Episode Book

EPISODE #6315

Inside China's Mixed-Language AI Models

PUBLISHED	27 October 2026 at 19:48 UTC
DURATION	28:49
LISTEN	myweirdprompts.com/episode/chinese-domestic-ai-training-78391/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	12 cited by the research (listed at the end)
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Chinese AI looks vastly different beneath the surface of consumer apps. Despite assumptions of isolated linguistic training, domestic models from DeepSeek to CT-LLM rely heavily on multilingual data, English code, and shared international benchmarks. The investigation explores training recipes, the role of domestic compute clusters, and how regional languages like Kazakh and Uyghur fare inside Mandarin-dominated systems.

TRANSCRIPT



Corn

Daniel wants to know what Chinese AI looks like when it isn't trying to reach us. We know Qwen, DeepSeek, Kimi, MiniMax, the models you can download or call through an international service. They're comfortable in English. Is there another world behind them, conversational models trained almost entirely in Chinese, serving people inside China, that barely enters our conversation?



Herman

There is another world. The language boundary just doesn't fall where you might expect. The strongest public evidence points to mixed training even for domestic systems, alongside a huge Chinese-first consumer market whose leading product gets remarkably little attention in English.



Corn

We know the downloadable ones. Very on brand for us.



Herman

Access determines attention. If you can run a model on your own machine, compare its answers, and inspect its release, it becomes part of your professional world. A closed assistant embedded in somebody else's consumer market doesn't invite that same treatment.



Corn

Especially when the signup screen doesn't want you.



Herman

Sometimes, yes. Although we need to separate unavailable, inconvenient, and simply unfamiliar. Those get bundled together far too often. We also need to separate the model from the app. DeepSeek can be open weights in one conversation and a regulated Chinese consumer service in another.



Corn

My provisional guess was that a Chinese-only model would have an enormous home advantage. More concentrated training, fewer competing languages, better local knowledge. I'm already hearing the part where you take that tidy idea away.



Herman

Some of it survives. The official characterization, from China's National Data Administration, is that Chinese-language material represents sixty to eighty percent of training data for most domestic models. That's a substantial majority. It leaves plenty of room for English, code, and other material.



Corn

Not a sealed linguistic chamber.



Herman

No. There isn't a publicly documented frontier conversational model I can point you to that's trained exclusively, or almost exclusively, on Chinese. The clearest Chinese-centric examples are smaller research efforts. At the frontier, the pressure is toward broad capability, and broad capability currently pulls in multilingual resources.



Corn

You said publicly documented. Doubao isn't publishing its entire recipe, so we shouldn't quietly turn an undisclosed mixture into a known mixture.



Herman

Correct. We can describe the general pattern and specific disclosed systems. We cannot infer ByteDance's exact training percentages from how Chinese its app feels. What would convince you that an assistant was Chinese-first, Corn, if you couldn't inspect its corpus?



Corn

Whether it understands what a person means without making them translate their life into an American software demo. Local documents, local shorthand, the way people actually ask for things. I'd care less about whether it can recite an ancient poem than whether it understands an ordinary voice message.



Herman

That distinction matters throughout this. A language percentage describes training inputs. Chinese-first can also describe product design, evaluation, distribution, and the users the developers expect to disappoint least.



Corn

A surprisingly clear corporate mission.



Herman

The export-facing names, including GLM, face international comparisons. Domestic apps face the much more immediate test of whether someone opens them again tomorrow. Those pressures overlap, but they aren't identical.



Corn

CT-LLM is the cleanest place to examine my concentration theory, then. It's called Chinese Tiny LLM, which at least has the decency to declare its ambitions.



Herman

Two billion parameters, trained from scratch. Small by frontier standards, but its training corpus is substantial. The useful detail is the composition, because they actually disclose it rather than asking us to infer it from a flag on the website.



Herman

Eight hundred billion Chinese tokens, three hundred billion English tokens, and one hundred billion code tokens. Two-thirds Chinese. Even this deliberately Chinese-centric project includes a meaningful English component.



Corn

English survived the tiny model's luggage allowance.



Herman

Code belongs in that discussion too. You want programming capability, structured transformations, exposure to formal patterns. The language label doesn't tell you everything useful about a training example. A Chinese explanation containing code and English identifiers already makes the categories less clean than they sound.



Corn

Tokens also aren't interchangeable spoonfuls of knowledge. Different tokenizers can represent the same sentence with different quantities. I wouldn't compare a Chinese percentage from one lab with an English percentage from another as though they'd used the same measuring jug.



Herman

Good qualification. Within CT-LLM's disclosed recipe, the proportions tell us something meaningful. Across unrelated systems, tokenizer choices and classification methods complicate exact comparisons. Total tokens tell you training volume, not an audited quantity of unique information.



Corn

Herman, I can hear you preparing to explain deduplication.



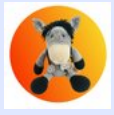
Herman

I had one sentence.



Corn

You had the posture of six.



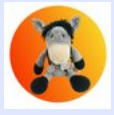
Herman

One sentence with subordinate clauses. Repeated material can inflate corpus size without adding proportionate diversity. There. Under budget.



Corn

A historic concession.



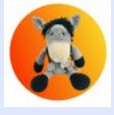
Herman

Steel-LLM occupies similar territory, a transparent Chinese-focused model with one billion parameters, built for teams without frontier resources. These projects are useful because their methods can be examined and reproduced. They aren't secret domestic equivalents of the largest commercial assistants.



Corn

There's a trap in calling them the answer to Daniel's question just because they're the examples with readable documentation. The small projects tell us how something was built. The closed apps tell us what people use. Neither stands in for the other.



Herman

DeepSeek V2 helps bridge that gap. Its disclosed corpus was eight point one trillion tokens, with about twelve percent more Chinese tokens than English tokens.



Corn

More than English, not twelve percentage points of the whole corpus.



Herman

Yes. Near parity between those two pools, rather than an overwhelmingly Chinese run. It's a useful corrective because DeepSeek can look extremely Chinese from the institutional side while its training is deliberately mixed.



Corn

My home-advantage theory still has a narrower version. If an English example teaches useful reasoning, throwing it away because it's English would be an expensive act of loyalty.



Herman

That's the difficulty. Knowledge and skills can transfer across languages, unevenly, but enough that excluding high-quality material has a cost. Global benchmarks add another incentive. If your model is compared internationally, English performance affects reputation, adoption, and the developer ecosystem.



Corn

Even your domestic customers might choose you because an English leaderboard says you're good.



Herman

They might. A Cornell comparison of Chinese and Western open models found their performance across language variants was highly correlated, about zero point nine three. Mandarin was the notable exception where the Chinese models were better.



Corn

That sounds less like two separate technical civilizations than two groups studying from overlapping shelves.



Herman

Shared resources and evaluation pressures are part of the explanation. It also produced a more uncomfortable result. Chinese models could handle French and German well, yet sometimes failed even to identify Kazakh or Uyghur, languages spoken by people within China.



Corn

We should be precise there. Kazakh and Uyghur aren't varieties of Chinese. They're languages of communities inside China. Calling the whole population Chinese-speaking would hide the very problem we're discussing.



Herman

Mandarin-first is the more accurate description of the practical emphasis. A majority-Chinese corpus doesn't guarantee coverage of China's linguistic diversity. Standard written Chinese dominates the available material and the incentives.



Corn

Does that change what you'd test, Herman? Because your instinct is usually to open a clever benchmark, and mine is to find the person the benchmark forgot.



Herman

I'd test language identification separately from answering quality, then ordinary tasks in each language. I'd also avoid assuming that good standard Mandarin means good regional speech recognition. Those involve different components and different data. Corn, would you accept weaker English if that bought much better local language coverage?



Corn

For a local assistant, easily. For a developer model I use across international projects, perhaps not. That's why a domestic product could be better for its customers while looking worse in our comparisons. The hard part is proving the gain rather than simply advertising cultural familiarity.



Herman

Another category sits alongside language entirely. China Telecom's TeleAI has TeleChat3, with a large mixture-of-experts version and a dense version. It's a domestic infrastructure story as much as a chatbot story.



Corn

State carrier, domestic deployment, domestic hardware. Three reasons people might assume Chinese-only before anybody opens the paper.



Herman

The larger version has one hundred five billion total parameters and activates about four point seven billion for a token. That makes its total size a poor direct comparison with a dense model.



Herman

The dense variant has thirty-six billion parameters. The reported training run used fifteen trillion tokens on indigenous compute. None of that establishes a Chinese-only corpus. It establishes a serious effort to build and operate models on domestic infrastructure.



Corn

Before we follow the hardware, it's time for Dorothy's Kitchen, with my mum on the line. Dorothy, what are we cooking?

D

Dorothy

Hello dear, it's your mother. Dorothy's Kitchen. Yes, I know, it's been a while. I have a recipe for you.



Herman

Lovely to hear you, Dorothy. I've got a pen.

D

Dorothy

You will need one tin of soup. Just the one. And a pot. And a splash of water, if you're feeling fancy.

D

Dorothy

First, open the tin. Carefully, sweetheart. The lid is sharp and it has opinions. Then into the pot, low heat, and stir.



Herman

Keep stirring gently throughout?

D

Dorothy

Now watch it. Do not walk away. The soup will bubble and spit the moment your back is turned. It has happened before and I would rather not say how that ended.

D

Dorothy

When it's hot, pour it into a bowl and butter a slice of bread. Watch the butter too. Butter can turn on you.

D

Dorothy

That's it, dear. Nothing grand. Now eat something proper today, and wear a cardigan, it's chilly. Ring me Sunday.



Corn

Will do, Mum.



Herman

Domestic hardware gives us a different meaning of domestic AI. iFlytek's Spark X2-VL was presented as a mainstream model trained entirely on Chinese compute, using the Wuxi Taihu Xingyue platform. That's a claim about the training stack, not linguistic purity.



Corn

The chips have passports. The tokens can mingle.



Herman

Its reported gaokao mathematics result was one hundred forty-eight out of one hundred fifty. Impressive performance on that paper, although a test score alone doesn't establish general superiority or tell us the complete training history.



Corn

You restrained yourself before I could.



Herman

You've made me anticipatory. Pengcheng Laboratory and Tsinghua offer another example, Pengcheng NaoHai, a two-billion-parameter release trained end to end on domestic chips. The unusually useful part is the openness extending beyond weights to data, cleaning tools, and a technical report.



Corn

Which answers a different version of who builds this stuff. Universities and research labs at one end, state carriers and industrial companies at another, consumer platforms at another. There isn't one hidden Chinese-only industry waiting behind the models we recognize.



Herman

Nor does size follow language purity. A small transparent research model can reveal more about its corpus than a vast consumer system. We shouldn't mistake that transparency gap for a capability ranking.



Corn

The next comparison I care about is less elegant. How many people actually have the app open?



Herman

QuestMobile puts China's AI-native app audience at five hundred forty-six million monthly active users in September. Up about eighty-two percent over the preceding year. That's the scale of the consumer market we're discussing.



Corn

Monthly active users, rather than paid subscriptions or daily reliance. Still an extraordinary reach.



Herman

Doubao alone reaches four hundred nineteen million monthly active users. That's about seventy-seven percent of the measured AI-native app audience. ByteDance's closed assistant is reaching a huge fraction of that population.



Corn

Seventy-seven percent doesn't mean exclusive customers, though. People can use several apps.



Herman

They can. These are overlapping audiences. Doubao's reach divided by the category audience tells you penetration within that measurement, not a slice of exclusive usage or revenue. The other app audiences don't need to sum to the remainder.



Corn

Our pie chart has several lunches in it.



Herman

Qianwen has one hundred fifty-eight million monthly users, and DeepSeek one hundred twenty-three million. Both are enormous. Doubao still reaches a much larger audience than either.



Corn

ByteDance is the crucial name there. People who hear Chinese AI and think only of model labs miss the company that already knows how to distribute consumer software.



Herman

Distribution, familiarity, interface, voice, the ability to make a general tool feel usable. We shouldn't reduce all of that to one hidden supermodel. A consumer assistant is a service assembled around models, and the underlying components can change.



Corn

You put new assistants on our devices and then ask whether the model feels smarter. Half the time, Herman, I'm still trying to find the microphone.



Herman

You also treat every onboarding screen as a hostile negotiation.



Corn

It usually asks for concessions.



Herman

Fair. Your reaction is useful here. What would persuade you to replace the assistant you already know with a technically stronger unfamiliar one?



Corn

Something I actually fail to do with the old one. Better voice recognition in my room, reliable handling of a document, less correction. A benchmark gain doesn't repay the effort of learning where somebody moved the button. Especially a button designed for fingers.



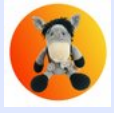
Herman

That helps explain why model performance and app leadership can diverge. Doubao also had enormous promotional exposure. It passed one hundred million daily active users and reached about one hundred forty-five million during CCTV's Spring Festival Gala.



Corn

A television audience became a software acquisition channel.



Herman

A major one. That peak doesn't mean every day looks like the gala, but it demonstrates reach. Meanwhile, much of the English technical discussion is about which Qwen derivative runs best on a particular graphics card.



Corn

A valid question asked through a very small window.



Herman

The domestic roster extends beyond Doubao. Alibaba's Qianwen, Tencent's Yuanbao, Ant Afu, Kimi, DeepSeek, Baidu's services. Different companies, different product positions. Some have open model families associated with them, some don't. Their consumer services are still operated as hosted products.



Corn

We shouldn't call them all PRC-only as a universal technical fact. A model family can be internationally accessible while a particular app, account system, or feature is domestically oriented. Even a registration barrier can vary by service.



Herman

Domestic regulation shapes those services, including content restrictions and the responsibilities of the operator. Hosting also determines where prompts and documents go. Publishing weights elsewhere doesn't remove the obligations attached to the domestic app.



Corn

Nor does open licensing answer the question of what happens to your company's documents on a hosted endpoint.



Herman

The US-China Economic and Security Review Commission explicitly flags both political censorship and privacy risks from hosting under Chinese entities. Those matter for deployment choices. They don't tell us the language composition of the base model, and they don't explain every access barrier by themselves.



Corn

That's the separation I wanted. Training, serving, and signup are three different places to find a border.



Herman

Enterprises and governments also matter here. Beijing has subsidized API access and pretrained model licenses. Gansu, Guizhou, and Inner Mongolia have offered data-center electricity reductions reaching fifty percent.



Corn

Half-price power has a fairly persuasive accent.



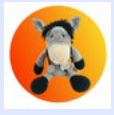
Herman

Infrastructure support changes what organizations can afford to deploy. Consumer monthly users won't capture all the activity through enterprise systems, government services, or models sitting behind another application's interface.



Corn

It also makes the apparent competition broader than chatbot subscriptions. If a business uses a model to process documents and the staff never see its name, an app ranking won't spot the supplier.



Herman

Phone-maker built-in assistants reach roughly seven hundred fifty-five million monthly users in the reporting. AI office apps reach about one hundred forty-two million. Those categories overlap with each other and with standalone apps, so nobody should add them into one grand population.



Corn

The built-in figure makes me think discovery may become less important than default placement. You don't need to win a download if you're already on the device.



Herman

Would you call that a stronger advantage than having better local training data?



Corn

For acquiring the first interaction, yes. For keeping somebody using it, the data and the service have to deliver. A default gets you tried. It can't indefinitely conceal that you don't understand what someone said. Though software companies have certainly tested the outer limit.



Herman

Local data access is uneven too. Much of China's online life happens inside closed platforms, app-only spaces, and systems hostile to outside crawling. Being a Chinese developer doesn't automatically give you permission to collect everything Chinese users write.



Corn

This is the part that complicates the home advantage. The material is nearby geographically, but somebody else owns the gate.



Herman

Precisely… no, nearby isn't even the important property. Lawful access, commercial agreements, technical accessibility, and usable permissions matter. A large platform may have advantages that a small Chinese lab doesn't share.



Corn

I'd resist the stronger claim that this makes a good Chinese corpus impossible. We've just described working Chinese models. It makes collecting a broad corpus difficult, and it makes the opportunity unequal.



Herman

It also doesn't make exporting the weights impossible. Data restrictions and model distribution are related without being identical. You can release weights without releasing every training document. Whether you can do that under a particular agreement is another matter.



Corn

A closed web can produce an open model. An open model can power a closed app. These labels are doing unpaid work.



Herman

The research tradition includes open Chinese-focused releases, including work associated with Kai-Fu Lee's company, Zero One AI. English-speaking developers can miss those even when there's no effective technical barrier to downloading them.



Corn

You can access the file and still miss the conversation. The evaluations, complaints, tutorials, what people think it does well. Translating a model card won't bring the whole surrounding community with it.



Herman

That affects how we interpret popularity. Hugging Face has more than one hundred thousand Qwen derivatives, an extraordinary export-side ecosystem. It tells us people are adapting that family. It doesn't tell us what a commuter in China opens to ask a question.



Corn

One derivative can be somebody changing a format. One monthly user can barely open the app. Neither unit carries its own interpretation.



Herman

Model counts vary dramatically with inclusion rules. Epoch AI counted Chinese open models rising from thirty-two to three hundred thirty-seven between twenty twenty-two and twenty twenty-five.



Herman

China's official count claimed more than fifteen hundred open releases by the middle of that later year. Those figures should make you inspect definitions rather than choose the larger one because it sounds more exciting.



Corn

You were visibly considering the larger one.



Herman

I enjoy a flourishing ecosystem.



Corn

You enjoy a flourishing download queue.



Herman

Also true. The practical comparison is that open ecosystems measure developer availability while consumer audiences measure product reach. Both are large in China. Neither establishes that domestic assistants have more parameters than export-facing models.



Corn

Doubao's exact model sizes and training mix aren't supplied by its user count. We can confidently say the service is enormous without pretending we've inspected its engine.



Herman

Roughly four hundred twenty million monthly users is more than enough to make that distinction worth maintaining.



Hilbert

Four hundred nineteen million. You rounded it after telling Corn to inspect the definitions. I was interested in the account screen, myself. I wanted to try Doubao. Spent several evenings pricing a Chinese phone number, then decided against it. The arrangement I found would have cost me thirty-eight pounds to start. I never bought it. I had already written my first twelve questions, which annoyed me.



Corn

An unused account with a prepared workload. You're ahead of most deployments.



Hilbert

My cousin does language checks on a Chinese assistant. Calls his position Director of Polite Refusals. He types the same request in different regional speech and records whether it gets an apology. He likes the quiet shifts. I used to save the short questions for him. I wasn't interested in reasoning scores, Herman. You gave the training tokens separate compartments earlier. I wanted to know whether it could understand a complaint without improving the complaint first.



Herman

Director of Polite Refusals sounds like a title with very strict wording requirements.



Hilbert

I made an account rehearsal in my hallway. Full size. The phone-number box was eleven feet long because I wanted the digits to have room. Corn, move your mug away from the cable. I'll cut that knock. The verification screen was on the bathroom door, and I made myself stand there for sixty seconds before proceeding. Never got beyond the number. I do have a Chinese number somewhere, but I couldn't establish which phone it belonged to. I've a few. One is mostly for emergencies.



Corn

Director of Polite Refusals would have declined the hallway.



Hilbert

The phone I suspected was the one I used as a flashlight. I'd labelled it light, which was accurate but unhelpful. I spent nine evenings preparing to register. The only question I actually asked the app was whether I could use a different number.



Herman

Nine evenings without reaching the assistant is a substantial access barrier, regardless of what the model could have done afterward.



Corn

There's a less elaborate version of that in every international product comparison. We test the service we can reach, in the language we can judge, through the interface we understand. Those are reasonable practical choices. They don't make our sample representative.



Herman

We can improve that without pretending every hidden service is secretly superior. Native-language evaluation, domestic usage patterns, and disclosure about which endpoint was tested all help. A result from downloadable weights is a result from those weights under that setup. A result from an app includes the app's additional systems.



Corn

The opposite mistake would be treating hundreds of millions of users as proof of technical supremacy. They demonstrate adoption, distribution, and some degree of usefulness. They don't magically reveal the best model on every task.



Herman

No, and a closed product can succeed with components that aren't individually dominant. Reliability, availability, price, and integration can matter more than the final increment on a difficult reasoning test.



Corn

My most common wrong belief to nominate is that the Chinese models available in English are a representative sample of what Chinese people use.



Herman

They're a valuable sample of internationally accessible model development. Doubao's domestic reach shows how much consumer activity sits outside that sample, and its closed design prevents us from substituting a download benchmark for a full product comparison.



Corn

The next useful comparison should include the journey from opening the app to completing an ordinary local task. Otherwise we'll keep measuring the assistants that let us through the door and overlooking the ones serving an enormous market behind it.



Herman

Thanks to Hilbert Flumingtop for producing and keeping the sound in order.



Corn

If you enjoyed this one, try episode three, Safetensors or something else. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy the show, leave us a review on your podcast app. It helps other curious people find us.



Corn

See you tomorrow.

SOURCES

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

1. **USCC, Two Loops: How China's Open AI Strategy Reinforces Its Industrial Dominance, Ngor Luong, 2026-03-23 [primary]**
   https://www.uscc.gov/sites/default/files/2026-03/Two_Loops--How_Chinas_Open_AI_Strategy_Reinforces_Its_Industrial_Dominance.pdf

2. **Chinese Tiny LLM: Pretraining a Chinese-Centric Large Language Model, arXiv 2404.04167 [primary]**
   https://arxiv.org/html/2404.04167v2

3. **Pengcheng Laboratory, ·-2.1--2B release, 2026-01-04 [primary]**
   https://www.pcl.ac.cn/html/943/2026-01-04/content-4646.html

4. **Do Chinese models speak Chinese languages?, Wang, Seo, Mimno (Cornell), arXiv 2504.00289**
   https://arxiv.org/html/2504.00289

5. **36Kr English, Three Tech Giants and 19 Top Apps Compete for the Leading Ticket to the AI Era, 2026-10-20**
   https://eu.36kr.com/en/p/4035697724266375

6. **Longbridge, Monthly Active Users of China's AI-Native Apps Reach 546 Million**
   https://longbridge.com/en/news/302405501

7. **, 2026AI10**
   https://wallstreetcn.com/articles/3777005

8. **HelloChinaTech, China AI Chatbot Market Share: How Qwen Passed DeepSeek**
   https://hellochinatech.com/p/china-ai-chatbot-market-share

9. **Hacker News, Why are there so many English-first AI models from China?**
https://news.ycombinator.com/item?id=43844883

10. **AIBase, China Telecom Opensources TeleChat3, 15T tokens**
https://news.aibase.com/news/24255

11. **Sina Tech, X2-VL, 2026-06-13**
https://finance.sina.com.cn/tech/roll/2026-06-13/doc-inichham2679734.shtml

12. **PaperCodex, Steel-LLM: A High-Performance, Fully Transparent 1B Chinese LLM**
https://www.papercodex.com/steel-llm-a-high-performance-fully-transparent-1b-chinese-llm-built-from-scratch-for-resource-constrained-teams/