WHEN DOES CONSERVED-FLOW ROUTING HELP? A COUPLING CONDITION FOR MIXTURE-OF-EXPERTS, WITH A LEGALITY GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) models route each input through a few experts chosen by a learned gate, one layer at a time and independently. We ask when it is useful to instead use a *coupled* router, which commits to a single, globally consistent path of experts. We study a conserved-flow router: a differentiable Tero–Nakagaki (Physarum) dynamic that relaxes to one min-cost path through a graph whose edges are exactly the legal expert transitions, so illegal routing is impossible by construction rather than merely penalized. Our central finding is a boundary condition. When a task's output *marginalizes* over its intermediate steps, meaning each step can be scored independently, the coupling provides no benefit: on MetaQA 2-hop knowledge-graph QA the flow router *ties* a plain per-step softmax over the same legal graph. As the steps become mutually constraining, the flow's advantage over independent gating grows monotonically, a crossover we chart on a controlled constraint-density sweep. In a small MoE transformer on a coupled compositional task, the flow router cuts the illegal-transition rate from 0.34 to 0.01 at one dominant expert per layer; its accuracy reaches 0.99 on some seeds but is unstable across seeds and below a top-1 gate on average, which we report seed by seed. To make the approach practical, we use implicit differentiation through the flow fixed point and a matrix-free solver, and verify the resulting gradients against finite differences. The contribution is the routing mechanism, a specific condition for when it helps, and a hard legality guarantee, rather than a new accuracy record.

1 INTRODUCTION

Sparse Mixture-of-Experts (MoE) models (Shazeer et al., 2017; Fedus et al., 2022) scale capacity by activating only a few experts per input, chosen by a learned gate. The standard gate decides *independently* at every layer: it picks the best-scoring expert(s) for this layer without committing to a consistent route through the network, and nothing prevents it from chaining experts in an order the task forbids. A natural alternative is a *coupled* router that commits to one globally consistent path of experts. We study such a router built from a conserved flow: a differentiable Physarum dynamic (Tero et al., 2007; 2010) that relaxes to a single min-cost path through a graph whose edges are exactly the legal expert transitions, so illegal routing is impossible by construction.

The more important question is when this added coupling is actually useful. Our experiments point to a boundary condition. When a task's output *marginalizes* over its intermediate steps, so each step can be scored independently, a conserved-flow router does not improve over a plain per-step softmax on the same legal graph. MetaQA 2-hop knowledge-graph QA (Zhang et al., 2018) has this structure, and the two methods *tie*. When the steps constrain one another, the conservation coupling, which uses a shared flow budget to enforce one globally consistent path, helps, and the margin increases as the coupling becomes stronger (Fig. 1). Contributions:

- **A condition for coupled routing**: a conserved-flow router beats independent per-step gating *iff* the steps are coupled; we show this crossover with a controlled constraint-density sweep and use it to predict the MetaQA 2-hop tie (§4).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

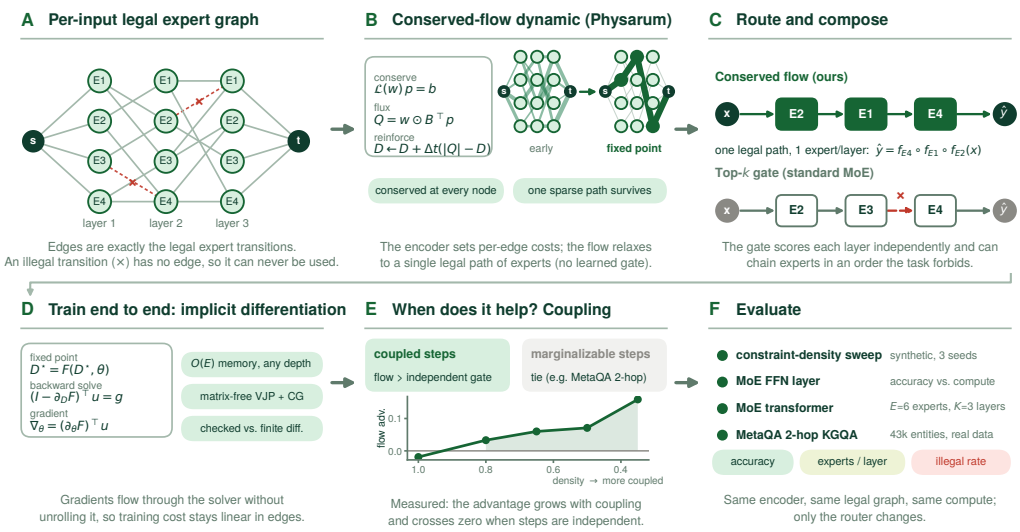

Figure 1: **Overview.** (A) Each input defines a graph containing only legal expert transitions, so an illegal transition has no edge. (B) A conserved Physarum flow over this graph relaxes to one sparse min-cost path. (C) The model applies the experts on that path in sequence, while a standard top-k gate chooses each layer independently and can therefore form an illegal sequence. (D) The flow is trained end to end by implicit differentiation through its fixed point. (E) The coupling condition: the flow beats an independent gate when the steps constrain one another and ties it when the output marginalizes over them (inset: measured advantage vs. legal-transition density, Fig. 3). (F) Testbeds and metrics.

- **A trainable MoE transformer, reported seed by seed**: on a coupled task, the flow router reduces the illegal-transition rate by more than an order of magnitude relative to a top-1 gate on every seed. Its accuracy is the best of the tested routers on some seeds, but it varies substantially across seeds and is below the gate on average (Fig. 5).
- **Hard, input-conditional legality by topology**: the router's flux is zero on every illegal transition, a guarantee rather than a penalty (§2).
- **A negative result on MetaQA 2-hop**: on a marginalizable task, the flow does not beat a matched no-flow baseline. This is consistent with the proposed boundary condition.
- **Practical training**: we use implicit differentiation through the flow fixed point and a matrix-free solver, with the gradient checked against finite differences (§3).

The flow dynamic itself is not new; it is the well-studied Tero–Nakagaki model. Our contribution is to use it as a trainable expert router, establish the resulting legality guarantee, and identify the condition under which the added coupling helps.

2 Method

Graph and costs. A signed incidence $B \in \mathbb{R}^{N\times E}$ encodes the per-query legal graph. An encoder maps the input to per-edge logits z; edge cost $L_e = \mathrm{softplus}(-z_e) + \varepsilon$.

Conserved-flow dynamic. We use the Tero–Nakagaki model of *Physarum* transport networks (Nakagaki et al., 2000; Tero et al., 2007; 2010). With conductance D_e and weight $w_e = D_e/L_e$, flux $Q_e = w_e(B^\top p)_e$ obeys Kirchhoff conservation $\mathcal{L}(w)\,p = b$, where $\mathcal{L}(w) = B\,\mathrm{diag}(w)\,B^\top$ is the weighted Laplacian and b injects $+1$ at the source and -1 at the sink (grounded). Reinforcement $D_e \leftarrow D_e + \Delta t(|Q_e| - D_e)$ iterates to the fixed point $D_e^\star = |Q_e^\star|$, at which a single sparse path survives.

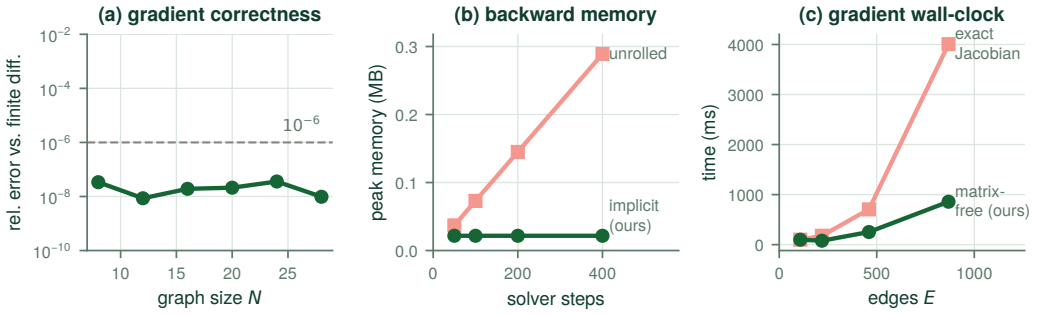

Figure 2: Differentiating and solving the flow at scale. (a) The implicit gradient matches a finite-difference gold standard to relative error $\sim 10^{-8}$ across graph sizes. (b) Backward memory is flat in solver depth for implicit differentiation (forward runs under `no_grad`) while unrolled autograd grows linearly. (c) The matrix-free (VJP-only) implicit gradient avoids the $O(E^2)$ Jacobian, so its wall-clock scales far better with edges.

Selection and composition. Node throughput $t_v = \frac{1}{2}\sum_e |B_{v,e}||Q_e|$; per position $a_k = t^\gamma / \sum t^\gamma$. Walking the path, data nodes inject a value and expert nodes apply an MLP, giving function *composition* (not a weighted average).

Two guarantees. $e \notin E \Rightarrow Q_e = 0$ (legality by topology); and $\sum_{\text{in}} Q = \sum_{\text{out}} Q$ (conservation coupling).

3 Scalability: Differentiating and Solving the Flow

The solve is one implicit layer $z^\star = F(z^\star, \theta)$ with $F(D) = |Q(D)|$, which we differentiate as in deep equilibrium models (Bai et al., 2019). By the implicit function theorem, for cotangent $g = \partial\mathcal{J}/\partial z^\star$ the parameter gradient is $\partial\mathcal{J}/\partial\theta = (\partial F/\partial\theta)^\top u$ with $(I - (\partial F/\partial z)^\top)u = g$, solved matrix-free (VJP-only). This requires $O(E)$ backward memory and does not grow with solver depth. We check the implicit gradient against finite differences (cosine 1.0, rel-err $\sim 10^{-8}$) and against the exact Jacobian. A small conductance floor regularizes the otherwise-degenerate hard fixed point. A matrix-free conjugate-gradient forward solve (Laplacian-vector products from the edge list, $O(E)$/iter) replaces the dense $O(N^3)$ solve, matching it to $\sim 10^{-3}$ and solving $N \approx 2400$ neighborhoods in ~ 0.4s.

4 Experiments

When does the flow help? We sweep the density of a $K{=}3$ transition graph (3 seeds): low density = many illegal transitions = steps strongly coupled; density 1.0 = every transition legal = unconstrained/marginalizable. We compare the conserved flow against an *independent per-step softmax* over the identical legal graph (same encoder, same candidates; only the conservation solve differs). Figure 3 shows a monotonic crossover: the flow's advantage is $+0.159$ MSE at the most coupled setting (density 0.35: flow 0.586 vs. independent 0.745) and falls to -0.019 when unconstrained (density 1.0: 0.509 vs. 0.490). In this sweep, the flow wins *iff* the steps are coupled.

MetaQA 2-hop: the predicted tie (real data). MetaQA (Zhang et al., 2018) KG: 43,234 entities, 9 relations, 134,741 triples; 2-hop questions, 8k train / 1.5k test, matched GRU encoder with per-hop query heads. A 2-hop answer marginalizes over the bridge entity, so the boundary predicts no benefit from coupling. Consistent with this prediction, the conserved flow ties a no-flow per-step softmax (Table 1). Both attain 0 hallucination by construction (the legal graph, not the flow, provides that).

As an MoE router (coupled regime). The same mechanism is a sparse expert router: the flow selects one legal *path of experts* and composes them sequentially. On a constrained sequential expert-composition task (density ≈ 0.6, 4 seeds), the conserved flow beats both standard top-k

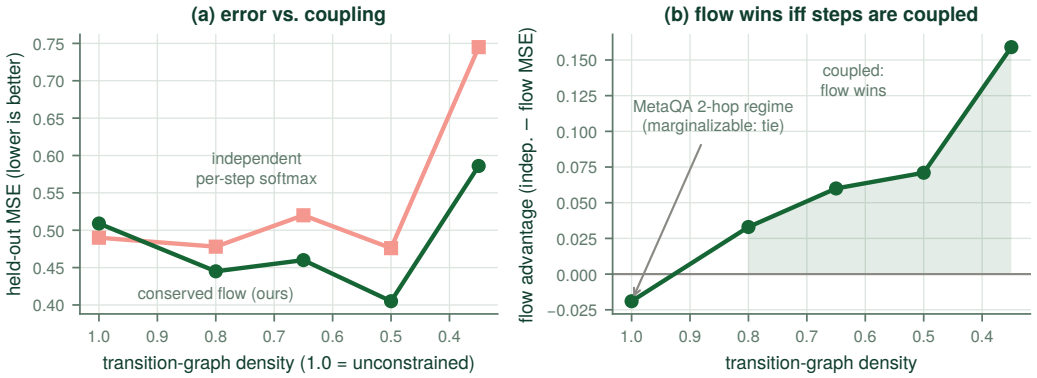

Figure 3: Conserved-flow routing outperforms independent per-step scoring when the reasoning steps are coupled. (a) Held-out MSE vs. constraint density (left = more coupled); (b) the flow's advantage crosses zero as the task becomes marginalizable. MetaQA 2-hop lies at the unconstrained end, where the sweep predicts the tie reported below.

Method	hits@1 (raw)	hits@1 (degree-norm)	hallucination
Conserved flow (ours)	0.864	0.948	0 (by construction)
No-flow ablation (same encoder/graph)	0.885	0.944	0 (by construction)

Table 1: MetaQA 2-hop. The conserved flow does not improve accuracy over a matched no-flow baseline on this marginalizable task, consistent with Fig. 3. The legality guarantee comes from the topology and holds for both. By contrast, embedding-based KGQA methods that score answers over *all* entities (e.g. EmbedKGQA (Saxena et al., 2020), built on ComplEx (Trouillon et al., 2016)) carry no such guarantee and can answer outside the legal set.

gating and Routing-Free MoE on both accuracy and legality (Table 2), which is the coupled regime where Fig. 3 predicts an advantage.

MoE FFN layer: accuracy and compute. We stack K MoE FFN layers over $E{=}8$ experts and compare the conserved-flow router against a hard top-k gate (standard MoE with load balancing) and a dense layer, in the coupled regime (3 seeds; Fig. 4). At the minimum compute setting of one active expert per layer, the flow router attains the best held-out MSE (1.170 ± 0.033), beating the top-1 gate at the same compute (1.294 ± 0.033) and the dense layer (1.278 ± 0.030, which activates all 8 experts). The dense model averages all expert outputs, while the flow instead composes the selected path. On legality, the flow's flux is supported only on legal expert transitions by construction; decoding its connected max-flux path is legal by construction, and even a per-layer argmax decode violates legality only 0.150 ± 0.010 of the time versus 0.899 for the top-1 gate. At the same sparsity budget, this gives lower error while keeping expert transitions within the hard constraint graph.

A real MoE transformer. We move from an isolated FFN block to a small transformer (token embedding, multi-head attention, and an $E{=}6$-expert MoE feed-forward block at each of $K{=}3$ layers) trained end-to-end on a coupled compositional sequence task, where the correct expert at a layer depends on the expert chosen at the previous layer (30k training sequences, 20 epochs, one run per seed). Because the flow's behavior varies across seeds, we show every seed (Fig. 5) as well as the summary (Table 3).

Legality is consistent across seeds. The flow's illegal-transition rate is 0.007 on average versus 0.336 for the top-1 gate, and never exceeds 0.034 on any seed. The residual nonzero rate is a decoding artifact: we read the route with a per-layer argmax of node throughput, which can pair experts from two different legal paths when the flow has not fully collapsed onto one; the flow's flux itself is zero on every illegal edge, and decoding the connected max-flux path is legal by construction.

Router	held-out MSE ↓	illegal-transition rate ↓
Top-k gate (standard MoE)	0.506 ± 0.031	0.295 ± 0.021
Routing-Free MoE	0.583 ± 0.024	0.262 ± 0.019
Conserved flow (ours)	$\mathbf{0.392 \pm 0.026}$	$\mathbf{0.103 \pm 0.007}$
oracle (upper bound)	0.001	0.000

Table 2: Constrained sequential expert routing, 4 seeds (MSE / illegal-transition rate at noise 1.0). The conserved-flow router beats top-k gating and Routing-Free MoE on both, in the coupled regime. Illegal-transition rate is 0.000 at lower noise for the flow.

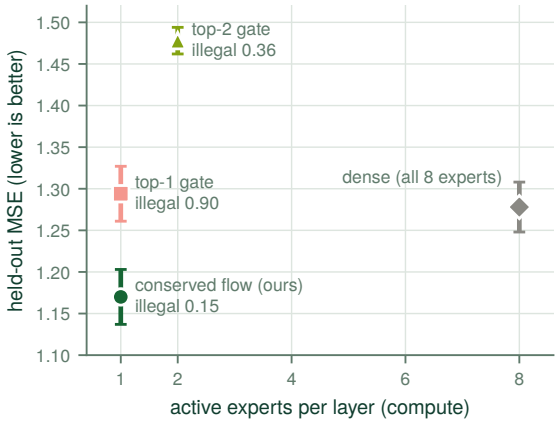

Figure 4: MoE accuracy/compute results in the coupled regime (3 seeds). Conserved-flow routing has the lowest error at one active expert per layer, while the top-k gate requires more active experts to reduce error and still produces illegal transitions (illegal-transition rate annotated).

Accuracy varies substantially across seeds. When training succeeds, the flow is the most accurate router we tested (up to 0.992, versus a top-1 mean of 0.847), and it beats the top-1 gate on 5 of 10 seeds. On the remaining seeds it trains to a worse solution, in one case failing entirely, so its mean (0.756; median 0.820) is below the gate's. One possible source is instability in the inner solve early in training, when edge costs are near-uniform and the reinforcement dynamic can settle on a poor path. Stabilizing this behavior (for example by annealing the conductance floor or the sharpening exponent γ) is an important remaining problem for this router. The gates are not stable at this scale either (top-1 falls to 0.45 on one seed and top-2 ranges from 0.11 to 0.99), and the dense layer, which averages all six experts, fails on this compositional task.

Shared-router legality baseline. A PathMoE-style *shared-router* baseline (router params tied across steps) attains the *worst* legality under noise (illegal-hop ≈ 0.5) versus the flow's ≈ 0: soft/param-sharing constraints can make legality worse, not merely fail to guarantee it. This result motivates using topology when expert transitions must satisfy a hard constraint.

5 Related Work

Sparse MoE routing. Standard MoE (Shazeer et al., 2017; Fedus et al., 2022) uses a learned top-k gate per layer, chosen *independently*; load-balancing losses keep experts utilized. PathMoE (Gu et al., 2026) constrains the path space by *tying* router parameters across positions (a soft, statistical constraint); Chain-of-Experts (Wang et al., 2025) chains experts sequentially within a layer with a learned router at each step; Routing-Free MoE (Liu et al., 2026) drops the gate for threshold self-activation. All route by learned scores with no hard guarantee on which expert may follow which. Our router is a conserved flow whose *topology* makes illegal expert transitions impossible, whose objective is a min-cost path (not a sampler), and which is differentiable and jointly trained; our

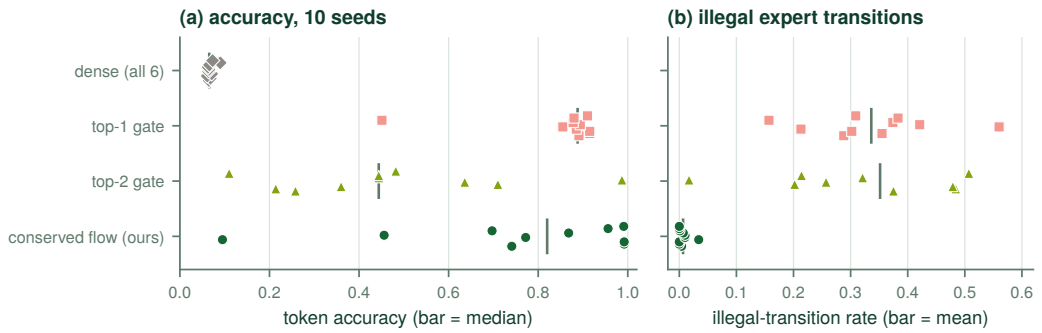

Figure 5: MoE transformer on a coupled compositional task, one point per seed. (a) Token accuracy: the flow router reaches the highest accuracy of any router on some seeds but is unstable across seeds. (b) Illegal expert transitions: the flow stays near zero on every seed while both top-k gates route illegally on a large fraction of inputs.

Router	token acc ↑	median acc	experts/layer	illegal-transition ↓
dense (all 6)	0.069 ± 0.008	0.066	6	—
top-1 gate	0.847 ± 0.133	0.888	1	0.336 ± 0.106
top-2 gate	0.464 ± 0.246	0.444	2	0.352 ± 0.178
conserved flow (ours)	0.756 ± 0.274	0.820	$1^{\dagger}$	0.007 ± 0.010

Table 3: MoE transformer, mean $\pm$ standard deviation over 10 seeds (per-seed values in Fig. 5). †The flow layer weights expert outputs by node throughput, which concentrates on one expert per layer at convergence; the top-k gates apply exactly k.

contribution over these is (i) the hard, input-conditional legality guarantee and (ii) the *condition* (coupling) under which such routing beats an independent gate.

Flow / graph reasoning. GFlowNets (Bengio et al., 2021) (flow *sampling*, not min-cost); PathRAG (Chen et al., 2026) (non-parametric flow pruning); D-RAG (Gao et al., 2025) (differentiable subgraph *set*); MoG (Yuan et al., 2026) (a topology-aware router over expert graphs); Physarum differentiable-LP layers (Meng et al., 2021) (LPs, not expert routing). These works already establish related outcomes, including PathMoE's path constraints, CoE's sequential composition, and GFlowNet's flow over a DAG. Our claim is narrower: the conserved-flow routing mechanism, its topology-based guarantee, and the coupling condition identified here.

6 Limitations

The flow's benefit is conditional. On marginalizable tasks it does not beat independent scoring, so the mechanism is only useful when the reasoning steps are coupled. The legality source is the KG (clean on KGQA, open where no native graph exists); γ and the conductance floor are tuned by hand. In the MoE transformer the flow's accuracy is unstable across seeds (Fig. 5), and its layer weights experts by throughput rather than applying exactly one, so a strict hard-path evaluation at equal compute is still needed. The coupled-regime improvement is shown on controlled synthetic constraints, while MetaQA gives the predicted tie. A real-data *coupled* instance (e.g. 3-hop, which our scalable solver makes tractable) remains an important experiment. Results are single-domain (movies); WebQSP/CWQ are needed for external validity.

7 Conclusion

A conserved-flow router commits a Mixture-of-Experts model to one globally consistent, min-cost path of experts, and its topology makes illegal expert transitions impossible by construction. We focused on *when* this kind of routing is useful and found a consistent boundary: when a task's output

marginalizes over its intermediate steps, the coupling buys nothing and an independent per-step gate suffices (MetaQA 2-hop, a tie); when the steps are mutually constraining, the flow's advantage over independent gating grows with the coupling. In a trainable MoE transformer the flow keeps illegal routing near zero on every seed and reaches the best accuracy of any router when it trains well, but its training is not yet reliable. Implicit differentiation and a matrix-free solver keep the training cost manageable. The practical takeaway is specific: coupled routing is useful when the steps constrain one another, while independent routing is sufficient when they do not. The topology-based legality guarantee is also useful in settings where an MoE must obey a hard constraint on which expert may follow which.

AI USE STATEMENT

The research direction and the central idea (using a conserved Physarum flow as an expert router, and asking when coupled routing helps) originated with the author. We used generative AI tools (an LLM-based coding assistant): to implement and debug the experiment code, to search for and verify related work, and to draft and edit the text of this paper. We reviewed all AI-assisted work. Every number reported here comes from code in our repository, the implicit-differentiation implementation is checked against finite differences, and every reference was checked against the published paper. We take responsibility for the final content of the paper, including the text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments use small models and public data (MetaQA; Zhang et al., 2018) and run on a single GPU or CPU. The method is specified in §2 and §3; for every table we report the number of seeds and mean $\pm$ standard deviation, and the implicit gradient is verified against finite differences (Fig. 2). Code for all experiments and figures will be released upon publication.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems*, 2021.

Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. PathRAG: Pruning graph-based retrieval augmented generation with relational paths. In *AAAI Conference on Artificial Intelligence*, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Guangze Gao, Zixuan Li, Chunfeng Yuan, Jiawei Li, Wu Jianzhuo, Yuehao Zhang, Xiaolong Jin, Bing Li, and Weiming Hu. D-RAG: Differentiable retrieval-augmented generation for knowledge graph question answering. In *Conference on Empirical Methods in Natural Language Processing*, pp. 35398–35417, 2025.

Zijin Gu, Tatiana Likhomanenko, Vimal Thilak, Jason Ramapuram, and Navdeep Jaitly. Path-constrained mixture-of-experts. *arXiv preprint arXiv:2603.18297*, 2026.

Yilun Liu, Jinru Han, Sikuan Yan, Volker Tresp, and Yunpu Ma. Routing-free mixture-of-experts. *arXiv preprint arXiv:2604.00801*, 2026.

Zihang Meng, Sathya N. Ravi, and Vikas Singh. Physarum powered differentiable linear programming layers and applications. In *AAAI Conference on Artificial Intelligence*, 2021.

Toshiyuki Nakagaki, Hiroyasu Yamada, and Ágota Tóth. Maze-solving by an amoeboid organism. *Nature*, 407:470, 2000.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Atsushi Tero, Ryo Kobayashi, and Toshiyuki Nakagaki. A mathematical model for adaptive transport network in path finding by true slime mold. *Journal of Theoretical Biology*, 244(4):553–564, 2007.

Atsushi Tero, Seiji Takagi, Tetsu Saigusa, Kentaro Ito, Dan P. Bebber, Mark D. Fricker, Kenji Yumiki, Ryo Kobayashi, and Toshiyuki Nakagaki. Rules for biologically inspired adaptive network design. *Science*, 327(5964):439–442, 2010.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, 2016.

Zihan Wang, Rui Pan, Jiarui Yao, Róbert Csordás, Linjie Li, Lu Yin, Jiajun Wu, Tong Zhang, Manling Li, and Shiwei Liu. Chain-of-experts: Unlocking the communication power of mixture-of-experts models. *arXiv preprint arXiv:2506.18945*, 2025.

Zheng Yuan, Chuang Zhou, Linhao Luo, Siyu An, Di Yin, Xing Sun, and Xiao Huang. MoG: Mixture of experts for graph-based retrieval-augmented generation. *arXiv preprint arXiv:2605.31010*, 2026.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *AAAI Conference on Artificial Intelligence*, 2018.