Public-data experiment
Release identity correction

The following eight pages are the original report for the separate public-data checkpoint. Its measurements are retained unchanged. References in that report to a 'PUBLIC RELEASE' or the '0.3.0 release' describe its intended packaging, not the published v0.3.0 checkpoint.

This experiment
Draft label: v0.3.0-public
Checkpoint SHA-256:
f64a77638758fe0da5a8400b5142a1b3a1a2f481b8fb7f8099f5fb88265b9992

Canonical v0.3.0 release
Checkpoint SHA-256:
ec83247177a6919c17944093fdd02d3258906dc809c16b326e0dfd879275efeb

The standard installation command downloads the canonical mixed-data checkpoint. It does not download this public-data experiment. The repository and both sets of weights require GitHub access; the public-data release remains a draft. 'Public-data' describes the training sources, not access to the weights.

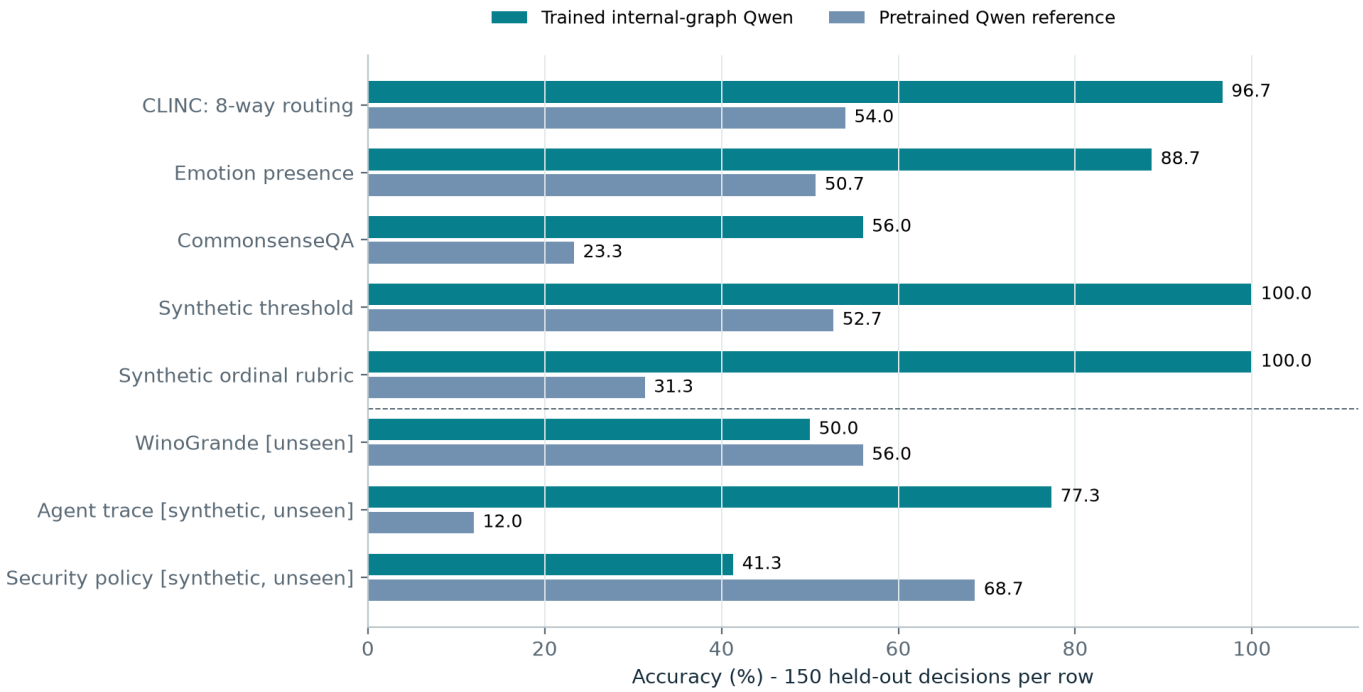
The report's reference to reproducing this run from the v0.3.0 source tag is also ambiguous. The canonical tag belongs to the other checkpoint. Identify this run by the fingerprint above and its own source snapshots, configs, and manifests. The generic README training command does not reproduce these exact weights.

Results and source data
openauditor.ai/public-run
Original eight-page report, without this correction

openauditor: training and evaluation

One Qwen3-0.6B checkpoint, internal graph modules, supervised learning and native RL, Version 0.3.0; data provenance is recorded in the run metadata.

Strongest where the training distribution is matched; weakest on unseen instruction regimes. Synthetic threshold reaches 100.0% and Synthetic ordinal rubric reaches 100.0%. The lowest row, Security policy, is at 41.3% against a pretrained reference of 68.7%. Training moved below the untouched reference on: WinoGrande, Security policy.



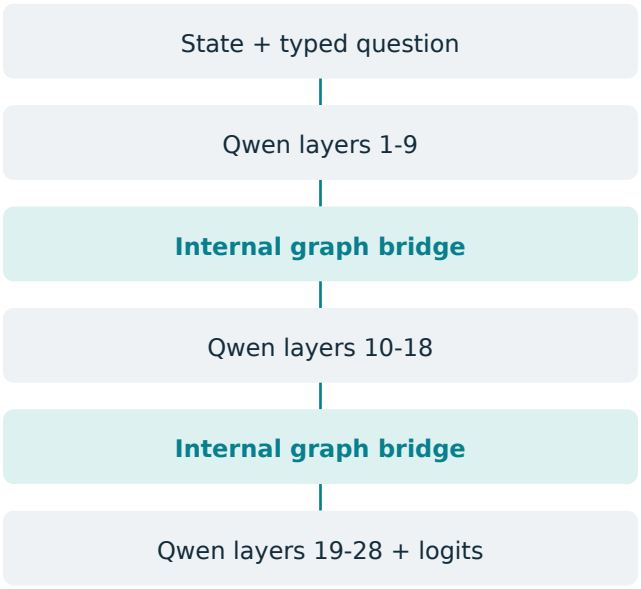
Reference: the same pinned pretrained Qwen3-0.6B, without our training, scored through the same candidate interface and calibrated on validation only. This is not Jev, a separately trained no-graph control, or an optimized prompting baseline.

What is ready	What is not established
Merged weights; BYO JSONL training; Choice, Score and Noul API; local serving; resume files; public-data recipe.	Jev parity, broad instruction generalization, graph message-passing advantage, calibration on new domains.

CLINC is an 8-option subset task, not full 151-class intent detection. Synthetic scores do not imply real-world reasoning. Agent-trace accuracy (77.3%) is a synthetic held-out family, not real agent traffic. Small point estimates are not significance claims.

One computation, one data format

The graph is embedded in Qwen's forward pass. It is not a second model that rescales the LLM's completed answer.



Inside each bridge
16 learned latent nodes, width 128.
Read prompt representations.
Learn top-4 incoming neighbors.
Run 2 message-passing rounds.
Write into the final decision token.

Later Qwen layers consume the graph-modified hidden state. These slots are not verified knowledge-graph entities.

11,309,824 jointly trainable parameters of 608,577,024. Rank-16 LoRA modifies the language projections while both graph bridges train. The base is BF16 and graph math is FP32. Export merges adapters into one weight file.

Data used by this run

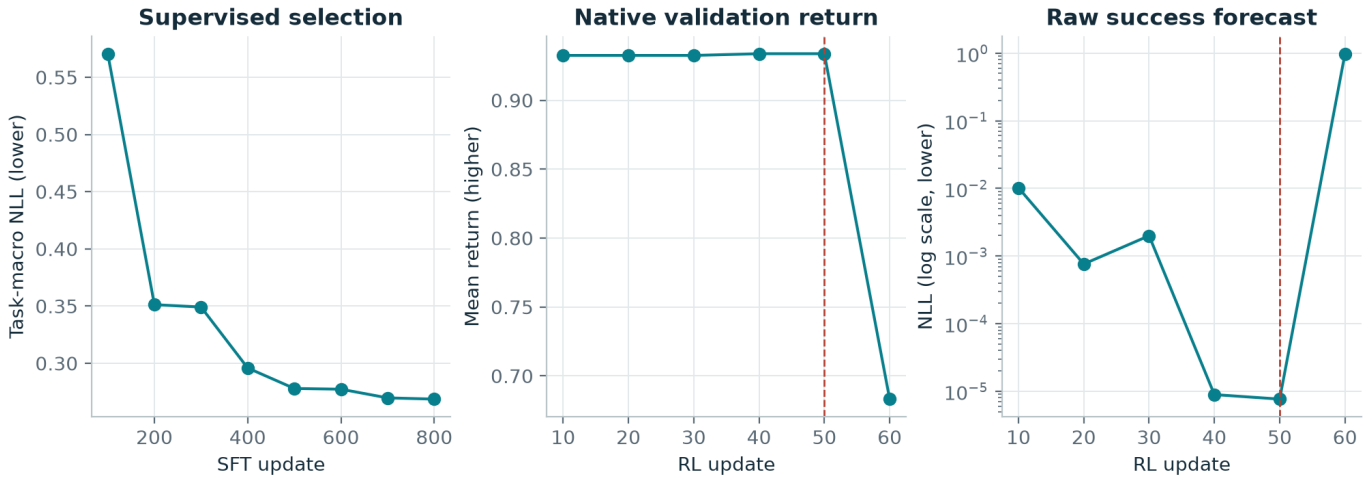
Partition	Decisions	Boundary
Train prepared	15,292	CLINC, GoEmotions, CommonsenseQA, synthetic rubric/threshold, verifier-labelled route and workflow episodes
Validation	967	483 selection / 484 calibration; disjoint groups
Test	966	7 task families, untouched
Unseen tasks / larger graphs	733	WinoGrande, synthetic agent trace and security policy, 9-node environments

The supervised stage consumed **12,800 decision draws** (800 updates x 16), not all 15,292 prepared decisions. Neither Doom nor Wikipedia was included in training. No session, chat or transcript data was used.

Upstream datasets keep their own licenses and attribution: CLINC150 (CC-BY-3.0), GoEmotions (Apache-2.0), CommonsenseQA (MIT), WinoGrande (CC-BY, held out). Pretraining exposure by Qwen is unknown, so held out means absent from this fine-tune.

What actually trained

800 supervised updates, then 60 native-RL updates. Validation selected RL update 50; the deployed checkpoint was then frozen.



Selection curves above use validation only. Dashed lines mark RL update 50. The forecast curve is uncalibrated NLL on native episodes; it is not test performance.

Real environment interaction; exploration signal measured, not assumed

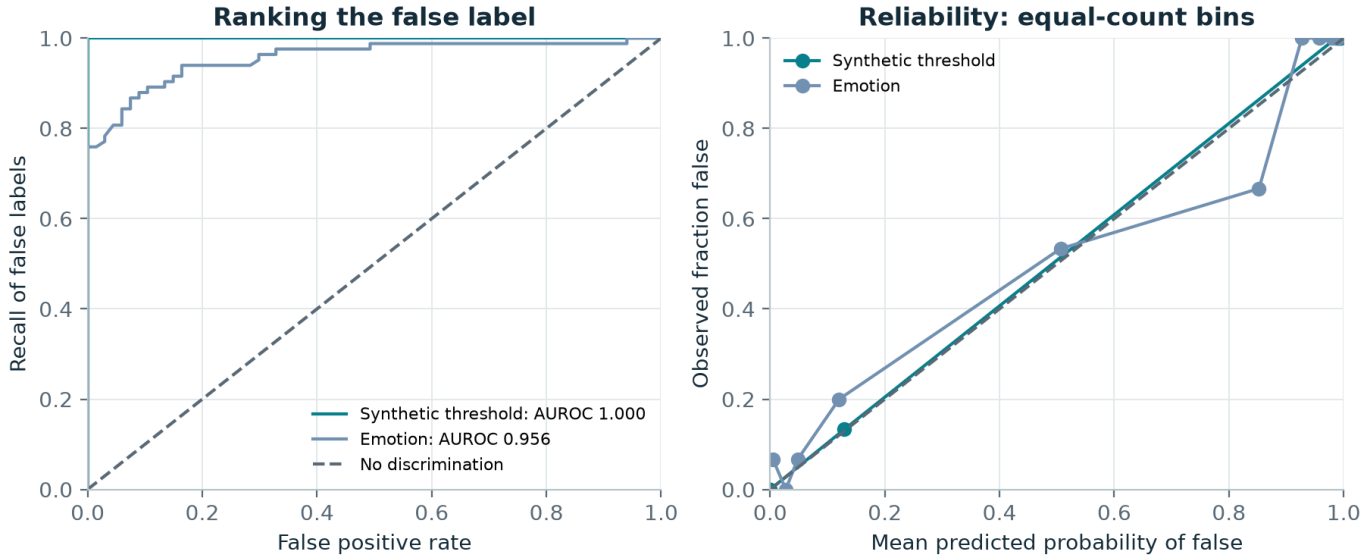
Four independent rollouts of each route/workflow task produce REINFORCE leave-one-out advantages. **0 of 60 updates** had a nonzero policy-gradient term; the others had equal rollout returns. Forecast and supervised-rehearsal losses still trained both language and graph weights.

Deployed-function selection	SFT candidate	RL candidate
Calibrated task-macro NLL	0.2696	0.2949
Allowed NLL degradation	Reference	+0.05 (passed by +0.0253)
Native validation return	Not a matched RL-benefit test	0.9338 at update 50
Deployed stage	-	environment rl

Choice / Score / Noul temperatures: 3.314 / 0.050 / 3.314, fitted on calibration groups only. This is an open RL/calibration method inspired by RLCD goals, not TypeSafe's undisclosed algorithm. The run does not isolate a causal benefit from RL or demonstrate convergence.

Are Noul probabilities usable?

Two binary test families, 300 decisions, scored against a constant training-prevalence predictor.



Reliability uses equal-count bins, distinct from the 15 equal-width confidence bins used for ECE. No confidence intervals are asserted. The test set did not fit a threshold.

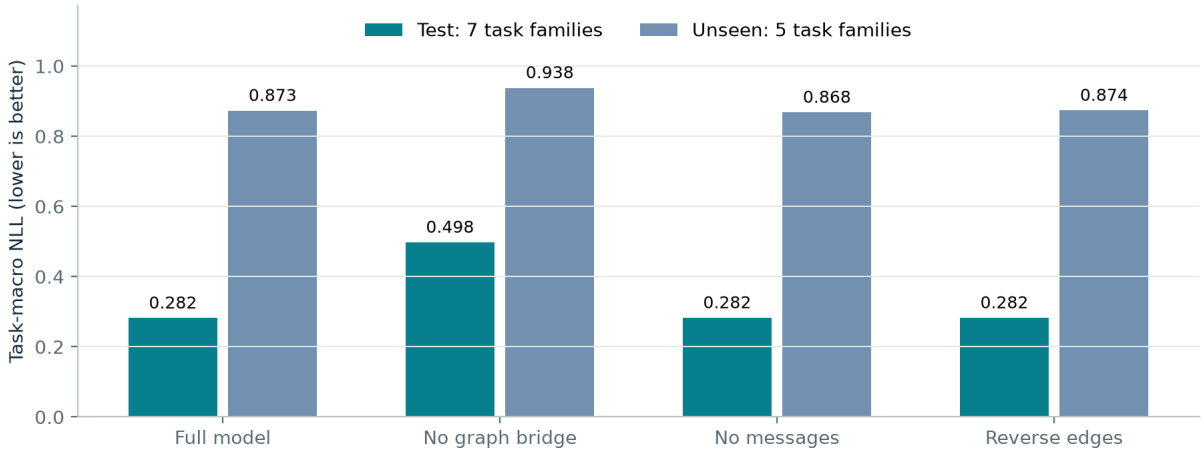
Metric	Synthetic threshold	Emotion	Constant baseline
AUROC for the false label (higher)	1.000	0.956	0.500
Average precision (higher)	1.000	0.971	prevalence
NLL (lower)	0.005	0.290	0.694 / 0.692
Brier, 2 classes (lower)	0.000	0.166	0.500 / 0.499
Balanced accuracy at 0.5	1.000	0.886	0.500
Accuracy at 0.5	100.0%	88.7%	48.7% / 55.3%

Proper scoring against a constant predictor is the honest test. The trained model beats the constant baseline on NLL for Synthetic threshold and Emotion. ECE alone does not measure discrimination; a constant predictor can have low ECE.

These probabilities describe these test distributions. Choice/Score confidence is entropy concentration, not probability that an answer is correct. Calibrate and choose thresholds on new validation data before using a model to trigger actions.

Does message passing earn its keep?

All interventions use the same co-trained checkpoint, frozen temperatures and evaluation seeds.



Removing the graph bridge changes test macro NLL by +0.216; the largest change is +0.216 for no graph. A degradation this size shows the trained model relies on the bridge, which is necessary but not sufficient for a graph benefit claim.

Native task success: greedy API policy

Intervention	6-node route	6-node workflow	9-node route	9-node workflow
Full	16/16	16/16	15/16	10/16
No graph	16/16	14/16	16/16	3/16
No messages	16/16	16/16	15/16	9/16
Reverse edges	16/16	16/16	15/16	10/16

The full stochastic policy solves 16/16 route and 16/16 workflow test episodes; on larger graphs it solves 15/16 route and 10/16 workflow episodes. Success forecasts are trained for that stochastic policy, not the greedy policy above.

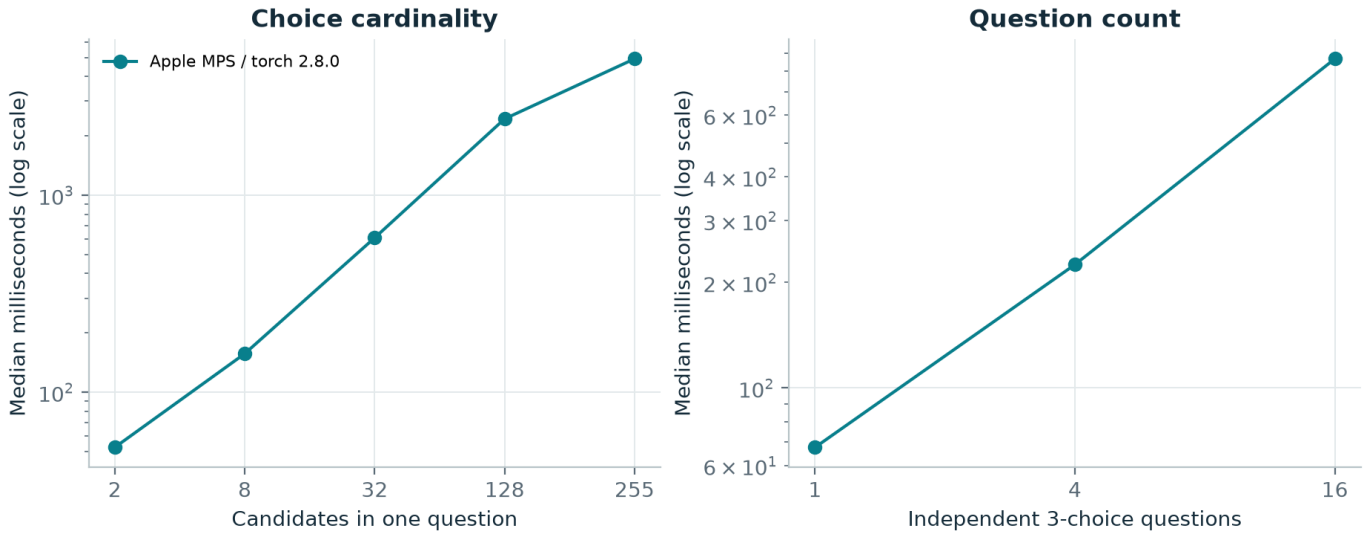
Counterfactual result, with the important control

Full: 16/16 paired dependency-chain tests solved, prediction changed with the edges in 100.0%. No graph: 16/16 paired dependency-chain tests solved, prediction changed with the edges in 100.0%. No messages: 16/16 paired dependency-chain tests solved, prediction changed with the edges in 100.0%. Reverse edges: 16/16 paired dependency-chain tests solved, prediction changed with the edges in 100.0%.

Ablations test reliance within a co-adapted model. They are not a controlled comparison against a separately trained, equal-budget no-graph architecture. A convincing follow-up needs multiple training seeds, a matched control, and harder fresh relational holdouts.

Fast decisions; two transfer demos

One frozen checkpoint on Apple MPS, plus structured-state Doom and real Wikipedia links.



Warm Python API medians, 3 repeats per case, including tokenization and normalization; model loading, HTTP and network time excluded. Logarithmic y-axes. This is descriptive, not a controlled hardware comparison, and not comparable to vendor API timings.

Probe	Observed result	Interpretation
1 question / 3 choices	67.7 ms	Local decision latency
255 choices	4,919.0 ms	Repeated state encoding becomes costly
Doom: eliminate enemy	0/3 kills; no-shot 0/3	Limited task success
Doom: do not fire	0/3 compliant	Instruction-following failure
Wikipedia	0/2 successful; Bicycle to Physics: cycle, Cat to CS: cycle	No reliable navigation demonstrated

Doom uses visible-enemy engine coordinates, not pixels. The control rate is 8.75 decisions per game-second (frame skip 4); recorded videos use game time. Both missions used the same seeds and no game-specific action override. Decisions took about 212 ms. Navigation selected actual outgoing links (up to 255), with no oracle fallback.

Question permutation plus an unrelated extra question caused a maximum probability drift of 0.0000. Independent candidate prompts repeat state computation; the implementation does not provide TypeSafe's claimed parallel-sampling architecture.

What we can and cannot match

These are capability-boundary checks, not a scored head-to-head. Jev access was unavailable and no paid calls were made.

Public Jev claim / demo [1,2]	Evidence from this run
Typed Choice, Score and Noul decisions	Same broad interface categories implemented and tested; not API-equivalence certification.
Calibrated decision probabilities / RLCD	Proper losses + native RL + validation temperatures. Calibration is domain-limited; algorithm equivalence is not claimed.
70-500 ms end-to-end API latency	67.7 ms to 4,919.0 ms warm local probes on Apple MPS. Network/cold-start excluded, so no speedup ratio is valid.
Structured-state Doom and instruction changes	0/3 kills; no-fire compliance 0/3.
Wikipedia link navigation; up to 255 choices	255 options supported; 0/2 tested routes reached the target.
Multi-question structured decisions	Mixed typed questions and probe independence pass; state processing is repeated per candidate.

The next experiment should target the failures

1. Generality before more parameters. Add varied instruction-conditioned tasks with counterexamples: obey vs refuse, engage vs hold, progress vs loop. Keep new test families untouched; today's failed examples are now diagnostic data, not fresh tests.

2. Make the graph hypothesis falsifiable. Use relational tasks that require multi-hop state tracking and compare an equal-budget no-graph training control. Retain the bridge only if repeatable gains justify its cost.

3. Measure Jev on the same contract. With access and budget approval, freeze prompts, candidate sets, labels and scoring rules before querying either system. Separate local compute latency from service end-to-end latency.

TypeSafe's published evaluation construction is not the same as our labeled-task accuracy. No numerical score in this report can be read as exceeding its published benchmarks. Schema-valid decisions can still be semantically wrong.

Artifacts and provenance

The 0.3.0 release contains the canonical model and this evidence. Every input is public or generated by the repository.

Artifact	Contents / purpose
openauditor-model.tar.gz	One merged model.safetensors, tokenizer, configs, Qwen license and provenance.
openauditor-training.tar.gz	SFT/RL best + last adapters, optimizer/RNG resume state, logs, source snapshots, data hashes and dataset manifest.
openauditor-evaluation.tar.gz	Frozen predictions (no input text), task/native/ablation metrics, benchmark metadata, Doom traces and navigation traces.
Report + charts + report_data.json	This PDF, PNG/SVG plots and aggregate source-derived values, with source-file fingerprints.
SHA256SUMS + artifact inventory	Archive checksums and per-file hashes. Verify before loading.

Canonical model-tree SHA-256:
f64a77638758fe0da5a8400b5142a1b3a1a2f481b8fb7f8099f5fb88265b9992

Backbone: Qwen/Qwen3-0.6B, revision c1899de289a04d12100db370d81485cdf75e47ca. Training device: Apple MPS, torch 2.8.0. The training, evaluation and demo commands are in README.md and TRAINING.md; the public data recipe is openauditor data, which pins every upstream dataset revision.

Use README.md for inference and BYO JSONL, TRAINING.md for preparation, resume, calibration and evaluation. Reproduction of this run needs only the pinned public datasets, the pinned Qwen revision and this repository at the release tag.

Release boundary: original code is MIT; preserve Qwen's Apache-2.0 terms and upstream dataset attribution. No session, chat or transcript data was used, so no privacy review of the weights is pending. No memorization or membership-inference audit is claimed.

Primary references

1. TypeSafe - Introducing System One Models and Jev (15 Sep 2026)
2. TypeSafe documentation - Introduction
3. Qwen3-0.6B model card and license
4. GreaseLM - prior language/graph fusion
5. TransNAR - prior transformer/neural-algorithmic fusion