Exam Format

(1) **Present** briefly, (2) **critically evaluate**, (3) **place in larger context** (∼10–15 min) + discussion. Bridges **(Local) Search** + **ML/Transformers**.

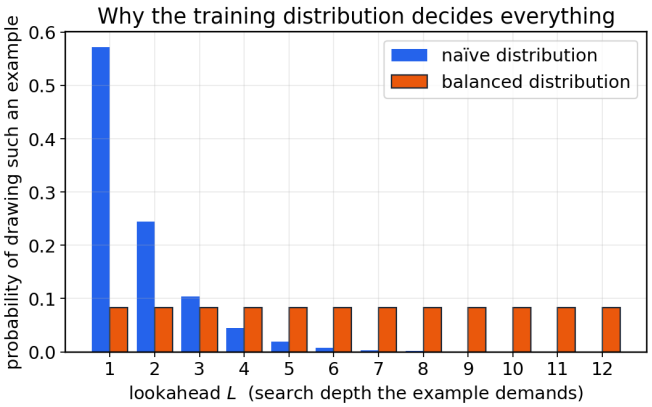
What the Paper Does

RQ: LLMs search poorly — (a) too little data, (b) too few params, or (c) *architectural limit*?
Testbed: graph connectivity on a **DAG**: given (edges, start, goal), predict the *next vertex*. Difficulty = **lookahead**

$$L = \min\{\,|P|,\ \max_i |S_i|\,\}\ (P = \text{path},\ S_i = \text{disjoint distractors}).$$

Min = stop at the cheaper of (A) reach goal in $|P|$, (B) exhaust distractors in $\max|S_i|$.
3 training distributions:

- **Naïve** (Erdős–Rényi): $P(L)$ *exponentially decays* (Fig 8) ⇒ shortcuts win.
- **Star**: k spokes; mid-ground.
- **Balanced**: uniform L + scale-free degrees → shortcuts engineered out. *Only here* the model learns search.



Architecture: tiny GPT-2-style, 1-hot token ⊕ position (*concatenated*); **no causal mask** (edges *shuffled* ⇒ needed edge can be anywhere). 6 L interp model (§3.1, Fig 5); 8 L for graph-size scaling (Fig 6). **Streaming/infinite** data ⇒ rules out "too little data".

Whether a transformer learns search depends on the training distribution, not just the model.

Learned Algorithm — Path-Merging

Reconstructed via **mechanistic interpretability** (activation patching, cost $\sim Ln^2mF$ ⇒ tiny models only).

- Each vertex token stores its **reachable set**.
- Each layer unions sets along edges ⇒ **doubles per layer** (parallel, all vertices simultaneously).
- $\sim \log_2 L$ layers suffice (parallel transitive closure).
- **Mechanism:** $Q_jK_i^\top$ on one-hots = *identity matching* (1 iff source ID = query's frontier slot). Two flavours: token-matching & position-based.
- **Non-maximal in practice** (Fig 5 bottom $\ll 1$): merges aren't always max ⇒ set grows slower than 2^ℓ ⇒ trained $L \le 12$ stretches only to $L = 13, 14$, then fails.

Doubling trace — chain $1 \to 2 \to \ldots \to 9$ (what vertex 1 knows it can reach):

After layer	reachable set of v. 1	max dist
input	$\{1\}$	0
layer 1	$\{1,2\}$	1
layer 2	$\{1,2,3\}$	2
layer 3	$\{1,2,3,4,5\}$	4
layer 4	$\{1,\ldots,9\}$	8 $\Rightarrow 2^\ell$

Doubling works because all vertices expand in parallel — vertex 1 merges with vertex 3 whose set is *already* $\{3,4,5\}$. Serial would grow by $+1$.

Key Results

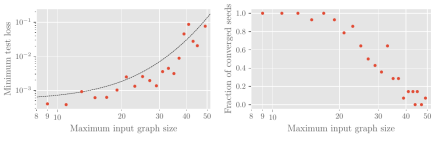
Bigger graphs (Fig 6; 8 L fixed, 14 seeds): converged fraction → 0; min test loss rises at *increasing* rate.
Bigger model (Fig 7; graph = 31, 0.9M→60.4M params): scaling **width** buys speed-to-shortcut basin (∼loss 1), *no ordering* to global min. "*Faster at being wrong.*" *Note:* Fig 7 scaled **width**; path-merging says **depth** is the search-range axis ⇒ depth-at-scale **untested**.
CoT (§6): **DFS = 3 layers**, **Selection-Inference = 4 layers** (constant suffices because search moves into trace). Still degrades w. graph size; scaling still doesn't fix it. SI: *selection easy, inference hard* on big graphs.
Fig 14 (DFS 47M) — the overfitting signature:

```
loss
  | train \__________  keeps DROPPING
0.4| test  \__      ___ rises back to pack (∼0.38)
  |            \_____/
0.15|                  DIPS (∼0.15) at ∼ 2×10^7 ex.
  +-------------------------→ #examples
```

train↓+test↑ = **distribution specialisation** (streaming ⇒ not memorisation; overfit to balanced mix vs. backtrack-8 test slice).
Robustness: decoder-only + **RoPE** (§A.5, Figs 11/12) → same failure. **NL proof search** (Fig 9, §3.1.2) → same failure, esp. in FLOPs ⇒ format isn't the cause (still tiny from-scratch models, *not* real LLMs).

3 misreadings to avoid: (i) "they tested LLMs" — no, ≤60M from-scratch. (ii) "can't represent search" — can (Merrill&Sabharwal); can't *learn* it. (iii) "scaling failed" — only *width* (d) at tiny scale; depth-at-scale untested.

Key numbers to memorise: interp model = **6 layers,** d=16, **input 128 tokens** (∼41 vertices, $L \le 20$); scaling fix-model = **8 layers**, vary graph 8→50, 14 seeds; scaling vary-model = graph 31, params **0.9M, 1.3M, 6.0M, 60.4M**, 4 seeds; CoT = DFS **3 L** (15 seeds), SI **4 L**; non-max merges ⇒ trained $L \le 12$ stretches to $L = 13, 14$ then fails.



Transformer Mini-Recap

Block = self-attention + feed-forward + residual + layernorm; stack N blocks. **Self-attention** per head:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\big(QK^\top/\sqrt{d_k}\big)V$$

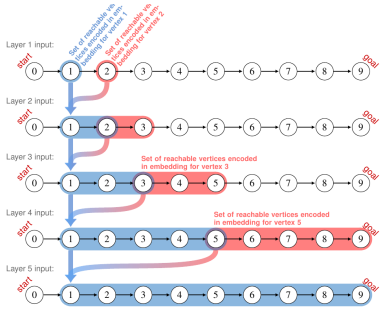
$Q = XW_Q$, $K = XW_K$, $V = XW_V$; X = token embedding + position. **Multi-head** = several attention ops in parallel, concatenated. **Position** is permutation-broken via positional encoding (absolute / relative / **RoPE** = rotation of Q/K encoding relative position). **Encoder** (BERT): bidirectional attention, MLM training. **Decoder** (GPT): *causal mask*, next-token training. This paper: GPT-2-style *without* causal mask (encoder-like).

Connection to Search

Task = classical (path-based), *not* local search.

	BFS/DFS/A*	Path-merging
Explore	sequential	all-vertices parallel
Steps	$\propto L$ (linear)	$\propto \log_2 L$
Correctness	provable	approximate (non-max)
Guarantee	yes	none

Training = local search: SGD ≡ hill climbing in weight space. Seed variance ≡ *random restarts*; shortcut solutions ≡ *local optima*; balanced data ≡ *landscape reshape*; curriculum ≡ *annealing*; looped transformers ≡ changed *move set*. The paper's negative results are loss-landscape phenomena, not capacity ones.



Critical Evaluation (Tier 1)

- **Learnability ≠ expressivity — title overreaches.** A *training* difficulty (huge seed variance, distribution sensitivity, failed convergence), not proof transformers can't *represent* search. **Merrill & Sabharwal (2024):** CoT-transformers = **Turing-complete**. Capacity is there; SGD doesn't find it.
- **Extrapolation gap**: ≤60M params, scaled **width**, depth held fixed. Path-merging says depth is the search-range axis ⇒ **depth-at-frontier-scale untested**. "60M doesn't help" ⇏ "10^{11} won't".
- **Data vs. architecture — emphasis, not contradiction.** Data binds small-scale; architecture may bind large-scale (Fig 6 even Balanced collapses); heavy distribution-sensitivity *is* a weak inductive bias. Honest critique: **title > evidence**, not logical contradiction.

Tier 2 / 3 Critiques

- **Equivalence to "reasoning" is asserted, not demonstrated.** Connectivity = proof search in *implicational propositional logic* only — no negation, disjunction, quantifiers, backtracking under uncertainty.
- **NL experiment** (Fig 9): still small transformers from scratch on synthetic English (*wumpus/vumpus*), not pretrained LLMs.
- **Interpretability cost** $\sim L\,n^2\,m\,F$ ⇒ method works only where it can run (tiny models); silent where the headline lives (frontier models). Self-undermining limit.
- **Authors' own hedge** (§7): scaling to much larger sizes *may* yield emergent search.

Fair: interpretability method + distribution-sensitivity findings are genuine, durable contributions.

What Reviewers Said

ICLR reviewers praised testbed + mech-interp but pushed on (i) limited situating w.r.t. chess/CoT/LLM-planning literature, (ii) gap to real-world LLMs, (iii) whether the negative scaling result is informative or just retreads the well-known shortcut/overfitting failure mode. Cite **Merrill & Sabharwal** as the strongest pre-existing tension.

Larger Context

- **Modern LLMs "can search"** (o1/o3, DeepSeek-R1, Gemini Deep Think, Claude extended thinking) — but by **externalising** search:
  - **Test-time compute / long CoT** (RL-trained reasoners).
  - **Explicit search**: Tree-of-Thoughts (DFS/BFS over partial solutions), MCTS; AlphaProof/AlphaGeometry → **IMO silver 2024**.
  - **Tool use** (SAT/SMT, code interp): the **neuro-symbolic** "LLM-modulo" (Kambhampati).
  - **Retrieval / agentic loops**: RAG, ReAct.

  ⇒ *confirms* the paper: single forward pass still doesn't robustly search.
- **Lecture bridge — Search:** compare path-merging to BFS/DFS/A* (completeness/optimality/serial vs. parallel). *Task* = classical search; *training* = local search.
- **Lecture bridge — ML:** shortcut learning = inductive bias / spurious correlations; seed variance = bias–variance + loss-landscape; balanced data = remove deceptive optima.

Oral-Exam Tactics

- *"Just tell me what the paper is about."* → 60-sec version: RQ → testbed (DAG ≡ proof search) → positive result (balanced learns) → algorithm (path-merging) → two negatives (scale, CoT). **Then stop**, leave hooks.
- *"Why graph connectivity?"* → lower bound on reasoning; *volunteer* the narrowness limit (signals maturity).
- *"Sketch the algorithm."* → chain $1 \to 9$, set doubles each layer, $\sim\log_2 L$, all-parallel; contrast w. BFS (serial, linear). Mention *non-maximal*.
- *"But ChatGPT can search now."* → *confirms* paper: externalised via CoT/MCTS/tools, not in the forward pass.
- *"What's your critique?"* → lead w. **learnability ≠ expressivity**; balance w. credit for mech-interp.
- *Rapid-fire defs:* **lookahead** $L = \min\{|P|, \max_i |S_i|\}$; **activation patching** = perturb input feature, re-run, watch dot-product/logit shift; **no causal mask** = edges shuffled, every token may need every other; **FLOPs** $\approx 2\cdot$params$\cdot$tokens.

Key References (one-line each)

- **Merrill & Sabharwal '24** — CoT-transformers Turing-complete ⇒ representability exists.
- **Sanford et al. '24** — graph connectivity needs $\sim\log n$ layers (matches path-merging theory).
- **Kambhampati '24** — LLMs can't plan; LLM-modulo (neuro-symbolic).
- **Creswell et al. '23** — selection-inference framework (the paper's §6.2 task).
- **Saparov & He '22** — LLMs are greedy reasoners; direct ancestor.
- **Schaeffer et al. '23** — "emergent abilities" as metric artefact; cuts against naive emergence-by-scale.

One-liner: *"The task is classical search but the real result is about local search — training is hill climbing trapped in shortcut optima unless data reshapes the landscape. Fixes (seeds, curriculum, looped transformers) are restarts and annealing. The title overreaches; evidence supports a calibrated learnability claim, not architectural impossibility."*