Episode Book

EPISODE #6243

Inside the Mixture of Experts Router

PUBLISHED	24 October 2026 at 12:07 UTC
DURATION	29:57
LISTEN	myweirdprompts.com/episode/mixture-of-experts-routing-explained-66526/
HOSTS	Corn, Daniel, Herman, Larry, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.20
SOURCES	15 cited by the research (listed at the end)
BOOK GENERATED	25 October 2026 at 06:19 UTC

SYNOPSIS

Mixture of experts models are often pitched as an efficiency trick — activate a fraction of the network, get more capability per compute. But the interesting part is the gate itself: what happens when a token arrives and only a handful of experts get picked? This episode digs into the routing mechanism, from the small learned linear map that scores each expert to the top-k selection that decides who does the expensive work. Along the way, we look at how DeepSeek V3 separates selection scores from mixing weights, uses a dynamic per-expert bias instead of a heavy balancing loss, and constrains routing around the physical hardware experts live on. We also ask whether experts actually specialize in anything we'd recognize — biology, programming, a particular language — or whether the picture is far less intuitive than that.

TRANSCRIPT



Corn

Choosing which weights touch your token can determine what a model does, what hardware has to talk, and what an observer might learn about your question. Daniel is here with us, and today he wants to get inside that decision rather than revisit the usual mixture of experts efficiency pitch.



Daniel

I want to understand what actually happens at the gate. A token arrives, some experts get picked, and others do nothing. What is making that choice? Then, when we inspect those experts, do we find biology and programming, or something much less intuitive?



Herman

The choice starts with something almost disappointingly small. Usually a learned linear map. The sophistication is in the hidden state arriving at that map, and in the expert networks trained behind it.



Corn

You look disappointed that it fits in one sentence.



Herman

Give me time. I can make a linear map quite inconvenient.



Daniel

Keep the model landscape as a quick frame, too. Are we talking about a handful of experts now, or hundreds?



Herman

Both, historically, and hundreds in several large open models. DeepSeek V3 activates eight routed experts out of two hundred fifty-six. Llama Four Maverick selects one out of one hundred twenty-eight. That difference matters, but the routing mechanism deserves the microphone first.



Corn

Eight consultants, none billing by the hour.



Herman

Inside an MoE block, the expert networks generally replace the ordinary feed-forward sublayer. Attention still mixes information between token positions. Then each position has a contextual hidden vector, and the router decides which expert feed-forward networks should transform that vector.



Daniel

The expert sees the token with its surrounding context already incorporated?



Herman

Incorporated into the representation, yes. It doesn't receive the whole sentence as a separate little chat session. It gets that position's hidden state. Each selected expert transforms it, and their outputs get combined using routing weights.



Corn

That distinction kills a very persuasive wrong picture. No miniature biologist reads your paragraph while a miniature programmer waits outside. Several numerical transformations contribute to one position, and the next layer can choose a different combination.



Herman

Corn, would you expect a word to keep returning to the same expert?



Corn

Before this conversation, probably. Especially a very distinctive word. Something like mitochondrion feels as though it ought to have a regular desk.



Herman

Its representation changes with context and depth. The same written token can take different routes. Conversely, unrelated written tokens can arrive at similar representations and take similar routes.



Daniel

Does every layer have its own router and its own experts, then? We aren't following one permanent expert through the whole model?



Herman

Typically, each MoE layer has its own set. Expert seven in one layer has no automatic identity relationship with expert seven elsewhere. Some architectures also retain dense feed-forward layers. DeepSeek V3, for example, doesn't make every transformer block an MoE block.



Corn

Expert seven has been transferred without references.



Herman

Also, these are learned parameters, not roles assigned in advance. Training starts without somebody labelling a network biology. The router and experts adapt together under the language modelling objective, along with whatever balancing machinery the architecture uses.



Daniel

Right. Make that gate concrete. I've got a hidden vector in front of me. What calculation turns it into a choice?



Herman

For the standard recipe, each expert has a learned scoring vector in the router. Take the hidden state and compute its dot product with each scoring vector. That produces one raw score per expert. Those scores are often called logits.



Corn

Every expert gets assessed, but only the winners do the expensive work.



Herman

The router is cheap relative to running all those feed-forward networks. A softmax can convert the scores into positive weights whose total is one. Then top k selection retains the highest scoring experts. Depending on the implementation, the retained weights are renormalized before combining the expert outputs.



Daniel

Is softmax doing any understanding there, or just turning the scores into something usable?



Herman

Just transforming scores. It makes them relative to the whole set, with higher logits receiving disproportionately more weight. It doesn't introduce semantic understanding. The learned scoring directions and the incoming representation determine the preferences.



Corn

I think I'd been giving the gate too much credit. It sounded like a separate classifier deciding what job needed doing. Really, most of that decision has already been prepared upstream.



Herman

Yes... although I'd narrow that. The router still learns its own scoring directions. It isn't merely reading a prewritten label. But its decision boundaries are simple compared with the network that constructs the hidden state.



Daniel

Top k is discrete. How does a hard choice learn? You can't smoothly differentiate changing your mind about which expert wins.



Herman

The selected weights and selected experts participate in ordinary gradient calculations. The hard selection itself doesn't have a smooth derivative across a change in membership. Training works with that sparse computation, sometimes with additional devices to encourage exploration or balance. We shouldn't describe it as every unselected expert receiving a full learning signal from every token.



Corn

An expert that never gets picked never gets better at being picked.



Herman

That feedback loop is a central training problem. A small early advantage can attract more tokens, which supplies more practice, which attracts more tokens. Meanwhile, unused experts consume memory and contribute little.



Daniel

That sounds like the balancing mechanism could end up making the decision as much as the token itself does.



Herman

It can influence the decision strongly. DeepSeek V3 gives us a particularly clear example, because its selection scores and its final mixing weights are deliberately separated.



Corn

Here comes the mechanism you want to adopt.



Herman

I do like it. You will object to the compulsory participation.



Corn

I object to being an auxiliary loss.



Herman

In V3, the affinities come from sigmoid rather than softmax. Each expert gets a bounded affinity independently at that stage. The routed choices are then selected using those affinities plus a dynamic per-expert bias.



Daniel

Independently means an expert's affinity doesn't automatically fall because another expert's rises?



Herman

At the sigmoid stage, correct. The selected original affinities are subsequently normalized for combining outputs. The extra bias affects who gets selected, but it is excluded from those final mixing weights.



Corn

Admission policy and influence after admission are separate.



Herman

Suppose an expert is overloaded. Its selection bias can be lowered. An underused expert gets a boost. That moves traffic without pretending the boosted expert had a stronger original affinity for the token.



Daniel

Why is that preferable to just adding a balancing penalty to the training objective?



Herman

A balancing loss asks gradient descent to satisfy another objective alongside predicting text. That can interfere with learning useful representations. Adjusting selection biases gives the system a more direct traffic control. V3 calls this auxiliary-loss-free balancing, although it retains a very small sequence-level balancing loss.



Corn

Free, with a small loss attached.



Herman

That naming needs the footnote. The main bias update speed is zero point zero zero one for most of training. The complementary sequence-wise loss coefficient is zero point zero zero zero one. Those are training controls, not semantic thresholds.



Daniel

The small sequence loss prevents one particular sequence from piling onto a tiny set, even if the overall batch looks balanced?



Herman

That's the purpose. Aggregate balance can hide extreme local concentration. Also, V3 keeps one shared expert active alongside its routed experts. That supplies a path available to every token rather than making every bit of common processing compete for a routed slot.



Corn

Shared expert means always used here, not an expert that merely happens to be popular across languages. Those sound similar, but they're different claims.



Herman

Important distinction. Later, when we talk about multilingual sharing, we mean observed reuse. Here, we mean an architectural component.

L

Larry

Do you ever feel like your model's mixture of experts is sending the wrong token to the wrong expert? Then you need RouterRite, the first over the counter routing remedy that unclogs your gating network and rebalances your load with a dynamic per expert bias term. RouterRite promises up to eight of the top experts firing for every token, no dropped tokens, and a softmax so smooth your specialists will never close another bracket out of turn.

L

Larry

Ask your doctor about RouterRite. Side effects may include mild monosemanticity, sudden expertise in LaTeX punctuation, a persistent bias update speed of zero point zero zero one, group routing that sends every thought to at most four nodes, and in rare cases, the uncontrollable urge to replace your entire feed forward network with two hundred fifty six routed experts and one shared expert. RouterRite is not for everyone. If you experience a load balancing loss lasting more than four hours, seek immediate semantic help.



Corn

Herman, that had your coefficient in it.



Herman

I supplied neither copy, dosage language, product positioning, nor the phrase semantic help. Returning to the shared expert, it handles common processing while routing can distribute more specialized work. There's another constraint before the final choices, too. V3 limits a token's routed traffic to at most four nodes.



Daniel

Physical machines are now affecting what gets picked?



Herman

The grouping is designed around deployment communication. Experts are distributed across hardware. Unrestricted selection could scatter each token across too many nodes, creating expensive communication. Group routing constrains the eligible choices so the model can exploit specialization without sending every token everywhere.



Corn

The hidden state has opinions. The network fabric has a veto.



Herman

That is part of the algorithm. Routing is shaped by representational affinity, load control, and communication constraints. V3 also reports dropping no tokens during training or inference, rather than discarding assignments that exceed an expert's capacity.



Daniel

Dropping there means skipping an expert computation, not deleting a word from the user's prompt?



Herman

Correct. Capacity-limited MoE systems can decline overflow assignments within a layer. The token can still continue through other paths, including the residual stream. No dropped tokens means those intended routed computations aren't discarded.



Corn

A fairly consequential difference hidden behind one cheerful phrase.



Herman

Now, Corn, imagine inspecting the routing records. What would convince you an expert was doing biology?



Corn

Topic correlation wouldn't be enough. Biology prose has particular vocabulary and sentence patterns. I'd want to know whether it also fires for the same operation outside biology, and whether changing that expert changes a biological task rather than just changing the prose.



Herman

That's the right obstacle. The router is a linear map of hidden states. Similar states tend to receive similar scores. Small differences near a top k boundary can still switch membership, so similarity doesn't guarantee identical selections. But the geometry constrains the routing.



Daniel

Can we identify what in that geometry it's using?



Herman

Sometimes. Position is one signal. Routing representations can support reliable prediction of whether a token lies in the first or second half of a sequence. They carry information beyond topic.



Corn

It can know where you are without knowing what you mean.



Herman

Careful... without that observation proving what it knows about meaning. Positional information and semantic information coexist. Also, gate choices often correlate with experts producing larger output norms. The gate isn't running every expert first to measure that. It's a learned association visible when you inspect the computations.



Daniel

Larger output norm means a larger change contributed to the hidden state?



Herman

A larger vector magnitude, yes. Its eventual influence also depends on its mixing weight and direction. A large output can reinforce something, cancel something, or point somewhere unhelpful. Magnitude alone doesn't give us a function label.



Corn

You'd have bought the loudest expert.



Herman

You keep the quietest appliance, Corn, even when it's unplugged.



Corn

It has excellent uptime.



Herman

Across inspected layers, expert diversity often grows with depth, with an especially pronounced last-layer outlier. That already warns against assigning one simple specialization story to the whole network.



Daniel

What does the strongest positive evidence look like, then? Something more specific than saying the labels are wrong.



Herman

The Expert Strikes Back describes fine-grained task specialization, including closing brackets in LaTeX. That's a recognizable operation, but much narrower than mathematics. Other experts show sensitivity to parts of speech, grammatical roles, or particular semantic operations.



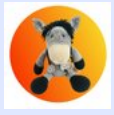
Corn

The mathematics department has hired a parenthesis.



Daniel

Is that literally a bracket detector, or something that helps generate the right bracket under the right circumstances?



Herman

The second is closer. Firing often on bracket tokens doesn't establish its entire computation. The relevant hidden states can encode an unfinished expression and the operation needed next. A fine-grained task specialist can help perform that transformation without owning every occurrence of the printed character.



Corn

This changes what I'd inspect. If I collect everything expert twenty touches and see lots of scientific equations, calling it a science expert would throw away the useful finding. The common operation might occur inside those equations and nowhere in the surrounding explanation.



Herman

Across six popular MoE models, routing paths carried strong predictive information about parts of speech. In a separate controlled analysis, roughly forty-four to fifty-four percent of measured specialization was classified as syntactic rather than semantic.



Daniel

That's not a universal percentage for all MoE models?



Herman

No. It belongs to that analysis and its definitions. The broader point is that syntax can account for a substantial share of what initially looks like semantic grouping.



Corn

Expert means an operation that found a useful patch of representation space. Sometimes we can name the operation. Sometimes we can only describe where the tokens cluster.



Herman

Sparsity can make those patches cleaner. Expert neurons tend to be less polysemantic than neurons in comparable dense feed-forward networks, with the difference widening under sparser routing.



Daniel

Less polysemantic means fewer unrelated things making the same neuron activate?



Herman

Broadly. One neuron can respond to several unrelated features, which complicates interpretation. Sparse routing can reduce that mixing at both neuron and expert level. It does not guarantee that every expert becomes a single understandable function.



Corn

A cleaner drawer can still contain something unlabelled.



Herman

Corn, would you rather have a model with cleaner individual experts if the overall routes remained hard to interpret?



Corn

For diagnosis, probably. For confidence in a whole answer, not automatically. I could understand one selected expert perfectly and still misunderstand why that hidden state reached it, or what the other selected experts did with it.



Herman

That distinction matters for model editing. A useful local handle doesn't make the full causal chain transparent.



Daniel

How much harder does generation make that? I could imagine inspecting my input and deciding which experts it will use, then being wrong as soon as it starts answering.



Herman

Prompt routing does not reliably predict rollout routing. During generation, the model consumes its own emerging text. Hidden states evolve, and the same user request can lead through different computations as the answer unfolds.



Corn

Your question doesn't reserve the experts for the answer.



Herman

The Myth of Expert Specialization in MoEs makes a stronger challenge to simple labels. In its comparisons, aligned expert overlap across models answering the same question was around sixty percent, no better than comparisons involving different questions.



Daniel

Aligned matters there. Expert numbers from independently trained models aren't naturally comparable.



Herman

They aren't. That result should be understood within the comparison procedure, not as matching arbitrary expert identifiers. What it undermines is the expectation of a stable, question-shaped team of specialists that reappears across models.



Corn

I also wouldn't turn sixty percent into sixty percent understanding. Shared activation can reflect common processing, frequent token shapes, or the measurement itself.



Herman

Deeper layers can show near-identical expert activation for semantically unrelated inputs, especially in reasoning models. That sits awkwardly beside the reports of fine-grained specialists, but the findings can coexist.



Daniel

Because they're looking at different models, layers, and types of specialization?



Herman

Those differences matter. Also, a repeated selection can still process different vectors differently. Identical expert membership doesn't mean identical computation inside the experts. We lose a lot by reducing a route to a list of numbers.



Corn

A routing trace tells you which workshop opened. It doesn't record every movement inside.



Herman

Load balancing complicates interpretation further. Its pressure can suppress shared directions in hidden states to maintain diverse routing. Under low-diversity data, that interaction may contribute to specialization collapse.



Daniel

Does that mean balance can work against specialization?



Herman

Sometimes. Complete concentration wastes capacity, but perfectly even traffic is not itself evidence of useful differentiation. The training system needs enough distribution to learn, without forcing artificial differences that damage prediction.



Corn

Equal attendance is a terrible exam result.



Herman

Multilingual routing makes that concrete. Some analyses find grouping along language families. High-resource languages tend to reuse more experts, while low-resource languages depend more on experts used predominantly for those languages.



Daniel

Are those language-specific experts enough to give a low-resource language its own protected capacity?



Herman

Not necessarily. Deep-layer routing collapse has been observed in low-resource languages, with traffic concentrating on a narrow subset. The corresponding English behavior was much less affected. Continual pretraining corrected that concentration in the reported experiments.



Corn

That's a nasty inversion. The language apparently has its own route, but that route may be too narrow to use the model well.



Herman

Early and late layers can be more language-specific, while middle layers behave more like shared capacity. We should resist treating that as a universal layer map, but it gives a more useful picture than one expert per language.



Daniel

Can we do anything causal with these discoveries, or are we mostly attaching descriptions to activation patterns?



Herman

Causal steering is possible in limited settings. One geometric routing experiment shifted hidden states toward a temporal expert's centroid and raised the measured temporal category probability by three hundred twenty-one percent.



Corn

Relative increase, not three hundred percentage points.



Herman

Relative. Suppressing a geographic expert reduced its measured geographic category probability by twenty-three percent. Those are particular category measurements, not a guarantee that you can install a reliable chronology switch.



Daniel

Still, that gets beyond watching a correlation. You're changing something and seeing a related behavior move.



Herman

It does. In that work, about fifteen percent of experts qualified as monosemantic specialists across the tested categories. That's a useful foothold, alongside a large remainder that didn't fit such neat labels.



Corn

I'd worry about changing the input geometry rather than editing an isolated faculty. Moving toward a centroid may alter several features at once.



Herman

That is the right caution. An intervention can establish influence without proving that our verbal label captures everything being changed.



Daniel

Put the expert counts around this now. If specialists are that small and mixed, does increasing their number make the structure easier to tease apart?



Herman

It can offer more distinct computational combinations, but interpretability doesn't follow automatically. The compact landscape is this. Mixtral selects two of eight experts. DBRX selects four of sixteen. Those are comparatively coarse menus.



Corn

You can still memorize those without a spreadsheet.



Herman

DeepSeek V3 selects eight of two hundred fifty-six routed experts, alongside its shared expert. Its total model size is six hundred seventy-one billion parameters, with about thirty-seven billion active per token.



Daniel

Active includes more than just those routed experts, though?



Herman

Yes, attention, shared components, and other active weights contribute. Expert fractions and whole-model active fractions are different quantities.



Corn

Counting guests tells you nothing about the kitchen.



Herman

Kimi K2 selects eight of three hundred eighty-four routed experts. GLM Four Point Five uses eight routed choices from one hundred sixty, plus a shared expert. Qwen Three Coder uses eight of one hundred twenty-eight without a shared expert.



Daniel

Llama choosing one sounds like the opposite design instinct.



Herman

Maverick's one of one hundred twenty-eight is a sharp contrast. There is no agreed optimal number of active experts. More, smaller experts offer combinatorial richness, but routing overhead, communication, training behavior, and kernel efficiency all matter.



Corn

The tempting claim is that eight choices must mean eight times the work. That only follows if the experts are equally large. Divide a larger expert into smaller ones and you can change the combinations while holding active parameter capacity roughly fixed.



Herman

Granularity theory formalizes that advantage. Choosing eight from two hundred fifty-six permits more than four hundred trillion possible selected sets. Under the theoretical comparison, matching that configuration count with a single-selection design requires an enormous menu.



Daniel

Possible sets aren't the same as useful sets, surely. The hidden states might only reach a fraction of them.



Herman

Precisely the limitation. Those are possible configurations, not observed behaviors or independently learned skills. The expressivity argument favors fine granularity under its assumptions. It doesn't prove that every additional expert pays for itself in a deployed model.



Corn

Four hundred trillion combinations, and some still forget the bracket.



Herman

More experts also mean more ways for routing records to reveal structure. If expert selections depend on hidden states that encode the query, those selections can become a privacy side channel.



Daniel

Even if you can't read the token contents directly, you might infer something from which experts are busy?



Herman

Potentially. That requires an observation channel and an inference method. Deterministic routing alone doesn't grant an attacker access. But usage patterns can leak information even when the observer cannot recover the whole text.



Corn

Interpretability for the operator can become interpretability for someone watching the operator.



Herman

What I'd want to examine is exactly what routing telemetry exposes, at what resolution, and across which users. Even coarse signals deserve attention when a model repeats the same distinctive paths.



Hilbert

Herman, would an expert register normally charge for experts that never receive a token? You separated selection from weighting earlier. I wish the borough's purchasing office had done that before charging me for the spare mouths.



Corn

An idle spare mouth sounds cheaper to feed.



Hilbert

It wasn't. I was the contractor operating their correspondence model. Mr. Edwin Vale in licensing wanted every expert registered separately. I printed the router list. He asked which ones gave advice. I said none of them gave advice separately. He gave each one an individual street-trading licence anyway. Two hundred fifty-six little licences. Nine pounds each. I had to display them on the machine.



Herman

Their definition of a trader seems to have skipped the trading.



Hilbert

Vale wasn't the difficulty. He kept coming down with the forms and a chair because my office only had one. The supervisor wanted portraits. I sent the same photograph of the server for every spare mouth. They rejected it for insufficient facial distinction. That held the invoice up for eleven weeks. The model was doing drainage letters throughout. Nobody had complained about the letters.



Corn

You should have routed the invoice to an unlicensed one.



Hilbert

I appealed the fee. Vale helped me word it. Very neat handwriting. They refunded the shared expert because it was classed as communal equipment, but the routed ones stayed on the register. I sold the server to a print shop. The licences weren't transferable. They posted the renewal notices to me in separate envelopes.



Herman

The spare mouths were registered against you rather than the machine.



Hilbert

Against the correspondence contract. I wrote cancelled on each notice. Vale said they needed a signature on each one. Two hundred fifty-six signatures. He brought a blue pen.



Corn

Separate envelopes are quite an achievement for something that only exists inside one computation.



Daniel

That distinction between components and independent entities does seem to keep causing trouble. If I say expert, people naturally expect something with a job title and a stable identity.



Herman

We can use the word, provided we keep asking what measurement supports the label. A routing preference, an activation pattern, and a causal effect on behavior are different levels of evidence. Ideally, you'd connect all of them.



Corn

I'd also keep the layer and model in the description. Saying an expert helps close LaTeX brackets in a particular setting is useful. Saying MoE has a bracket expert sounds like every architecture ships the same department.



Daniel

The clean answer to my mechanical question is much simpler than the clean answer to what the experts do.



Herman

The score computation is explicit. We can inspect the map, selection rule, bias adjustments, and output weights. Explaining why a particular hidden state landed there requires understanding what the rest of the model encoded.



Corn

That is where I withdraw my request for a biology desk. I still want a description of what an expert contributes, but I don't expect a human organizational chart.



Herman

One adjacent result worth keeping: a bandpass routing loss reduced the share of dead experts from roughly thirty to forty-five percent down to roughly zero to six percent in its experiments.



Daniel

Meaning the training rule can change how much of the supposed expert population actually participates.



Herman

Yes. Counting the available networks is only the beginning. Whether they learn distinct useful work depends on how traffic and learning reach them.



Corn

The open question I'd watch is whether those separate computational handles let us edit a behavior without disturbing everything sharing its route. Cleaner specialists make that plausible, but the spare mouths still belong to one model.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you want more of this, try episode three, Safetensors or something else; episode eighteen, Beyond the GPU; and episode nineteen, AI Images. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoyed getting inside the gate, leave us a review wherever you listen.



Daniel

Thanks for working through it with me. See you soon.



Corn

Until next time.

SOURCES

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

1. **DeepSeek-V3 Technical Report [primary]**
   arXiv:2412.19437v2 (2024-12, rev. v2).
   https://arxiv.org/html/2412.19437v2

2. **DeepSeek-V3.2 model card [primary]**
   Hugging Face, created 2025-12-01.
   https://huggingface.co/deepseek-ai/DeepSeek-V3.2

3. **Vinija's DeepSeek-V3 primer (routing/bias/group details).**
   https://vinija.ai/models/deepseekV3/

4. **Herbst, Wermert & Lee, The Expert Strikes Back**
   arXiv:2604.02178v2, ICML 2026.
   https://arxiv.org/abs/2604.02178v2

5. **Wang, Hayou & Nalisnick, The Myth of Expert Specialization in MoEs**
   arXiv:2604.09780v1 (2026-04-10).
   https://arxiv.org/abs/2604.09780v1

6. **Lo et al., A Closer Look into MoE in LLMs**
   arXiv:2406.18219v3, NAACL 2025 Findings.
   https://arxiv.org/abs/2406.18219v3

7. **Bershatsky & Oseledets, On the Spatial Structure of MoE in Transformers**
   arXiv:2504.04444, ICLR 2025 SLLM.
   https://arxiv.org/pdf/2504.04444

8. **Antoine et al., Part-Of-Speech Sensitivity of Routers in MoE Models**
   COLING 2025.
   https://aclanthology.org/2025.coling-main.431/

9. **The power of fine-grained experts: Granularity boosts expressivity**
arXiv:2505.06839.
https://arxiv.org/html/2505.06839

10. **Ternovtsii & Bilak, Cosine-Similarity Routing with Semantic Anchors**
arXiv:2509.14255v2.
https://arxiv.org/abs/2509.14255v2

11. **Ternovtsii & Bilak, Geometric Routing Enables Causal Expert Control**
arXiv:2604.14434v1 (2026-04-15).
https://arxiv.org/abs/2604.14434v1

12. **Chen et al., Understanding Multilingualism in MoE LLMs**
arXiv:2601.14050v1 (2026-01-20).
https://arxiv.org/abs/2601.14050v1

13. **Bar Joseph et al., Mixture of Experts for Low-Resource LLMs**
arXiv:2605.17598v1 (2026-05-17).
https://arxiv.org/abs/2605.17598v1

14. **Turing Post, Breakdown: Kimi K2, DeepSeek-R1, Qwen3, GLM-4.5 (Kimi K2 384 experts).**
https://turingpost.substack.com/p/breakdown-kimi-k2-deepseek-r1-qwen3

15. **heise iX, Neue große Open-Weight-Modelle im Überblick (GLM-4.5/Kimi K2/Qwen3-Coder table).**
https://www.heise.de/select/ix/2025/9/2519809511536754092