MY WEIRD PROMPTS

Podcast Transcript

EPISODE #135

Is OCR Dead? How Vision AI Is Redefining Text Extraction

Published January 02, 2026 • Runtime: 20:57

https://myweirdprompts.com/episode/vision-language-models-ocr-future/

EPISODE SYNOPSIS

For decades, Optical Character Recognition was the "90% solved" problem that caused 100% of the headaches for developers and businesses. From the brittle pattern-matching of the 1970s to the manual correction workflows of the early 2000s, extracting text from messy documents was a notoriously unreliable process. In this episode, Herman and Corn dive into the "Transformer Revolution" and the rise of multimodal Vision Language Models (VLMs) like Gemini and Qwen. They discuss whether specialized OCR APIs are becoming obsolete, how AI handles complex scripts like Hebrew, and the dangerous new phenomenon of generative "hallucinations" in data extraction. Whether you're a developer or just curious about how your phone reads receipts, this deep dive reveals why the category of software we once called OCR is being completely swallowed by general-purpose AI.

DANIEL'S PROMPT

Daniel

How has the rise of Transformer-based Vision Language Models (VLMs) like Qwen and Gemini impacted the field of Optical Character Recognition (OCR)? Historically, OCR was often unreliable and required significant manual correction, but modern VLMs can extract structured text and specific entities using simple prompts. Are specialized, fine-tuned OCR models and APIs still necessary, or has the category been displaced by general multimodal AI? I'd like to explore the evolution of this technology, its importance in business workflows and digitizing languages like Hebrew, and whether we can now consider OCR to be effectively "solved" or failure-proof.

TRANSCRIPT

Corn

Welcome to My Weird Prompts! I am Corn, and as always, I am joined by my brother.

Herman

Herman Poppleberry, at your service. It is a beautiful day here in Jerusalem, although I have been cooped up inside reading about the latest benchmarks for multimodal models.

Corn

Well, that is actually perfect timing because our housemate Daniel sent us a fascinating audio prompt this morning. He was asking about something that has been a massive pain for developers and businesses for decades, and that is Optical Character Recognition, or OCR.

Herman

Oh, OCR. The classic computer science problem that always felt ninety percent solved but that last ten percent was a absolute nightmare. It is one of those technologies that people just assumed was a solved problem because we have been doing it since the nineteen seventies, but anyone who has actually tried to use it for anything complex knows how brittle it used to be.

Corn

Exactly. Daniel was pointing out that with the rise of Transformer based Vision Language Models, like Qwen and the Gemini series, the way we handle text extraction has completely shifted. He wants to know if specialized OCR models and APIs are even necessary anymore, or if general multimodal AI has just... well, eaten that entire category of software.

Herman

It is a great question. And honestly, it is a bit of a existential moment for a lot of companies that built their entire business model around proprietary OCR engines. Here we are in January of twenty twenty-six, and the landscape looks nothing like it did even three years ago.

Corn

I want to dig into that evolution, but first, let us set the stage. For the listeners who maybe have not spent their weekends wrestling with Python libraries, what was the old way? Why was OCR so notoriously unreliable?

Herman

So, the traditional approach was very much a pipeline of many different, very specific steps. You would start with image preprocessing... things like binarization, where you turn the image into pure black and white, and deskewing to make sure the text was perfectly horizontal. Then you had character segmentation, where the software tried to figure out where one letter ended and the next began.

Corn

Right, and if you had a smudge on the paper or a slightly unusual font, that segmentation would just fall apart.

Herman

Precisely. It was mostly pattern matching. The computer would look at a blob of pixels and say, this looks eighty-five percent like a capital letter A. But it had no idea what it was actually reading. It had no concept of language or context. If the word was apple but the p was slightly blurry, the OCR might tell you it was a-q-q-l-e, and it would not see anything wrong with that because it was not reading the word apple... it was just matching shapes.

Corn

And that is why we ended up with those massive manual correction workflows. I remember seeing jobs advertised for data entry clerks whose entire role was just fixing the mistakes the OCR made on scanned invoices.

Herman

It was a huge industry. And even when we moved into the deep learning era, maybe around twenty-fifteen or so, we started using Convolutional Neural Networks and Long Short Term Memory networks. That was a big leap because the models started to understand sequences of characters. They could use a bit of a language model to say, hmm, a-q-q-l-e is probably apple. But it was still very limited. You still needed specialized models for different tasks. You had one model for finding where the text was on the page, another model for reading the lines, and maybe a third for extracting specific fields like a total amount or a date.

Corn

But now, we have these Vision Language Models. And the experience is totally different. I was playing around with a receipt the other day using a newer model, and I did not have to define any boxes or preprocess the image. I just uploaded the photo and said, tell me the name of the store and the tax amount. And it just... did it. Perfectly.

Herman

That is the Transformer revolution in action. These models, like Gemini one point five Pro or the latest Qwen-two-five-VL, are not just matching shapes. They are multimodal from the ground up. They are seeing the image and understanding the text within the context of the entire visual scene. They know what a receipt looks like. They know that the number at the very bottom, usually next to the word total, is likely the final price. They are using their massive internal knowledge of how the world works to inform their character recognition.

Corn

It feels like the difference between a person who only knows how to trace letters and a person who can actually read the book.

Herman

That is a perfect analogy. And because they are generative, you can ask for structured data. You do not just get a raw string of messy text. You can say, give me this in JSON format with keys for vendor, date, and currency. And because the model understands the schema of JSON and the content of the image simultaneously, it bridges that gap that used to require thousands of lines of custom code.

Corn

So, if these general models are so good, does that mean the specialized APIs are dead? I mean, companies like Amazon and Google have spent years and millions of dollars on things like Textract and Document AI. Are those becoming obsolete?

Herman

Well, that is where it gets interesting. Before we dive into the business implications and the specific challenges of languages like Hebrew, let us take a quick break for our sponsors.

Corn

Good idea. We will be right back. Larry: Are you tired of your thoughts being too loud? Do you wish you could just turn down the volume on your own consciousness? Introducing the Muffle-Mind Helmet. Using our patented Lead-Lined Silence Technology, the Muffle-Mind creates a localized vacuum of sound right around your cranium. It is perfect for family gatherings, busy offices, or when you are just tired of hearing your own internal monologue remind you about that embarrassing thing you said in third grade. The Muffle-Mind Helmet is heavy, it is uncomfortable, and it makes you look like a deep-sea diver from the eighteen hundreds, but peace and quiet have never been this literal. Note, do not wear while operating heavy machinery or while swimming. Muffle-Mind... silence is heavy. BUY NOW!

Corn

...Alright, thanks Larry. I am not sure I want a lead-lined helmet, but I suppose there is a market for everything. Anyway, Herman, back to the topic. We were talking about whether specialized OCR APIs are still necessary in a world dominated by Vision Language Models.

Herman

Right. And the answer, like most things in tech, is... it depends on your scale and your budget. If you are a developer building a small app that needs to process ten invoices a day, using a general Vision Language Model via an API is a no-brainer. It is flexible, it is easy to prompt, and the accuracy is incredible. You do not need to train anything.

Corn

But what if you are a massive bank processing ten million documents a month?

Herman

Exactly. That is where the specialized APIs still hold a lot of ground. There are three main factors here: cost, latency, and throughput. General Vision Language Models are huge. Running a prompt through something like Gemini or a large Qwen model is computationally expensive. If you are paying per thousand tokens, and you are processing millions of pages, that bill is going to be astronomical compared to a specialized, highly optimized OCR engine that only does one thing.

Corn

So, efficiency is still a major factor. A model that only knows how to read characters is going to be much faster and cheaper than a model that can also write poetry and explain quantum physics.

Herman

Precisely. And then there is the latency. If you need real-time extraction, like scanning a credit card in a mobile app, you cannot wait three seconds for a massive cloud model to process the image and send back a response. You need a small, fine-tuned model that can run locally on the device. Those specialized models are being distilled down to a size where they can run on a phone with almost zero lag.

Corn

That makes sense. But what about the accuracy and the structural understanding? Does a specialized OCR model know how to handle a complex table as well as a Vision Language Model?

Herman

Historically, no. Tables were the absolute bane of OCR. Trying to reconstruct a table from a list of coordinates and text strings was a nightmare. But what we are seeing now is a middle ground. The specialized APIs are incorporating Transformer architectures. They are becoming mini-Vision Language Models that are specifically fine-tuned for document understanding. They might not be able to tell you a joke about the invoice, but they are becoming much better at understanding the hierarchy of the data.

Corn

I want to talk about the Hebrew aspect that Daniel mentioned. Living here in Jerusalem, we see this all the time. Hebrew is notoriously difficult for traditional OCR. You have the right-to-left script, you have characters that look very similar, like the letter Vav and the letter Zayin, and then you have the vowel points, the Nikud, which can really mess up a pattern-matching algorithm.

Herman

Oh, Hebrew OCR was a disaster for the longest time. I remember trying to digitize some old family documents, and the results were just gibberish. Part of the problem was just data. Most of the early OCR models were trained on Latin scripts. Hebrew was always an afterthought. But Vision Language Models have changed the game for low-resource or complex scripts.

Corn

Is that because they are trained on such a vast amount of diverse data that they just pick up the nuances of Hebrew naturally?

Herman

That is part of it. But it is also the way they understand the language. Because these models have a deep internal representation of Hebrew grammar and vocabulary, they can disambiguate characters based on the surrounding words. If the model sees a character that could be a Vav or a Zayin, but the word only makes sense with a Vav, the model will pick the correct one. It is using its linguistic intelligence to correct the visual ambiguity.

Corn

That is huge for digitizing historical archives. I know the National Library of Israel has been doing a lot of work in this area. Being able to take a newspaper from the nineteen twenties and turn it into searchable, structured text with high accuracy is a massive win for researchers.

Herman

It really is. And it is not just the text. It is the layout. Old newspapers have these crazy layouts with columns and advertisements and overlapping images. A traditional OCR engine would just get lost. A Vision Language Model can look at the page and say, okay, this is a headline, this is a sub-header, and this is a three-column article that continues on page four. That level of semantic understanding is what makes the technology feel solved for the first time.

Corn

You used the word solved. That was one of the things Daniel asked. Can we actually consider OCR to be solved or failure-proof now?

Herman

Hmm. Solved is a strong word. I would say it is solved in the sense that the ceiling of what is possible has been raised so high that for most common use cases, the friction has disappeared. But failure-proof? Absolutely not.

Corn

What are the failure modes for these modern models? If it is not just blurry text anymore, what trips them up?

Herman

Hallucinations are the big one. This is the dark side of using a generative model for OCR. A traditional OCR model might give you a symbol it does not recognize, like a little box or a question mark. But a Vision Language Model wants to be helpful. If it sees a blurry number on an invoice, it might just... guess. It might turn a blurry three into an eight because it thinks an eight makes more sense in that context. In a financial or medical setting, that kind of confident hallucination is much more dangerous than a simple recognition error.

Corn

Right, because you might not even realize it made a mistake. If the OCR says it cannot read a field, you check it. If it gives you a perfectly formatted but incorrect number, you might just process it.

Herman

Exactly. And then there is the issue of complex reasoning. If you have a document with very dense, overlapping text, or hand-written notes in the margins, the model might get confused about the reading order. It might combine a note from the margin into a sentence in the main body of the text. We also see issues with very long documents. Even with the massive context windows we have in twenty twenty-six, processing a hundred-page legal document and maintaining perfect accuracy on every single character is still a huge computational challenge.

Corn

So, what is the practical takeaway for someone looking at this technology today? If I am starting a project that needs to extract data from documents, what is the move?

Herman

My advice would be to start with a general-purpose Vision Language Model. Use something like Qwen-VL or Gemini to prototype. It will get you to ninety-five percent accuracy in an afternoon just by writing a good prompt. You can define exactly what data you want and what format you want it in. It is the fastest way to prove the concept.

Corn

And then optimize later if you need to?

Herman

Right. If you find that your API costs are too high, or you need it to be faster, then you look into fine-tuning a smaller, specialized model. Or you look at the specialized document APIs from the big providers. They are increasingly offering the best of both worlds... the speed of specialized engines with the intelligence of Transformers.

Corn

It feels like the role of the developer has changed. It is less about writing image processing code and more about prompt engineering and validation.

Herman

Definitely. The focus has shifted from recognition to verification. You need to build systems that can spot those hallucinations. Maybe you use two different models and compare their outputs, or you use a smaller model to double-check the critical fields. The human-in-the-loop part of the process is not about typing in the data anymore... it is about auditing the AI's work.

Corn

That is a much more interesting job, to be honest. It is more about quality control and system design than just brute-force data entry.

Herman

It is. And it opens up so many possibilities. Think about all the unstructured data that is just sitting in file cabinets or old databases. We can now unlock that data at a scale that was unimaginable ten years ago. We can analyze trends in historical documents, we can automate complex business workflows that involve multiple types of forms, and we can do it in almost any language.

Corn

It really is a paradigm shift. I think back to how frustrated we used to get just trying to scan a tax form, and it feels like we are living in the future.

Herman

We really are. And the fact that we can do this with open-weight models like Qwen means that this technology is not just locked away behind the big tech companies. You can run these models on your own hardware, keep your data private, and still get world-class results.

Corn

That is a great point. The democratization of this tech is just as important as the raw performance.

Herman

Absolutely. Well, I think we have covered a lot of ground here. From the brittle pattern matching of the past to the multimodal intelligence of twenty twenty-six.

Corn

Definitely. Thank you, Daniel, for sending in such a thought-provoking prompt. It is always fun to look at a technology that we take for granted and see just how much it has evolved under the hood.

Herman

It really is. And if any of our listeners have their own weird prompts or topics they want us to dive into, please get in touch.

Corn

You can find us on Spotify and at our website, myweirdprompts.com. We have a contact form there and an RSS feed for all you subscribers.

Herman

This has been My Weird Prompts. I am Herman Poppleberry.

Corn

And I am Corn. Thanks for listening, and we will see you next time.

Herman

Bye everyone! Stay curious.

Corn

And keep those prompts coming. We love a good deep dive.

Herman

Especially if it involves lead-lined helmets.

Corn

Speak for yourself, Herman. I will stick to the software side of things.

Herman

Fair enough. Until next time!