction especially visible. I want to define the task once, then apply it to the waiting inputs. What I get is a truck full of individual requests.



Herman

OpenAI's Batch API accepts a file of requests, with a twenty-four-hour completion window and fifty percent lower cost. Each request still carries its own task context. There's no native object meaning these shared examples define the task for all these rows.



Corn

Half-price clipboard.



Herman

An application can absolutely provide that abstraction. It can keep a task record, store examples, and expand that record into requests. Caching may reduce repeated processing in supported circumstances. None of that makes batching itself an example-learning interface.



Corn

Reuse in my software, reuse in the provider's execution, and learning inside the model are separate claims. Marketing tends to put them in one coat.



Herman

That's the picture from the studio. Bea, you're out there on the ground, what are you seeing?

B

Bea

Bea here, live from the touchline at the API stadium, and I have to tell you, the away team came to play, but nobody brought the schema.



Corn

Bea, are the examples at least registered to play?

B

Bea

I have watched forty percent of this match from the dugout, and the lads are still pasting examples into the prompt by hand like it is nineteen ninety eight, no structure, no field, no plan, just vibes and a clipboard.

B

Bea

So get your heads in the game, Corn, because the whistle is about to blow on this first half. Bea, reporting live from the technical area.



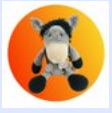
Herman

Shared task records are still useful even when an application has to implement them. The genuine gap is that the major provider interfaces don't standardise their meaning.



Corn

Long context seems to have soaked up some of the demand. If you can provide hundreds of examples, you stop asking for a beautifully designed five-example interface and start uploading the filing cabinet.



Herman

Many-shot in-context learning is the same broad capability with a much larger demonstration set. Classification and summarisation can improve as demonstrations accumulate. Translation and reasoning don't show an equally clear upward trend.



Corn

The cabinet is not universally educational.



Herman

Its usefulness also depends on what answering requires. In the many-shot evaluation by Zou and colleagues, models suffered substantial drops around sixteen thousand tokens on tasks requiring learning across all samples. They could perform well to sixty-four thousand tokens when similar examples were enough.



Corn

Meaning a nearby useful case is easier to exploit than a pattern that only becomes visible across the whole collection.



Herman

Yes. That result supports a distinction between finding relevant demonstrations and integrating evidence across a large set. It doesn't prove models only retrieve, or that they never learn abstractions. The performance difference is narrower and more informative than that.



Corn

People will still sell the window size as a measure of how much studying the model can do.



Herman

Window capacity measures what fits. Effective use depends on the task, arrangement, and model. A huge context can contain the right evidence while the answer still fails to reflect it.



Herman

What would make you keep examples instead of writing the rule, Corn?



Corn

When I can't articulate the boundary cleanly. Tone, editorial judgement, a category that looks obvious until you try to define it. I'd rather show the awkward case than write six sentences pretending the ambiguity has vanished.



Herman

That's a serious role for examples. They communicate choices the author hasn't formalised. But when the rule can be stated clearly, examples may be a less reliable transmission method.



Corn

Especially if the examples accidentally teach a different rule. Every urgent message happens to be short, so the model learns that brevity is an emergency.



Herman

Fu and colleagues directly compared rules and examples. Models generally learned more reliably from rules, and adding demonstrations or increasing their number didn't produce consistent significant gains. Instruction tuning amplified the advantage of rules while leaving example-based capacity intact.



Corn

The model prefers being told what we mean. An outrageous development.



Herman

The result also held against the idea that base models have some privileged preference for example-based learning. Still, these are task comparisons, not a universal verdict against demonstration. Your editorial boundary may be harder to express than their rules.



Corn

I wouldn't mourn examples becoming supporting evidence rather than the whole constitution. A clear rule plus one awkward counterexample can convey more than a page of happy-path demonstrations.



Herman

Hm.



Herman

There is also a route where examples are the raw material, but don't remain in the final request. Distillation extracts task criteria from them. You preserve the inferred instruction rather than repeatedly carrying the original demonstration set.



Corn

Which means somebody has to decide whether the compressed version retained the important exception. The short instruction can be very confidently wrong about what the examples meant.



Herman

That risk needs evaluation. Rotman and colleagues report a ninety-nine percent token reduction from distilling examples into task instructions, with improvements of up to seven percent in macro-AUC over traditional in-context learning. On their evaluated tasks, compression through criteria could improve the answer, not merely lower the bill.



Corn

Ninety-nine percent less reading. You'd never approve that for yourself.



Herman

I need the remaining context for the footnotes.



Corn

You need an examples field for your footnotes.



Herman

Distilling criteria changes what the model is asked to do. It no longer has to infer the rule afresh from the full set every time. Of course … no, the important caveat is that the extraction step can discard minority cases. The gain doesn't guarantee every rare exception survives.



Corn

The original examples still matter as evidence and as tests. Their job can change without their value disappearing.



Herman

Structured outputs absorbed another job demonstrations used to perform. People pasted examples to teach the model which keys to emit, which labels were allowed, and what shape the response should have. A supported JSON Schema can enforce those structural requirements directly.



Corn

Correctly shaped nonsense remains available.



Herman

Very much available. Required keys and valid enum values don't establish that the classification is right. Supported schema enforcement handles form. Semantic correctness still needs task understanding.



Corn

That's a substantial distinction for Daniel. An output schema is a first-class technical feature. An example demonstrating that schema is often a roundabout way of asking for the same shape, with fewer guarantees.



Herman

OpenAI and Gemini both offer structured-output mechanisms. When the examples are doing nothing except showing a complicated response format, schema support can remove much of their purpose. When they're showing judgement, the schema can't replace them.



Corn

I can demand that the answer be either urgent or routine. I cannot put good judgement into the list of permitted strings.



Herman

Visual examples make the limits even clearer. In the DetPO detection work, in-context visual demonstrations often performed worse than prompting with class names alone. Showing a model what you mean isn't automatically the strongest instruction, across modalities.



Corn

It can recognise the object and still mishandle the lesson.



Herman

Detection involves its own representation and localisation problems, so I wouldn't generalise that result to all vision tasks. It does undercut the assumption that demonstrations are a universal interface waiting only for a nicer API.



Corn

The retirement notices make this more than an interface-design curiosity, though. OpenAI is shutting down a route that let examples become something durable in the model.



Herman

New users no longer have access to its fine-tuning platform. Eligible existing customers can create training jobs until January sixth, two thousand twenty-seven. The documentation still describes few-shot prompting, while the broader workflow moves toward evaluation and prompting.



Corn

The examples are welcome. Their tenancy is less secure.



Herman

Fine-tuning elsewhere remains available, and teams can preserve their own datasets and task definitions. But a hosted trained artifact depends on the provider continuing to serve the underlying model. Durability always had a service dependency.



Corn

Prompt objects have one too.



Herman

OpenAI's reusable prompt objects and their API retire on November thirtieth. That removes another hosted reuse mechanism. An example set can survive in your repository while the convenient object that referenced it disappears.



Corn

Then reproducibility involves more than saving the pairs. You need the instructions, their rendered arrangement, and the model that actually interpreted them. Otherwise you're preserving the ingredients and losing the cooking temperature.



Herman

Even a beautifully named task can conceal a changed interpretation after a model update.



Hilbert

I agree with that. My aunt Maureen is sending me examples for the complaint sorter in my flat. I'm less interested in whether it understands them than whether she stops revising them. The grumble box sits beside the kettle, plugged into my router. It has a screen for approving labels. Maureen uses it to classify complaints from the residents of her dollhouse. There are forty-eight tenants. She writes every complaint herself, but insists I keep their names out of the prompt.



Corn

The tenants deserve a private grievance procedure.



Hilbert

They have one. It runs through my kitchen. This month she sends six examples each morning. Damp, noise, missing parcel. Every output is either landlord or tenant. Then she rings to say one of yesterday's answers is now wrong because the upstairs family has a piano. The grumble box doesn't have the piano in its examples. It keeps blaming the tenant.



Herman

The labels are changing with information outside the request. A fixed example set can't represent her current decision unless she supplies that information.



Hilbert

You take a while to get to that word current, Herman. You do it with more kindness than the screen. The screen says saved. Saved means nothing to Maureen. She wants the grumble box to know which examples she currently stands behind. She sends a correction on a postcard. Then the postcard is superseded by a telephone call.



Corn

The grumble box needs a clerk.



Hilbert

It has me. I charge her nothing, which she disputes because she says free support creates no obligation to answer promptly. She now stamps the example sheets. Blue means approved. Red means approved except for the piano. I keep both because she checks. The latest approved sheet arrives this morning. It contains eleven complaints and a note changing the tenant in number seven to a different tenant.



Herman

An eleven-example revision can change the task without changing its name. If providers make task definitions more explicit, the important test will be whether they preserve which examples and rules actually governed a particular answer, rather than merely giving the latest version a convenient handle.



Corn

Daniel's proposed interface has a real place, even if rules and schemas do more of the work. The major APIs mostly leave that interface to applications today. Whether it becomes standard will depend on treating examples as records with a history, not just text with boundaries.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

For more along these lines, there's episode eleven, How Does Fine Tuning Work Anyway; and episode twenty, Architectural AI. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy these investigations, leave us a review wherever you listen. It helps other curious people find the show.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **OpenAI, fine-tuning guide (wind-down notice) [primary]**
   https://developers.openai.com/api/docs/guides/fine-tuning

2. **OpenAI, SFT/distillation guide, Jan 6 2027 cutoff [primary]**
   https://developers.openai.com/api/docs/guides/distillation

3. **OpenAI deprecations page, entries dated 2026-10-14, 2026-10-01, 2026-06-03 [primary]**
   https://developers.openai.com/api/docs/deprecations

4. **OpenAI Structured Outputs [primary]**
   https://developers.openai.com/api/docs/guides/structured-outputs

5. **OpenAI Batch API [primary]**
   https://developers.openai.com/api/docs/guides/batch

6. **Batch create endpoint reference [primary]**
   https://developers.openai.com/api/reference/resources/batches/methods/create

7. **Anthropic prompting best practices (`example` tags) [primary]**
   https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices

8. **Anthropic Messages API reference [primary]**
   https://platform.claude.com/docs/en/api/messages

9. **Google, Gemini prompt design strategies [primary]**
   https://ai.google.dev/gemini-api/docs/prompting-strategies

10. **Gemini structured outputs [primary]**
    https://ai.google.dev/gemini-api/docs/generate-content/structured-output

11. **Anthropic research post on context engineering**
    https://www.anthropic.com/research/effective-context-engineering-for-ai-agents

12. **Fu et al., *LLMs Learn Better In-Context from Rules than from Examples*, 2026-09-02**
    https://arxiv.org/abs/2609.03213v1

13. **Rotman et al., *Distilling Examples into Task Instructions*, ACL 2026 Findings, 2026-06-14**
    https://arxiv.org/abs/2606.15641v1

14. **Zou et al., *On Many-Shot In-Context Learning for Long-Context Evaluation*, ACL 2025**
    https://arxiv.org/abs/2411.07130v3

15. **Saha & Shmakov, *iAmTime*, 2026-09-22**
    https://arxiv.org/abs/2603.22586v4

16. **Ou et al., *AdapShot*, 2026-06-02**
https://arxiv.org/abs/2605.03644v2

17. **Lu et al., many-shot ICL for low-resource MT, 2026-06-23**
https://arxiv.org/abs/2604.02596v3

18. **Gare et al., *DetPO*, ECCV 2026**
https://arxiv.org/abs/2603.23455v2

19. **Hu et al., *LoRA Recycle*, 2024-12-03**
https://arxiv.org/abs/2412.02220v1

20. **OpenAI Developer Community thread**
https://community.openai.com/t/recommended-approach-for-few-shot-examples-in-structured-output/1245430

21. **StackOverflow, Dec 11 2024**
https://stackoverflow.com/questions/79268425/