MACROIC: A FOUNDATION MODEL FOR IN-CONTEXT MACROECONOMIC POLICY

Anonymous authors
Paper under double-blind review

ABSTRACT

Macroeconomic policy analysis relies on many structural models, each with its own preferences, frictions, and calibration. Each model must be solved or estimated separately, and none can adapt its decision rule to a new economy from a few observations, so knowledge about policy rules does not accumulate across models. We introduce MACROIC, a foundation policy model that addresses this gap by pretraining once on simulated trajectories from heterogeneous structural models and adapting through in-context learning. A shared vocabulary of economic roles allows one causal transformer to predict actions from an economy's history without a model identifier or test-time weight updates. Its advantage over explicit rule estimation is concentrated in short demonstrations: on held-out monetary models, including additional models evaluated after the training recipe was fixed, a short context gives the frozen policy higher prediction skill than local rule estimation, a learned-prior estimator, and a pretrained tabular regressor, while explicit estimation catches up with longer histories. Replacing the added structural trajectories with reduced-form simulations lowers short-context skill under the same continuation recipe. Separately, the shared policy retains low consumption-equivalent regret on a suite of planner economies. These results delimit the regime in which structural policy pretraining is effective: predicting unfamiliar decision rules from limited demonstrations while retaining precision on previously learned control tasks.

1 INTRODUCTION

Macroeconomic policy is analysed with many structural models rather than one. Central banks and researchers compare policy rules across model databases because no single specification is trusted for every question (Wieland et al., 2012; Christiano et al., 2018). Each model encodes its own preferences, frictions, and policy rules, and each calibration adds further variation (Smets & Wouters, 2007). Their decision rules are nevertheless solved or learned one model at a time. The cost of evaluating a rule therefore grows with the number of models, and what is learned from one model is not reused in the next, even though the models share much of their economic structure. Pretrained models have reduced this kind of cost in other fields where many related tasks share a common interface.

We ask whether trajectories from structural models can support such a reusable policy: can one pretrained network adapt its actions to an unfamiliar economy from a short demonstration, without updating its weights? The length of that demonstration changes the problem. With a long history, fitting a local decision rule can be effective. With only a few observations, as for a newly specified model or a new policy regime, the learner must also use information acquired from other economies. Structural simulation offers a way to learn this information: vary a model and its calibration, solve its source policy, and generate a trajectory. The resulting distribution contains both variation in decision rules and the economic dynamics under which those rules operate. Whether this distribution transfers better than simpler synthetic regression tasks is an empirical question.

We introduce MACROIC, a compact causal transformer trained on trajectories from heterogeneous structural models. A shared representation maps variables by economic function, such as consumption, capital, output, or the nominal interest rate, rather than by their model-specific names. At inference, the network reads the current observation and preceding state, action, and reward tokens.

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.

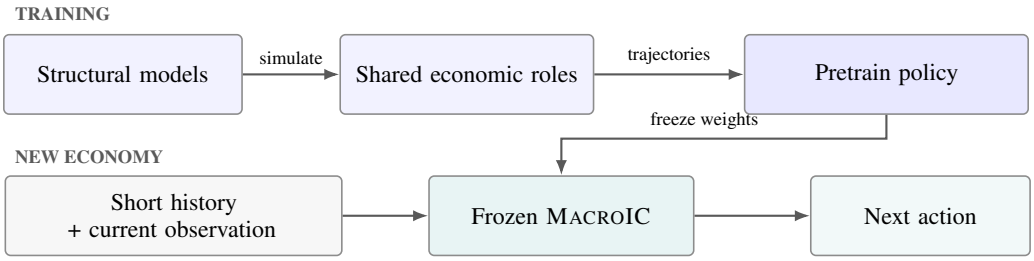

Figure 1: **Structural pretraining and in-context use.** Training trajectories are generated by numerical planner solutions or specified monetary rules, then mapped to shared economic roles. Core pretraining is followed by continuation with core replay. At test time, the same role mapping encodes a short history and the current observation; the policy predicts the next action with fixed weights. Transfer is evaluated by predicting source-rule actions on new monetary models. Planner rollouts provide a separate test of retained control precision.

It receives no explicit model identifier or calibration token. To broaden pretraining, we continue a policy trained on a core structural library using additional monetary models and replay of the core data (Figure 1).

The central empirical finding is a short-demonstration advantage over explicit rule estimation, including estimation with a prior learned from other economies and a pretrained tabular regressor. It holds on the ten-model development panel and repeats on eight further targets evaluated after the recipe was fixed; at full context the methods converge. Replacing the added structural corpus with reduced-form dynamics under the same recipe weakens short-context prediction.

Prediction and control require separate evidence. The monetary experiments measure imitation of specified interest-rate rules, not optimal monetary policy. On planner economies scored in welfare units through exact nonlinear transitions, the shared policy retains its precision after continuation; in an external simulator it improves on the source rule but trails a locally fitted rule.

The paper makes the following contributions.

- **Problem and system.** We treat structural-model trajectories as a pretraining distribution for a reusable policy and build MACROIC: a role-based interface aligning heterogeneous models in one observation–action space, a causal transformer predicting actions from a short demonstration with fixed weights, and a continuation recipe that broadens the library while replaying the core.
- **A demonstration-length result.** The frozen policy exceeds per-window ridge estimation, a learned-prior estimator, and a pretrained tabular regressor when the demonstration is short, and converges to them at full context. The result holds on the development panel and on eight further targets evaluated after the recipe was fixed.
- **Evidence on the pretraining distribution.** Replacing the added structural corpus with matched reduced-form or i.i.d. synthetic data under the same recipe reduces short-context transfer, so within the tested designs the gain traces to the structural trajectories rather than to generic dynamics or rule variety alone.
- **An evaluation separating prediction from control.** Exact-transition planner rollouts scored in consumption-equivalent units show retained precision after continuation, and an independent simulator with a full comparator set shows where in-context transfer carries over and where locally fitted rules remain stronger.

2 RELATED WORK

Synthetic pretraining and in-context learning. Prior-data fitted networks learn prediction procedures from tasks sampled by a data generator (Müller et al., 2022; Hollmann et al., 2023). TabPFN extends this approach to practical tabular prediction (Hollmann et al., 2025). Transformers can also learn simple function classes in context from synthetic examples (Garg et al., 2022). These results motivate learning from a distribution of economic decision problems. We test an external pretrained regressor and alternative synthetic training distributions.

In-context decision-making. Algorithm Distillation learns from the learning histories of reinforcement-learning agents (Laskin et al., 2023). Decision-pretrained transformers predict task-specific optimal actions from interaction data (Lee et al., 2023), while Decision Transformer models actions conditioned on returns (Chen et al., 2021). MACROIC instead imitates trajectories from fixed source policies. It is closer to in-context imitation, including Instant Policy's use of procedurally generated robot demonstrations (Vosylius & Johns, 2025). We study transfer across economic models through a shared role-based interface.

Learning with economic models. Neural methods approximate economic decision rules and equilibrium solutions (Maliar et al., 2021; Azinovic et al., 2022); economic simulators also support task-specific agent training (Zheng et al., 2022; Mi et al., 2024; 2025). DSGE-generated data have been used to train macroeconomic forecasters (Chib et al., 2026), and equilibrium world models learn solutions subject to equilibrium conditions (Schaab & Scheidegger, 2026). We study a different use of simulation: pretraining a shared action predictor that adapts through demonstrations, with closed-loop welfare evaluated separately on the planner suite.

3 STRUCTURAL POLICY PRETRAINING

Tasks and source trajectories. A task τ consists of a structural model, a calibration θ_τ, and an observation–action interface. Its source policy generates a trajectory $\mathcal{D}_\tau = ((o_t, a_t^{\mathrm{ref}}, r_t))_{t=1}^{T}$, where o_t contains the declared observations, not necessarily the full structural state. Actions include consumption and the nominal interest rate. Planner trajectories approximate optimizing decision rules; monetary trajectories follow specified interest-rate rules. Both supply demonstrations, but only the former support our planner-welfare comparison.

The core library contains 16 environments spanning growth, business-cycle, overlapping-generations, and monetary models, with independently sampled calibrations and second-order Dynare solutions (Adjemian et al., 2024; Schmitt-Grohé & Uribe, 2004). We add models from the Macroeconomic Model Data Base (Wieland et al., 2012), solved at first order. For these monetary models we vary rule coefficients and shock scales, not every structural parameter. Observations are lagged common aggregates: the nominal rate, inflation, the output gap, and output where available. The corpus and import transformations are documented in Appendix A.

A shared vocabulary of economic roles. Variables with the same economic function occupy the same role slots, even when their source names differ. Availability masks distinguish unobserved roles from observed zeros. Values are transformed to compatible deviation units and divided by role-specific scales. No target model contributes to the normalizer used for its evaluation. The core benchmark uses calibration-derived steady-state centering; a causal-centering control is reported in Appendix H. Masks can reveal aspects of the environment and are held fixed in within-episode comparisons.

In monetary models, the lagged rate can appear in the observation, while the next source-rule rate occupies the action role. A planner model instead predicts consumption from its declared stocks and shocks using the same token format. The interface aligns these quantities; it does not require the different source policies to optimize a common objective. A corpus audit rejects observations containing the contemporaneous target action (Appendix A).

A frozen in-context policy. Let L count visible observation positions, including the query. The preceding demonstration therefore contains $L-1$ labelled transitions. We write

$$\mathcal{H}_{t,L} = ((o_s, a_s, r_s))_{s=t-L+1}^{t-1}, \qquad \widehat{a}_t = f_\psi(\mathcal{H}_{t,L}, o_t; \iota_L), \tag{1}$$

where ι_L denotes the token positions used to encode the window. Observations, actions, and rewards are interleaved under causal attention; the query action and reward are never inputs. The network has six layers, width 256, eight heads, approximately five million parameters, and a maximum context of 128 temporal positions. It uses learned absolute-position and period-index embeddings. Truncating a window reindexes these embeddings, an important detail for interpreting context-length comparisons (Appendix E).

Training minimizes squared action-prediction error on active action slots, with an optional structural-residual term:

$$\mathcal{L}(\psi) = \mathcal{L}_{\text{act}}(\psi) + \lambda \mathcal{L}_{\text{struct}}(\psi). \tag{2}$$

The second term evaluates tagged source-model conditions while substituting the predicted action into the recorded trajectory. It applies only where those conditions are available. Imported monetary models contribute the action loss alone. The standard recipe uses $\lambda = 10$; the core context advantage persists at $\lambda = 0$ (Appendix C). At inference, new observations and actions update the context, not ψ.

Continuation with replay. Core pretraining and continuation each use $30{,}000$ optimization steps. Continuation starts from a core checkpoint, uses a learning rate of 10^{-4}, and samples from

$$P_{\text{cont}} = p_{\text{old}} P_{\text{core}} + (1 - p_{\text{old}}) P_{\text{added}}. \tag{3}$$

Within each selected group, the sampler selects an environment and a trajectory window. Core+8-C adds eight monetary models with $p_{\text{old}} = 0.5$. Core-continued spends the same second-stage budget on the core corpus alone. Wide-75 adds all 87 eligible monetary models with $p_{\text{old}} = 0.75$, giving 103 training environments. Additional recipe and budget comparisons are in Appendix I.

4 Evaluation

Development and prospective monetary tests. The development panel contains ten monetary targets and excludes them and ten declared close variants from training and normalization. It was used to select corpus composition and replay share. We then fixed the recipe and sampled eight additional targets from models outside the development exclusions and corpus C, without using their target prediction scores. Core+8-C and Core-continued were retrained with these eight targets also excluded from normalization. Their training support and core donors already excluded the new targets. We call this the *prospective test*. It tests the fixed recipe on new evaluation targets, not on necessarily new economic mechanisms. Three targets had appeared as training data in earlier alternative-corpus pilots; the full selection history is disclosed in Appendix B.

All monetary tests predict the source policy's next action from preceding source-generated observations and actions. For target e and context length L, query-only skill is

$$S_e(L) = 1 - \frac{\sum_{w \in \mathcal{W}_e} (\widehat{a}_{w,L} - a_w^{\text{ref}})^2}{\sum_{w \in \mathcal{W}_e} (a_w^- - a_w^{\text{ref}})^2}, \qquad \overline{S}(L) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} S_e(L), \tag{4}$$

where a_w^- is the last observed action before the fixed query in window w. Positive skill improves on persistence. Query times, targets, and persistence denominators are fixed across L; truncation changes both history and its position encoding. The main comparisons use this query-only score. Earlier breadth sweeps score all teacher-forced positions and are labelled separately in Appendix D.

Comparators and uncertainty. Local ridge fits a rule to the $L - 1$ observed state–action pairs, using the four lagged roles and an unpenalized intercept. Learned-prior estimation first fits rules on non-target training windows and estimates their mean. It then regularizes each new window's fit toward that mean, with strength selected separately for each context length outside the evaluated targets. Both the prior and its tuning are recomputed with the prospective targets excluded. This comparator tests whether pretraining adds to a learned mean rule. TabPFN-v2 (Hollmann et al., 2025) receives the same labelled pairs and query features as the regressions, using public regressor weights and an ensemble of eight, without target-panel tuning. MacroIC additionally uses the sequence representation and its declared reward channel; the comparison is not an architecture-matched ablation. We report means over targets and training runs, with 95% paired-bootstrap intervals for selected differences. The bootstrap resamples windows within targets and is conditional on the fixed target lists. Five continuation runs use four distinct core donors. Exact protocols and complete curves are in Appendix D.

Planner welfare. On seven planner economies, the policy drives exact nonlinear transitions after a 64-period source warm-up. The calibration's numerical reference rule is rolled through the same transitions from the same inherited stocks under the same shocks, so policy error is not confounded

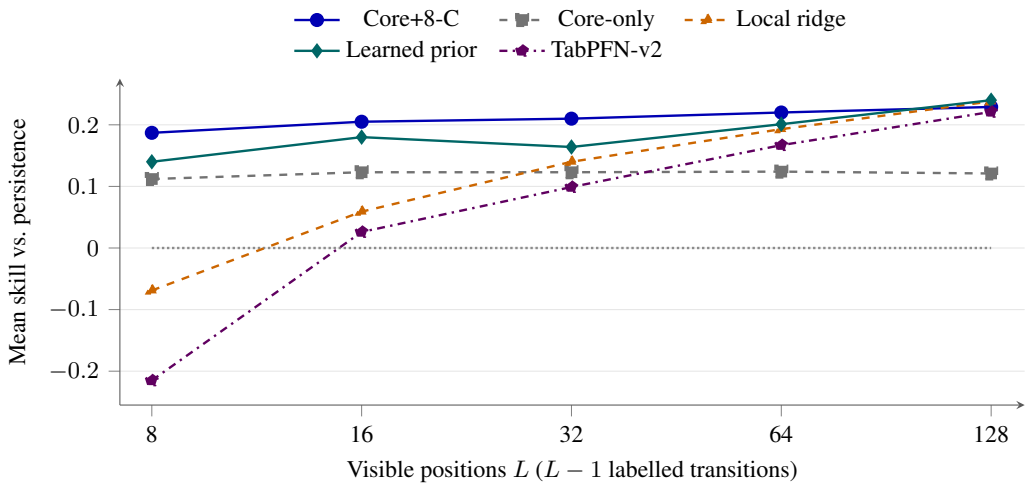

Figure 2: **Prediction on eight prospective targets.** With short demonstrations, Core+8-C outperforms local ridge, learned-prior estimation, and TabPFN-v2. Explicit estimation catches up at full context. The core-only continuation remains weaker throughout. Values are mean query-only skill over targets; network curves average five continuation runs. Paired intervals and exact values are reported in Tables 10 and 11.

with the perturbation solution's transition error. The reference is attainable but numerically approximate.

Let x_t^π be the endogenous state and u_t^π the task-specific flow utility under policy π, with takeover indexed by $t = 0$. For horizon $H = 40$ and common reference-terminal prices $\widehat{q}$, the score difference is

$$\widehat{J}(\pi) - \widehat{J}(\mathrm{ref}) = \sum_{t=0}^{H-1} \beta^t \left[u_t^\pi - u_t^{\mathrm{ref}}\right] + \beta^H \widehat{q}^\top (x_H^\pi - x_H^{\mathrm{ref}}). \tag{5}$$

The evaluator converts this difference to a consumption-equivalent loss in parts per million (Lucas, 1987). Positive regret indicates lower welfare; a negative value does not certify improvement over the true optimum. Reported medians first aggregate episodes within a run and then aggregate runs. Replaying the reference actions gives zero regret. Terminal prices, numerical conversion, and continuation checks are documented in Appendix F.

Subsistence and internal habits make preference-dependent policy differences more visible. Their utility gaps are $g_t = c_t - \bar{c}$ and $g_t = c_t - hc_{t-1}$, respectively. At fixed technology and discounting, the interior deterministic steady-state capital and consumption levels do not identify these preference parameters (Proposition 1). Thus, the default centering reference does not itself reveal them. Every comparator observes the full declared Markov state, including lagged consumption under habits. This distinguishes the value of a demonstration from merely supplying a missing habit-state variable.

5 Results

5.1 Transfer as a function of demonstration length

On the prospective test, Core+8-C achieves mean skill 0.187 at $L = 8$, compared with -0.069 for local ridge, 0.140 for learned-prior estimation, and -0.215 for TabPFN-v2 (Figure 2). It improves on persistence on all eight targets at this length. The gap over the strongest comparator, learned-prior estimation, is 0.047 $[0.022, 0.075]$ at $L = 8$ and 0.046 $[0.019, 0.076]$ at $L = 32$. At $L = 16$, the estimate is positive but its interval includes zero: 0.024 $[-0.001, 0.051]$, so the advantage is not resolved at every short length.

Table 1: **Development-panel query skill by context length.** Mean query-only skill on the ten targets used for recipe selection; higher is better. Each query receives $L - 1$ labelled transitions. Network columns average five continuation runs. A dash denotes an unreported TabPFN evaluation. The prospective test is shown separately in Figure 2.

L	Core+8-C	Local ridge	Learned prior	TabPFN-v2	Wide-75
8	0.132	−0.318	0.010	−0.388	0.097
16	0.134	−0.181	0.027	−0.245	0.121
32	0.144	−0.005	0.085	−0.007	0.136
64	0.150	0.135	0.147	–	0.137
128	0.164	0.182	0.190	0.160	0.139

Table 2: **Changing the added training data.** Mean query-only skill on the prospective set, three matched continuation runs per arm. The structural and reduced-form arms use the same core donors and training recipe. Marginal data scales are only approximately matched. The independent-observation arm and development-panel results are in Table 13.

L	Structural	Reduced-form dynamics	Paired difference [95% CI]
8	0.186	0.061	+0.125 [+0.092, +0.161]
16	0.205	0.134	+0.071 [+0.046, +0.099]
32	0.210	0.174	+0.036 [+0.014, +0.063]
64	0.222	0.196	+0.026 [+0.006, +0.047]
128	0.231	0.206	+0.026 [+0.001, +0.062]

With 128 positions, local estimation catches up: Core+8-C reaches 0.229, local ridge 0.238, learned-prior estimation 0.240, and TabPFN-v2 0.221. The paired intervals for all three comparisons include zero. Meanwhile, Core-continued achieves only 0.112 at $L = 8$ and 0.121 at $L = 128$. The prospective test therefore reproduces both the benefit of added monetary training data and the concentration of its relative advantage in short demonstrations.

The development panel shows the same pattern (Table 1). At $L = 8$, Core+8-C reaches 0.132, versus 0.010 for learned-prior estimation and −0.388 for TabPFN-v2. By $L = 128$, the learned-prior estimator is slightly ahead, 0.190 versus 0.164. The smaller short-context gap on the prospective set reflects the comparator's level: learned-prior estimation performs better there, so development-panel effect sizes should not be treated as typical of every new model. TabPFN, a general-purpose pretrained regressor, does not reproduce the short-context gain.

5.2 Dependence on the added pretraining distribution

We next keep the network, core donors, continuation budget, replay share, and normalizer fixed, but replace the eight added monetary models with synthetic alternatives. One alternative uses stable reduced-form linear dynamics with feedback from the sampled decision rule. Another applies the same rule family to independent observations. The alternatives match the interfaces, coefficient ranges, and episode counts, with approximate matching of scales and persistence.

At $L = 8$, the structural arm reaches 0.186 against 0.061 for reduced-form dynamics, a gap of 0.125 $[0.092, 0.161]$ (Table 2). At $L = 128$, the scores are 0.231 and 0.206. The development panel shows an even larger short-context contrast, 0.132 against −0.505. Reduced-form dynamics nevertheless raise full-context skill above the core-only continuation, so variation in rules and generic dynamics accounts for part of the gain; the structural corpus matters most when the demonstration is too short to compensate for a mismatched pretraining distribution.

This comparison isolates the *added monetary corpus*, not structural pretraining from scratch. It also does not perfectly match the data distributions: synthetic interest-rate standard deviations are 15–70% higher, and output scales differ for two models. With one generator design per arm, the result supports the chosen structural distribution over the tested replacements.

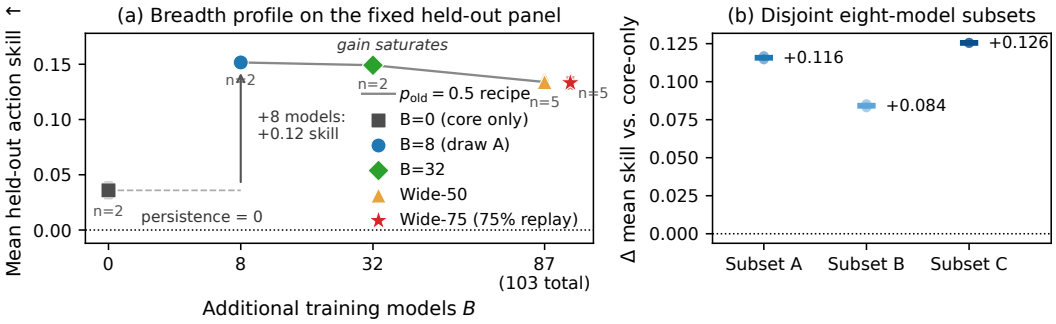

Figure 3: **Development-panel breadth and composition, whole-window skill.** Left: continuation with $p_{\text{old}} = 0.5$ as added breadth increases; core-only and Wide-75 have different group sampling and are marked separately. Right: all three tested disjoint eight-model subsets, with two matched runs each. These experiments selected corpus C; the prospective test evaluates the fixed recipe on additional targets.

Table 3: **Planner precision after continuation.** Consumption-equivalent regret in ppm, lower is better. Each entry is the median across five continuation runs of episode-median regret on the same seven-economy suite. These are the original checkpoint groups; the prospective test uses retrained groups with a revised normalizer.

Economy	Core-continued	Core+8-C	Wide-75
RBC baseline	0.6	0.8	1.3
Ramsey base	0.7	1.4	2.6
Ramsey CARA	1.5	2.0	3.2
Ramsey CRRA	0.7	1.8	3.3
Internal habits	0.9	2.0	4.2
Subsistence	0.6	4.0	11.4
Balanced growth	2.4	4.1	5.9

5.3 Corpus composition and context dependence

The gain is not confined to one added corpus. On the development panel, three disjoint eight-model subsets improve the matched two-run core-only continuation from 0.036 whole-window skill to 0.152, 0.120, and 0.161 (Figure 3); the full five-run Core-continued group averages 0.027 on the same metric (Table 14). These are all three tested subsets, disjoint blocks of one seeded shuffle rather than independent corpus draws. Retraining core-only and Core+8-C with normalizers fitted on their own training supports preserves the breadth gap (0.042 versus 0.161 in the matched two-run comparison), so the shared normalization pool is not required for the improvement. Adding 32 or 87 models does not further improve this panel under the tested budget. We therefore use Core+8-C as the principal model; Wide-75 is a breadth and retention check, not a superior predictor.

The policy also relies on the supplied demonstration. At fixed full length, Core+8-C's query skill drops from 0.164 to 0.019 when a single transition is repeated, and to -1.526 and -2.334 under shuffled histories and histories from another calibration of the same environment (Appendix E). These interventions preserve query observations and positions, unlike length truncation, and demonstrate reliance on contextual evidence and sensitivity to misleading histories.

5.4 A shared policy retains planner precision

Core+8-C retains median regret between 0.8 and 4.1 ppm across seven planner economies (Table 3). Wide-75 retains 4.2 ppm under habits and 11.4 ppm under subsistence despite its broader training support. Continuation therefore yields positive monetary transfer with low planner regret, although the core-only continuation remains more precise on these tasks (Appendix I).

Table 4: **Value of demonstrations on the core suite.** Median consumption-equivalent regret in ppm, lower is better. The three original core-trained comparators use five runs each and the same 200 episodes per economy. All observe the declared Markov state; only the parameter-informed model receives the calibration. These are not the continued checkpoints of Table 3. Ramsey base is a negative control.

Economy	Param.-informed	With context	No context
RBC baseline	0.2	0.9	32.6
Ramsey base (negative control)	0.2	1.9	0.9
Ramsey CARA	0.4	2.4	4.0
Ramsey CRRA	0.1	1.6	8.8
Balanced growth	1.1	3.4	6.7
Subsistence	2.3	6.3	66.7
Internal habits	−1.5	1.9	224.6

Table 5: **EconGym closed-loop comparators.** 24 paired test episodes after a common 32-step warm-up; higher is better. The network row uses continuation seed 0, and local ARX fits the 31 labelled warm-up transitions. The 20-step return uses zero reward after termination. PPO is a scene-trained reference (600k interaction steps), not a demonstration-matched baseline.

Policy	Reward/step	Total return	20-step return
Noisy source rule	0.388	7.90	7.88
Noise-free source rule	0.384	7.92	7.91
Core+8-C (frozen)	0.406	8.28	8.27
Local ARX (warm-up fit)	0.426	8.58	8.57
Hold last rate	0.392	8.25	8.22
Random rule draw	0.397	8.23	8.21
Scene-trained PPO	0.540	19.08	11.44

The controlled core experiment shows when demonstrations improve welfare (Table 4). Under habits, the context-using and no-context policies have median regret 1.9 and 224.6 ppm; under subsistence, 6.3 and 66.7 ppm. Retraining with reward tokens permanently removed gives 1.2 ppm under habits and 2.6 ppm under subsistence, so the measured benefit does not require the reward channel. A causal-centering pilot also preserves the reported context advantage. The base economy serves as a negative control where context has no median advantage (Appendix H). Replacing the terminal approximation with a long numerical continuation preserves the reported ordering on habits and subsistence (Table 16).

These results do not imply that pretraining beats task-specific fitting in control. In a matched-prefix test, local linear rules are competitive even with eight demonstrated transitions and have no failures in the tested configurations. Quadratic fits can fail under habits, but linear fits generally have lower 95th-percentile regret than the network (Table 17).

5.5 External transfer to an independent simulator

We next test EconGym (Mi et al., 2025), an independent simulator absent from training, normalization, and prior-fitting supports. Each method takes over after the same 32-step source-rule warm-up; other agents remain rule-based and the reserve ratio is fixed at 0.08. We report mean reward per surviving step, total post-takeover return, and a 20-step return with zero reward after termination; the latter two retain the effect of different survival times. Data splits and causal role transforms are in Appendix J.

After the shared warm-up, frozen Core+8-C obtains mean post-takeover return 8.28 (Table 5), compared with 7.90 for the noisy source rule and 7.92 for its noise-free version. The latter gap is 0.36 $[0.24, 0.46]$, so simply removing source-action noise does not reproduce the improvement. A local ARX rule fitted once on the same warm-up is stronger, with return 8.58 and a paired network-minus-ARX gap of -0.30 $[-0.44, -0.16]$. The advantage over holding the last rate is unresolved on total return, even though reward per surviving step is higher. The fixed 20-step return gives the same

ordering between the network, source rules, and ARX; the other four network runs give total returns of 8.25–8.39 (Appendix J).

Open-loop prediction likewise improves on persistence but trails ARX on the all-origins aggregate: at $L = 16$, their skills are 0.189 and 0.618. The network is more competitive away from the rate's lower bound, but this conditional result does not reverse the unfiltered comparison (Table 19). The external result therefore shows portability without scene-specific weight updates, not dominance over local adaptation.

6 Discussion and Limitations

The results locate a specific use for structural pretraining: interpreting a short source-policy demonstration when local estimation has limited data. They do not establish optimal policy discovery from arbitrary histories or adaptation to changing objectives. The monetary targets share broad New Keynesian structure, and the prospective test adds evaluation targets rather than proving their mechanisms absent from training.

Role availability and default core centering provide information beyond raw history. Learned position embeddings mean that context-length curves combine a change in history with a change in position encoding, and large corruption losses expose brittleness. Planner regret is measured against an approximate reference, and local rules remain strong control baselines. Finally, the external evidence is one simulator scene, not validation on historical economies. These limitations leave scope for broader training distributions and more robust adaptation.

7 Conclusion

MacroIC reuses trajectories from structural economic models to predict unfamiliar decision rules through context. Its advantage over explicit estimation is concentrated in short demonstrations: it repeats on additional targets and weakens when the added structural corpus is replaced by the tested reduced-form generator. The shared policy also retains low planner regret, but does not generally outperform local fitting in closed loop. Structural simulation is thus a useful pretraining distribution for in-context economic policy, with value that depends on the available history and the evaluation task.

Ethics statement

Experiments use synthetic trajectories and model-specific objectives; the learned policies are not recommendations for real monetary or fiscal decisions, and omitted mechanisms can affect behavior and welfare.

Reproducibility statement

All details needed to reproduce the experiments are given in the appendix: Appendices A to C specify the corpus, interfaces, selection history, and training recipes; Appendices D to F and J document evaluation protocols; Appendix K states compute costs and the scope of reproduction.

AI use statement

We used generative AI tools for coding assistance when implementing the experiment and evaluation code, to edit the manuscript for readability, and to identify related literature. The authors tested all AI-assisted code and checked all text and references against the sources. We take responsibility for the final content of this work.

REFERENCES

Stéphane Adjemian, Michel Juillard, Frédéric Karamé, Willi Mutschler, Johannes Pfeifer, Marco Ratto, Normann Rion, and Sébastien Villemot. Dynare: Reference manual, version 6. Dynare Working Papers 80, CEPREMAP, 2024.

Marlon Azinovic, Luca Gaegauf, and Simon Scheidegger. Deep equilibrium nets. *International Economic Review*, 63(4):1471–1525, 2022.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Siddhartha Chib, Fei Tan, and Zhixun Zhang. Learning the macroeconomic language. *arXiv preprint arXiv:2512.21031*, 2026. Version 3.

Lawrence J. Christiano, Martin S. Eichenbaum, and Mathias Trabandt. On DSGE models. *Journal of Economic Perspectives*, 32(3):113–140, 2018.

George M. Constantinides. Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy*, 98(3):519–543, 1990.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022.

Roy C. Geary. A note on "a constant-utility index of the cost of living". *Review of Economic Studies*, 18(1):65–66, 1950.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *International Conference on Learning Representations*, 2023.

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Robert E. Lucas. *Models of Business Cycles*. Basil Blackwell, Oxford, 1987.

Lilia Maliar, Serguei Maliar, and Pablo Winant. Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, 122:76–101, 2021.

Qirui Mi, Siyu Xia, Yan Song, Haifeng Zhang, Shenghao Zhu, and Jun Wang. TaxAI: A dynamic economic simulator and benchmark for multi-agent reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024.

Qirui Mi, Qipeng Yang, Zijun Fan, Wentian Fan, Heyang Ma, Chengdong Ma, Siyu Xia, Bo An, Jun Wang, and Haifeng Zhang. Econgym: A scalable ai testbed with diverse economic tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2025), Datasets and Benchmarks Track*, 2025.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Andreas Schaab and Simon Scheidegger. Equilibrium world models. *arXiv preprint arXiv:2606.23463*, 2026.

Stephanie Schmitt-Grohé and Martín Uribe. Solving dynamic general equilibrium models using a second-order approximation to the policy function. *Journal of Economic Dynamics and Control*, 28(4):755–775, 2004.

Frank Smets and Rafael Wouters. Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3):586–606, 2007.

Richard Stone. Linear expenditure systems and demand analysis: An application to the pattern of British demand. *The Economic Journal*, 64(255):511–527, 1954.

Vitalis Vosylius and Edward Johns. Instant policy: In-context imitation learning via graph diffusion. In *International Conference on Learning Representations*, 2025.

Volker Wieland, Tobias Cwik, Gernot J. Müller, Sebastian Schmidt, and Maik Wolters. A new comparative approach to macroeconomic modeling and policy analysis. *Journal of Economic Behavior & Organization*, 83(3):523–541, 2012.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, 2022.

A CORPUS AND INTERFACES

Counts. All environment counts in this paper are recomputed from the packed corpora and the training configurations stored with each checkpoint, not from experiment-log labels. The broad corpus contains 123 packed environments: the 16 core environments and 107 imported monetary models. The held-out panel removes 20 environments (ten targets and ten declared close variants; Appendix B), so the wide checkpoints train on 103 environments: 16 core plus 87 added monetary models.

Core library. The core corpus is generated from 16 Dynare environments spanning four model classes and thirteen specification families (Table 6): neoclassical growth under six preference specifications (log-utility base, CRRA per-capita, CARA, balanced growth, and the two welfare-sensitive variants of Appendix G), six RBC variants (baseline divisible labour, a CRRA labour-choice variant, capital-stock shock, news shocks, government spending, recursive preferences with stochastic volatility), an overlapping-generations economy, and three New Keynesian monetary environments with staggered wage setting, matching frictions, and open-economy credit frictions. Each environment carries an interface specification declaring observed states (with lags), actions, reward, parameter sampling ranges, and tagged optimality conditions. For each episode the calibration is drawn uniformly from the declared ranges, the model is solved by perturbation at order two (Schmitt-Grohé & Uribe, 2004; Adjemian et al., 2024), and 200 periods are simulated after a 100-period burn-in; episodes violating the declared utility domain are rejected. Generation is deterministic in the seed.

Representative sampling ranges. For the CRRA economy: capital share $\alpha \in [0.25, 0.40]$, discount factor $\beta \in [0.94, 0.99]$, depreciation $\delta \in [0.05, 0.15]$, risk aversion $\sigma \in [0.8, 3.0]$, population growth $n \in [0, 0.02]$, shock persistence $\rho_z \in [0.7, 0.95]$, innovation scale $\sigma_z \in [0.005, 0.025]$. The welfare-sensitive economies add the floor share $\phi \in [0, 0.6]$ (subsistence) and habit persistence $h \in [0.3, 0.8]$. Shock scales are sampled like any other parameter, so the model cannot assume a fixed noise level. The complete ranges for all environments ship with the specification files.

MMB import. The additional monetary models are imported from the Macroeconomic Model Data Base: each `.mod` file is patched for Dynare 6 compatibility, solved at first order, and simulated under its own shock processes. The generation protocol varies the coefficients of the specified interest-rate rule and the shock scales around the source specification; it does not sample every structural parameter of each imported model. The observation interface uses the lagged common aggregates (nominal rate, inflation, and output gap, with output included when the model declares it), mapped to the same roles as the core monetary environments. The decision variable is the nominal rate, following the model's specified rule; these actions are reference rules, not certified optima. Rewards use a quadratic central-bank stabilization loss. Models whose common variables could not be mapped or whose simulation failed the health checks were excluded at import time; the packed corpus lists exactly the environments that survived.

Importer audit. The imported environments are MMB-derived, not verbatim reproductions, and the automated transformations are the same for every model: (i) source files that load estimated policy parameters from binary archives instead receive one smoothed default rule, with the rule coefficients then sampled per episode from the declared ranges (smoothing $\in [0, 0.85]$, inflation response $\in [0.3, 2.0]$, output response $\in [0, 0.5]$); (ii) the database's common rule expression is simplified to lags and leads of at most one period, dropping deep-lag terms that the sampled rules set to zero anyway; (iii) simulation statements commented out in the source are re-enabled at first order; (iv) each model's shock scales receive one common rescaling factor chosen so that simulated quarterly inflation volatility matches the corpus convention; and (v) episodes whose sampled rule renders the model indeterminate are rejected at solve time. Of 126 candidate models put through this pipeline, 115 passed the simulation smoke test and 107 survived interface mapping and the health checks into the packed corpus.

Dataset cards. Table 6 documents each core environment's observation interface, action, objective, solver order, number of sampled quantities (parameters and shock scales), and packed episode count. Table 7 summarizes the imported monetary block by lineage category; all imported models are solved at first order, use the nominal rate as the action, and expose the same common aggregates.

Table 8 lists the ten held-out panel targets with the lineage category each represents and the declared close relatives excluded from all training and normalization supports. All three tables are generated directly from the environment specifications and packed data.

Table 6: **Core environments.** States are counted role-slots (with declared lags); "params" counts sampled structural parameters plus sampled shock scales; episodes sum the packed train/validation/test splits.

Environment	Family	States	Action	Objective	Order	Params	Episodes
Born_Pfeifer_2018_MP	new_keynesian	8	NominalInterestRate	quadratic stabilisation loss	2	17	25,000
Caldara_et_al_2012	rbc_recursive	3	Consumption	Cobb-Douglas log utility	2	11	19,685
Collard_2001	rbc_labor	3	Consumption	CRRA utility	2	9	21,335
Faia_2008	new_keynesian_labor	7	NominalInterestRate	inflation stabilisation loss	2	14	25,000
Faia_Monacelli_2008	new_keynesian_open	5	NominalInterestRate	CRRA utility, power labor disutility	2	15	24,977
OLG_stoch	olg	3	ConsYoung	CRRA utility	2	8	25,000
RBC_baseline_stoch	rbc	2	Consumption	CRRA utility	2	6	23,963
RBC_capital_stock_shock_stoch	rbc_labor	2	Consumption	CRRA utility, log leisure	2	7	25,000
RBC_news_shock_model_stoch	rbc_labor	10	Consumption	CRRA utility, log leisure	2	9	13,442
RBC_state_dependent_GIRF_household	rbc_government	3	Consumption	CRRA utility, log leisure	2	11	25,000
Ramsey_base_stoch	ramsey	2	Consumption	CRRA utility	2	5	25,000
Ramsey_cara_stoch	ramsey_cara	2	ConsumptionPerCapita	CARA utility	2	7	25,000
Ramsey_crra_stoch	ramsey	2	ConsumptionPerCapita	CRRA utility	2	7	25,000
Ramsey_habit_stoch	ramsey_habit	3	Consumption	internal-habit utility	2	7	22,200
Ramsey_subsistence_stoch	ramsey_subsistence	2	Consumption	Stone-Geary utility (subsistence floor)	2	7	25,000
Ramsey_upgrade_stoch	ramsey_growth	2	ConsumptionPerEffectiveLabor	CRRA utility	2	8	25,000

Table 7: **Imported monetary models by lineage category.** Every model observes the common aggregates (nominal rate, inflation, output gap, and output where the model declares it) and acts through its interest-rate rule.

Category	Models	Solver order	Episodes
Brazil	1	1	3,594
Canada	2	1	6,627
Chile	1	1	3,343
Hong Kong	2	1	6,974
United Kingdom	1	1	3,333
estimated United States	47	1	162,014
estimated euro area	15	1	47,008
multi-country	1	1	3,342
real business cycle	1	1	3,282
regional euro area	3	1	10,304
small calibrated New Keynesian	33	1	110,545

Table 8: **Held-out panel targets and their excluded relatives.** Relatives are model variants declared close enough to constitute leakage; they are removed from every training and normalization support.

Target	Category	Relatives excluded from training
BRA_SAMBA08	emerging-economy DSGE (Brazil)	none
CA_LS07	small open economy (Canada)	none
DEREA_GEAR16	two-region euro area (Germany/REA)	none
EA_QUEST3	large-scale euro area policy model	none
NK_BGG99	financial accelerator	mmb_us_vi16bgg, mmb_ea_vi16bgg
NK_GK09lin	banking / financial intermediation	mmb_us_vi16gk, mmb_ea_vi16gk
US_IAC05	housing and collateral	none
US_JPT11	investment-shock driven US cycle	none
US_LTW17	fiscal-monetary interaction	mmb_us_ltw17gz, mmb_us_ltw17nu, mmb_us_ltw17rot
US_SW07	canonical medium-scale US DSGE	mmb_us_dngs15_sw, mmb_us_dngs15_swsp, mmb_us_dngs15_swpi

Roles and audits. Every variable enters through the shared role vocabulary (consumption, capital, output, labour, nominal rate, inflation, output gap, productivity, ...), declared slots for each role and occurrence, with availability masks distinguishing unobserved roles from observed zeros. Two automated audits guard the corpus: a leakage audit rejecting any environment whose declared state contains the contemporaneous action, and a health check verifying that recorded paths satisfy each environment's declared identities. Variables enter as deviations from reference levels divided by one

constant per role, pooled by a median across environments so no single volatile environment sets the units; Appendix B documents which environments contribute to these statistics in each experiment.

B SPLITS, SELECTION, AND INFORMATION CONTRACTS

Development panel. The ten targets are US_SW07, NK_BGG99, NK_GK09lin, US_IAC05, US_LTW17, CA_LS07, DEREA_GEAR16, US_JPT11, BRA_SAMBA08, and EA_QUEST3. Ten declared close variants are excluded alongside them: the DNGS15 re-estimations of Smets–Wouters (SW, SWSP, SWπ), the VI16 implementations of BGG and GK (US and EA), and the LTW17 variants (gz, nu, rot). These exclusions apply to policy training, normalization, and donors. The donors contain no imported MMB environments. The exclusion concerns the declared models and variants, not every mechanism they contain.

The panel was used to compare corpus compositions A, B, and C and to select the replay share of the wide model. All tested composition results are reported in Figure 3; additional seeds of C do not turn this panel into an independent test. We therefore distinguish it throughout from the prospective evaluation.

Prospective targets. After fixing the Core+8-C recipe, a seeded shuffle (seed 2026) selected eight targets from 79 candidates outside the development exclusions and the eight models in corpus C: EA_GE10, EA_SR07, NK_CK08, NK_PP17, NK_RA16, NK_RW06, US_FGKR15, and US_RA07. Selection used no prediction scores on these targets. Core+8-C and Core-continued were retrained with the same donors and continuation settings, excluding the eight targets from normalization as well as policy training. Baseline priors and tuning pools also exclude these targets. Wide-75 is not evaluated on this set, since its original training support includes the additional models.

This is a prospective test of a fixed recipe, not a claim that the selected models were absent from every earlier experiment. EA_SR07 appeared in pilot corpus A, and EA_GE10 and NK_RW06 in pilot corpus B. Those pilots were compared on the ten development targets; prediction scores on the new eight-target set were not used in that comparison. We also do not claim a comprehensive new-lineage exclusion beyond the specified target models. The prospective checkpoint groups are distinct from the original groups used for planner retention and EconGym.

Normalization. The original breadth and continuation experiments share role scales estimated from the 103 non-development environments. Thus, narrow conditions can use scales estimated from environments outside their gradient support, but no development target contributes. A matched two-run control estimates each condition's normalizer from its own training support: core-only whole-window skill is 0.042, versus 0.161 for Core+8-C. The prospective retrains additionally exclude the eight new targets from the reference pool. On the development queries, the retrained Core+8-C retains scores 0.132, 0.135, 0.145, and 0.165 at $L = 8, 16, 32, 128$, close to the original group's results.

Checkpoint groups. The principal five-run continuation groups use donor seeds $0, 0, 1, 2, 3$ for continuation seeds 0–4. They therefore represent four independent pretraining donors, not five fully independent training pipelines. The pretraining-distribution comparison uses three matched continuation runs per arm; its structural reference is restricted to those same three runs.

Core information contract. All methods receive the current declared Markov state, the same role mask, and the same normalization reference. The context-using variants additionally receive up to 64 source transitions; the parameter-informed variants receive the true calibration. Default centering uses the episode's steady state. Reward-hidden and causal-centering controls are reported in Appendix H. Seven core environments share identical masks, including subsistence and the base negative control. Masks therefore do not uniquely identify every task, although they can still convey information. The audit does not establish an identity-free representation, nor does it remove calibration information from the observed state distribution.

Table 9: **Checkpoint groups and gradient support.** Original and prospective versions of Core+8-C/Core-continued differ in normalization exclusions, not in their training-model lists. The prospective groups and generator arms are identified separately in the display data.

Group	Training recipe	Environments	Runs
Core-original	Core, 30k steps from scratch	16	5
Core-continued	Core donor + 30k, core only	16	5
Core+8-A/B/C	Core donor + 30k, $p_{\text{old}} = .5$	24	2/2/5
Wide-50	Core donor + 30k, $p_{\text{old}} = .5$	103	5
Wide-75	Core donor + 30k, $p_{\text{old}} = .75$	103	5
Wide-no-context	From scratch, $p_{\text{old}} = .75$	103	2
Prospective core/Core+8-C	Same recipe, new exclusions	16/24	5/5
Generator arms	Same donors + 30k, $p_{\text{old}} = .5$	24	3 each

C Model and training details

Sequence representation. A window is flattened into $[o_1 \,\|\, a_1 \,\|\, r_1 \,\|\, o_2 \,\|\, \cdots \,\|\, o_L]$. Token-type embeddings distinguish observations, actions, and rewards. Learned absolute-position and period-index embeddings encode positions within the window. Observation and action tokens contain role-slot vectors and availability masks. The six-layer transformer has width 256, eight heads, approximately five million parameters, and at most 128 observation positions. It predicts the action at each observation token. The current action and reward are masked from that prediction. The parameter-informed comparator prepends a calibration token; no-context comparators restrict attention to the current observation while retaining its position encoding.

Structural regularization. Each core specification tags a subset of its first-order conditions. These are compiled from the source `.mod` file into residuals over the model's variables. During training, the action is replaced by the network prediction, mapped back to economic units through the inverse normalization; other variables retain their recorded values. Gradients therefore pass through the predicted action to the parameters on which it depends, rather than through a newly solved equilibrium.

Residuals and their products with instruments are averaged within each window before squaring. This targets moment violations rather than the squared size of each realized forecast error. Each residual is normalized by its local action sensitivity. The penalty measures excess violation relative to the recorded path, clipped below at zero. Fractional powers and logarithms use domain guards at 10^{-8}. These guards stabilize differentiation; they do not certify feasibility or optimality of the predicted policy. Imported monetary models contribute only the action loss. Core retraining at $\lambda \in \{0, 1, 100\}$ changes habit and CRRA median regret by at most a few ppm relative to $\lambda = 10$ (for example, 2.6 versus 1.9 ppm on habits at $\lambda = 0$; Appendix H), and the context advantage remains at $\lambda = 0$.

Optimization. Core pretraining uses $30{,}000$ AdamW steps with cosine decay. Continuation initializes from a core donor, uses learning rate 10^{-4}, and trains for another $30{,}000$ steps. A two-level sampler first selects core data with probability p_{old}, then an environment and window within the selected group. Core-continued assigns all second-stage samples to the core group. The stored checkpoint configuration specifies each condition's environment list.

The broad from-scratch controls use $30{,}000$ steps, half the total budget of the continued groups. Their comparison therefore concerns complete recipes, not a compute-matched effect of initialization (Appendix I). Prospective retraining preserves the selected recipe. Generator replacement preserves architecture, donors, steps, replay share, and normalizer, changing only the added monetary data (Appendix D).

D MONETARY PREDICTION AND PRETRAINING-DISTRIBUTION EXPERIMENTS

D.1 PREDICTION PROTOCOL AND COMPARATORS

The baseline and context experiments score the last observation of each 128-position test window. They use up to 500 subsampled windows per development target (321–373 realized). At visible length L, each method receives the last L observations and the preceding $L-1$ labelled transitions. Queries and persistence denominators remain fixed. Scores are pooled over queries within each target, converted to skill, then averaged equally over targets and over network runs. The prospective test uses the same query protocol on its eight targets.

Local and learned-prior estimation. Local OLS and ridge regress the source action on the four lagged roles: nominal rate, inflation, output gap, and output, with an unpenalized intercept. The ridge penalty is selected outside the evaluated targets; it is 1 for both panels. The learned-prior comparator first fits an OLS rule in each of 13,787 training windows from 84 full-feature non-development monetary environments. Trimming removes ill-conditioned fits (Gram condition number above 10^8), per-coordinate coefficient tails outside the 1st–99th percentiles, and the highest 5% of residual variances, leaving 12,432 windows. A random-effects estimate provides a mean rule and coefficient covariance. The evaluated shrinkage estimator regularizes each new window's fit toward the learned mean, with strength selected separately for each L outside the target panel.

For the prospective test, the prior and its tuning are recomputed after excluding the eight additional targets. The selected strengths for $L = 8, 16, 32, 64, 128$ are $10, 10, 1, 1, 1$, compared with $10, 10, 10, 10, 1$ for development. A secondary analytic empirical-Bayes variant with a single global residual variance performs poorly on development, as does the prior mean without local fitting (skill -0.82). We use the tuned shrinkage estimator in the principal comparisons; its results should not be read as an upper bound on all Bayesian estimators.

TabPFN-v2. We use the public TabPFN-v2 regression model (Hollmann et al., 2025), with the default ensemble of eight and no target-panel tuning. Each window is a separate table containing the $L-1$ observed lagged-state/action pairs and one query row. The feature set and query are identical to those of local ridge. Unlike MACROIC, this interface does not represent the history as ordered observation/action/reward tokens. This comparison tests a general-purpose pretrained estimator, whereas the generator experiment below holds the policy architecture fixed.

Uncertainty and rounding. Reported intervals use 2,000 paired bootstrap draws, resampling windows within each target and preserving the paired predictions for each comparison. They measure uncertainty conditional on the target list, not uncertainty over new structural families. Training runs and shared donors are identified in Appendix B; five continuations do not imply five independent pretraining samples. Marginal scores and paired differences are rounded separately in the source reports. A confidence interval including zero does not establish statistical equivalence.

D.2 PROSPECTIVE EVALUATION

Tables 10 and 11 give complete curves for the eight new evaluation targets. The selected recipe was fixed before their prediction scores were observed; target membership and earlier pilot usage are documented in Appendix B. The original decision criterion required a positive, resolved gap over the strongest comparator at both $L = 8$ and $L = 16$. It is not fully met: the learned-prior gap at $L = 16$ includes zero. Positive gaps at $L = 8$ and $L = 32$, and the loss of a clear advantage at $L = 128$, describe the measured pattern without replacing that criterion after observing the results.

At $L = 8$, Core+8-C's per-target skills are 0.007 (EA_GE10), 0.117 (EA_SR07), 0.143 (NK_CK08), 0.282 (NK_PP17), 0.169 (NK_RA16), 0.253 (NK_RW06), 0.256 (US_FGKR15), and 0.273 (US_RA07). All are positive, although the EA_GE10 gain is small. The stronger learned-prior performance on this set makes the network's advantage smaller than on development. Differences in skill between the two panels should not be interpreted as improvement under a common difficulty distribution.

Table 10: **Complete prospective prediction results.** Mean query-only skill on eight targets. Core-only and Core+8-C use five retrained continuation runs with the new normalization exclusions. Higher is better.

L	Core+8-C	Core-only	Local ridge	Learned prior	TabPFN-v2
8	0.187	0.112	−0.069	0.140	−0.215
16	0.205	0.123	0.059	0.180	0.026
32	0.210	0.123	0.140	0.164	0.099
64	0.220	0.124	0.193	0.201	0.167
128	0.229	0.121	0.238	0.240	0.221

Table 11: **Prospective paired differences.** Core+8-C minus each comparator, with 95% paired-bootstrap intervals conditional on the eight targets.

L	vs. local ridge	vs. learned prior	vs. TabPFN-v2
8	+0.256 [+0.198, +0.325]	+0.047 [+0.022, +0.075]	+0.403 [+0.282, +0.593]
16	+0.145 [+0.108, +0.187]	+0.024 [−0.001, +0.051]	+0.178 [+0.136, +0.233]
32	+0.071 [+0.041, +0.106]	+0.046 [+0.019, +0.076]	+0.112 [+0.069, +0.167]
64	+0.027 [+0.005, +0.051]	+0.019 [−0.002, +0.043]	+0.054 [+0.028, +0.083]
128	−0.009 [−0.026, +0.009]	−0.011 [−0.028, +0.008]	+0.008 [−0.013, +0.032]

D.3 Development results

Table 1 in the main text gives the complete development-panel curves. Here we report their paired differences and the target-supervised reference.

The query-only Core-continued score at $L = 128$ is 0.004, compared with 0.164 for Core+8-C. A target-supervised reference trains the same continuation recipe with the ten targets restored to gradient support, retaining their close-relative exclusions (two runs). Its query-only mean skill is 0.211. The mean of the per-target fractions of the persistence-to-supervised MSE gap recovered by Core+8-C is 0.73 (range 0.37–0.96). This is a target-trained reference under one recipe, not a performance ceiling, and the reported mean fraction is not the ratio of mean skills.

Global ridge and a two-layer MLP trained on the Core+8-C support reach 0.044 and −0.032 query skill. These models use current roles, masks, and the previous action, rather than fitting to the new demonstration. They are controls for a pooled rule, not substitutes for the history-adapting comparators above.

D.4 Replacing the added corpus

The structural arm continues on corpus C. The reduced-form arm replaces each added environment with a stable linear dynamical system whose sampled decision rule feeds back into the transition. The independent-observation arm applies the same rule family to i.i.d. observations. Interface, rule-coefficient ranges, and episode counts are matched, and marginal scale and persistence statistics are estimated from training data. All arms retain the same structural core donors, normalizer, replay share, and continuation budget. Three matched continuation runs are used per arm, including the structural reference.

Matching is approximate. Inflation and output-gap standard deviations differ by 1–8%, interest-rate standard deviations in the replacements are 15–70% higher, and output scale is underestimated for two trend-dominated models. One generator design is evaluated per arm. These differences preclude assigning the entire performance gap to economic equilibrium restrictions alone.

Reduced-form data improve full-context performance relative to core-only continuation, but are less useful than the structural corpus at short lengths. The independent-observation replacement performs poorly at every measured length. This is a limitation of the tested generator, not evidence that generic function learning cannot support in-context adaptation. Because all arms retain structural core pretraining, the experiment identifies the effect of the added data distribution rather than the effect of all economic pretraining.

Table 12: **Development paired differences.** Core+8-C minus learned-prior estimation or TabPFN-v2, with 95% intervals. At full context, the learned-prior estimator has a small resolved advantage.

L	vs. learned prior	vs. TabPFN-v2
8	$+0.123$ $[+0.085, +0.160]$	$+0.520$ $[+0.431, +0.633]$
16	$+0.109$ $[+0.074, +0.142]$	$+0.379$ $[+0.293, +0.492]$
32	$+0.060$ $[+0.032, +0.089]$	$+0.152$ $[+0.100, +0.216]$
64	$+0.004$ $[-0.024, +0.029]$	–
128	-0.025 $[-0.051, -0.005]$	$+0.004$ $[-0.019, +0.030]$

Table 13: **Full generator-replacement results.** Three matched continuation runs per arm. Both target sets use query-only skill. The last column reports the structural-minus-reduced-form difference and its 95% interval; this three-run structural reference differs from the five-run reference in Table 10.

L	Structural	Reduced-form	i.i.d. observations	Structural − reduced-form
Development panel				
8	0.132	-0.505	-4.09	$+0.637$ $[+0.544, +0.758]$
16	0.136	-0.190	-4.21	$+0.326$ $[+0.267, +0.399]$
32	0.144	0.013	-4.15	$+0.130$ $[+0.095, +0.170]$
64	0.148	0.103	-4.02	$+0.045$ $[+0.020, +0.074]$
128	0.161	0.122	-3.91	$+0.040$ $[+0.020, +0.064]$
Prospective test				
8	0.186	0.061	-0.88	$+0.125$ $[+0.092, +0.161]$
16	0.205	0.134	-0.91	$+0.071$ $[+0.046, +0.099]$
32	0.210	0.174	-0.89	$+0.036$ $[+0.014, +0.063]$
64	0.222	0.196	-0.87	$+0.026$ $[+0.006, +0.047]$
128	0.231	0.206	-0.86	$+0.026$ $[+0.001, +0.062]$

D.5 Corpus composition and whole-window scoring

The breadth sweeps in Figure 3 score all teacher-forced positions in each window, not only its final query. We retain this whole-window metric for those comparisons and label it explicitly. Its values need not equal the query scores above: Core-continued/Core+8-C/Wide-75 reach $0.027/0.160/0.133$ on the whole-window metric, versus $0.004/0.164/0.139$ on the query metric. The composition comparison in Figure 3 uses two matched runs per condition; its core-only reference, 0.036, is the two-run subset of the five-run Core-continued group whose mean is 0.027. No paired query gap is computed from the whole-window values.

The three eight-model subsets are successive disjoint blocks of one seeded shuffle of the 87 eligible non-development monetary models, not independent draws. Subset A contains EA_SR07, NK_BGUS10, NK_CFP10, NK_CGG99, NK_LWW03, US_FMS134, US_KS15, and US_PM08. Subset B contains CL_MS07, EA_GE10, EA_PV15, NK_ADE25cpi, NK_GS14, NK_RW06, US_ACELswt, and US_CCTW10. Subset C contains NK_CW09, NK_MPT10, NK_RW97, US_ACELm, US_CCF12, US_FRB22mcapwp, US_FU19, and US_OR03. The 32-model condition uses the first 32 entries of the same shuffle. All nonzero-breadth conditions in the connected curve use the same continuation recipe and 50% core replay.

E Context length, positions, and corrupted demonstrations

Length protocol. Each prediction target is the final observation of a full test window. A context of length L contains the last L observations and $L-1$ past transitions. The target action, query observation, and persistence denominator are unchanged. However, truncation restarts the learned absolute-position and period-index embeddings at the beginning of the shorter window. Thus, these curves measure the model's actual performance with reindexed short windows, not the causal effect of removing history while fixing all position features.

Table 14: **Development per-target whole-window skill.** This is the metric used in the breadth experiments, not the query-only comparator metric. Wide-NC is a separately trained no-context model with a different recipe and two runs.

Target	Core-cont.	Core+8-C	Wide-75	Wide-50	Wide-NC
BRA_SAMBA08	-0.292	0.122	0.004	0.004	-0.982
CA_LS07	0.167	0.224	0.247	0.246	0.175
DEREA_GEAR16	0.106	0.092	0.134	0.133	-0.041
EA_QUEST3	-0.039	0.054	0.035	0.039	-1.327
NK_BGG99	0.107	0.306	0.315	0.320	0.203
NK_GK09lin	0.118	0.247	0.203	0.200	-0.298
US_IAC05	0.183	0.233	0.274	0.273	0.210
US_JPT11	-0.132	0.053	-0.055	-0.055	-0.637
US_LTW17	0.059	0.173	0.116	0.119	-0.127
US_SW07	-0.007	0.093	0.060	0.060	-0.117
Mean	0.027	0.160	0.133	0.134	-0.294

Table 15: **Query-only context curves and corruptions on development targets.** Five continuation runs per checkpoint group. Length truncation reindexes positions; corruption rows preserve the full length and query position. Wide-50 values exclude the separately trained no-context runs.

Condition	Wide-75	Core+8-C	Wide-50
$L = 1$	+0.033	-0.005	+0.048
$L = 2$	-0.135	+0.016	-0.090
$L = 4$	+0.050	+0.120	+0.058
$L = 8$	+0.097	+0.132	+0.103
$L = 16$	+0.121	+0.134	+0.122
$L = 32$	+0.136	+0.144	+0.136
$L = 64$	+0.137	+0.150	+0.140
$L = 128$	+0.139	+0.164	+0.139
Repeated transition	+0.028	+0.019	+0.020
Shuffled history	-1.844	-1.526	-1.800
Foreign history	-2.509	-2.334	-2.311

This distinction is visible even in the separately trained no-context model: its mean skill varies from approximately -0.29 to -0.31 across lengths despite identical current-observation content. A prediction-level check confirms different outputs when only query position changes. The size of this effect in the no-context model is not an upper bound on position effects in the context-using models. A fixed-position history-mask experiment would be needed to isolate history length from positional encoding.

Fixed-length interventions. All corruptions alter only the history before the query and preserve the query observation, token count, and positional indices. Shuffle permutes the preceding transitions. Foreign substitutes the history of another episode of the same environment, with a different calibration. Repeated-transition context repeats the final history triple across the window. These interventions test whether the existing policy uses the supplied demonstration; they can also create out-of-distribution inputs.

Core+8-C has skill 0.164 with the original full history, 0.019 with repeated transitions, -1.526 with shuffled history, and -2.334 with foreign history. Wide-75 shows the same qualitative dependence. These losses support reliance on contextual evidence but do not distinguish latent-state inference from inference about the source rule. The negative short-window scores at $L = 2$ for the wide models are reported rather than omitted. Their nonmonotonicity further cautions against interpreting a length curve as a unique identification procedure.

F WELFARE EVALUATION AND LOCAL-POLICY COMPARATORS

F.1 REFERENCE ROLLOUT AND SCORING

After a 64-period source-policy warm-up, learned actions drive the exact production, resource, and capital-accumulation equations. The reference is the calibration's unpruned numerical decision rule, rolled through the same exact transitions from the same inherited stocks under the same exogenous innovations. Comparing against the original perturbation trajectory instead would confound policy error with its transition-approximation error. The rolled reference is attainable, but it is not a certified exact optimum.

The pairwise score is defined in Equation (5), with takeover indexed by $t = 0$, horizon $H = 40$, and terminal prices $\widehat{q}$ shared across policies. The common additive terminal-value constant cancels. The evaluator converts the score difference to a permanent-consumption-equivalent loss and reports it in ppm; positive values indicate lower welfare than the rolled reference. For floor utilities the conversion is solved numerically by bisection rather than a homothetic closed form. Negative regret is retained and does not certify improvement over the true optimum.

Replay consistency. A policy that reproduces the reference actions has the same state and utility path, hence zero difference in Equation (5). This follows by induction through the common transition. It is enforced as an evaluator self-test, not treated as a separate theoretical contribution. The reported tolerance is 10^{-9} on the replay-regret check.

F.2 TERMINAL PRICES AND APPROXIMATION ERROR

Fix the terminal exogenous state and consider a continuation-value function $V(x)$, with reference terminal stock $x^\star = x_H^{\mathrm{ref}}$. The implemented approximation is

$$\widehat{V}(x) = V(x^\star) + \widehat{q}^{\top}(x - x^\star). \tag{6}$$

If ∇V is locally L_V-Lipschitz and $\|\widehat{q} - \nabla V(x^\star)\| \le \varepsilon$, the Taylor remainder and Cauchy–Schwarz inequality give

$$\beta^H \left|\widehat{V}(x) - V(x)\right| \le \beta^H \left(\frac{L_V}{2}\|x - x^\star\|^2 + \varepsilon\|x - x^\star\|\right). \tag{7}$$

Differentiability alone would not give the quadratic remainder. The bound separates curvature error from gradient-proxy error; we do not estimate uniform constants L_V or ε from the evaluation episodes. Numerical checks below assess the implemented approximation directly.

Internal-habit prices. For the interior planner problem with $g = c - hc_-$ and $K' = F(K, z) + (1 - \delta)K - c$, the Bellman equation is

$$V^*(K, c_-, z) = \max_c \left\{u(c - hc_-) + \beta\, \mathbb{E}[V^*(K', c, z') \mid z]\right\}.$$

Writing $R = F_K(K, z) + 1 - \delta$ and suppressing state arguments, the first-order and envelope conditions yield

$$\mu = u'(g) - \beta h\, \mathbb{E}[u'(g') \mid z], \qquad \partial_K V^* = \mu R, \qquad \partial_{c_-} V^* = -h\, u'(g). \tag{8}$$

The inherited habit stock has a negative marginal value: it raises the consumption floor. These are identities for the optimal value under the stated regularity assumptions, not identities for the value of an arbitrary approximate decision rule.

The implemented prices use stationary proxies evaluated at the reference's last scored utility gap. In particular, $(1 - \beta h)u'(g_{H-1}^{\mathrm{ref}})$ approximates the next-period μ, and $u'(g_{H-1}^{\mathrm{ref}})$ approximates the next-period marginal utility. The substitution is exact in the deterministic stationary limit, not generally under stochastic stationarity. Both policies use the same reference-based prices, so their price error enters through the difference in terminal stocks.

Table 16: **Terminal-value sensitivity.** Original checkpoint groups, 200 episodes per economy; medians across five runs of per-run episode-median regret in ppm. The alternative replaces the affine terminal value with a 600-period continuation of the reference rule from each policy's terminal stocks.

Economy	Checkpoint	Linear terminal	Continuation terminal
Internal habits	Core-continued	0.9	0.3
	Core+8-C	2.0	1.6
	Wide-75	4.2	3.9
Subsistence	Core-continued	0.6	3.1
	Core+8-C	4.0	5.7
	Wide-75	11.4	13.7

Finite differences. For 50 test calibrations per economy, each reference terminal stock is perturbed by $\pm 1\%$, followed by a 600-period zero-innovation continuation under the reference rule. Capital-price ratios have median 1.00, with median absolute deviations of 1.8% on the welfare-sensitive economies and 0.4% on the smoother ones. Habit-stock prices have the correct sign in all 50 episodes and median relative error about 2.5%. These are finite-difference diagnostics for the approximate reference rule; they are neither a proof of derivative accuracy nor a uniform bound in Equation (7).

Continuation sensitivity. Replacing the affine terminal value with a 600-period continuation from each policy's own terminal stocks changes regret levels but preserves the reported group ordering on habits and subsistence (Table 16). The same check preserves the controlled core context/no-context contrasts. Zero-innovation continuation is a sensitivity check, not the stochastic infinite-horizon value. Likewise, order-one versus order-two reference differences (0.5 ppm on habits and -0.6 ppm on subsistence) indicate numerical resolution, not distance from the true optimum.

F.3 Fitting a local policy from the same demonstration

We retain the physical 64-period warm-up, takeover state, shock path, horizon, and terminal treatment, but reveal only the final $K \in \{8, 16, 32, 64\}$ warm-up transitions. The network's wrapper hides earlier warm-up periods. Linear and quadratic ridge rules are fitted once on the same K state–action pairs and frozen during takeover. The development-selected penalties are 0.01 and 10^{-5} respectively. Each method receives the full declared Markov state.

Local linear fitting remains competitive even with eight transitions. Its p95 is below the network's in ten of twelve configurations, and it has no observed failures. The network has lower median regret than the linear fit in the habit economy at $K = 32$ and 64, but the paired episode win rates are close to one half. Quadratic fits can have extreme tails and failures under habits. Thus, the results do not support either systematic short-prefix welfare superiority or general tail superiority of the network over local linear estimation. Conditional quantiles and separately reported failures should be read together.

G Preference variation in the planner suite

Why these preferences. At a smooth interior optimum, the first variation of welfare along feasible policy perturbations is zero. Small action errors can therefore have little welfare effect in smooth growth models. Subsistence and internal habits make departures toward the utility boundary more costly while allowing the preference parameters to vary without changing deterministic steady-state capital and consumption.

For both variants, define utility on the positive gap $g > 0$ by

$$u_\sigma(g) = \begin{cases} g^{1-\sigma}/(1-\sigma), & \sigma \neq 1, \\ \log g, & \sigma = 1. \end{cases}$$

Under subsistence, $g_t = c_t - \bar{c}$, with $\bar{c} = \phi c_{\mathrm{ss}}$ and $\phi \in [0, 0.6]$ (Geary, 1950; Stone, 1954). Relative risk aversion measured in consumption is $\sigma c/(c - \bar{c})$, hence $\sigma/(1-\phi)$ at the deterministic steady

Table 17: **Short-prefix planner control.** Regret in ppm on 200 attempted episodes per economy, with median and p95 computed on shared valid episodes. Network statistics are medians across five runs. Linear/quadratic failure counts are reported separately; no network failure was reported. Win rate is the fraction of matched episodes on which the network has lower regret than the linear fit. This table uses prefix-wrapper evaluation and shared-valid statistics; its entries are not substituted for the original retention values in Table 3. The two aggregations differ: here the network value is the median over shared valid episodes of the per-episode median across the five runs, whereas Table 3 reports the median across runs of each run's own episode median, on all valid episodes of that run; at K=64 this yields 0.96 against 2.0 ppm under habits from the same checkpoints.

K	Core+8-C		Local linear		Local quadratic		Failures	Win rate
	Median	p95	Median	p95	Median	p95	lin./quad.	vs. linear
Ramsey base								
8	2.64	29	0.60	63	0.00	50	0/2	0.40
16	1.95	19	−0.02	7	0.00	1	0/0	0.33
32	1.50	18	0.01	4	0.00	0	0/0	0.30
64	1.41	15	−0.02	3	−0.01	2	0/0	0.30
Internal habits								
8	6.14	230	−1.34	275	5.35	6.4×10^6	0/10	0.48
16	1.50	164	0.01	106	0.05	35	0/0	0.48
32	0.77	128	1.16	88	0.06	17	0/0	0.51
64	0.96	149	1.51	77	0.50	1.9×10^9	0/8	0.50
Subsistence								
8	13.18	469	−3.30	191	0.25	424	0/1	0.39
16	7.33	402	−0.40	202	0.12	118	0/0	0.45
32	4.57	280	−0.67	211	0.00	60	0/0	0.41
64	3.04	198	−0.23	144	−0.20	62	0/1	0.42

state. Under internal habits, $g_t = c_t - hc_{t-1}$ with $h \in [0.3, 0.8]$ (Constantinides, 1990). Lagged consumption belongs to the Markov state and is observed by every comparator. Habits are internal, so today's consumption affects tomorrow's utility floor as well as tomorrow's capital.

Proposition 1 (Preference parameters and the deterministic steady state)**.** *Consider production $Y_t = e^{z_t}K_t^\alpha$, accumulation $K_{t+1} = Y_t + (1-\delta)K_t - c_t$, discount factor β, and either $g_t = c_t - \bar{c}$ or $g_t = c_t - hc_{t-1}$. Suppose the deterministic steady state has $z^* = 0$, is interior with $u'(g^*) > 0$, and, under habits, satisfies $1 - \beta h > 0$. Then*

$$1 = \beta\left[\alpha(K^*)^{\alpha-1} + 1 - \delta\right], \qquad c^* = (K^*)^\alpha - \delta K^*. \tag{9}$$

At fixed (α, β, δ), these levels do not identify utility curvature or the gap parameters $\bar{c}$ and h.

Proof. For subsistence, the Euler equation is $u'(g_t) = \beta\mathbb{E}_t[u'(g_{t+1})R_{t+1}]$, where $R_{t+1} = \alpha e^{z_{t+1}}K_{t+1}^{\alpha-1} + 1 - \delta$. At the deterministic steady state the positive marginal utility cancels, giving $1 = \beta R^*$. The resource constraint gives the second equation.

For internal habits, the marginal value of a unit of consumption expenditure is $\mu_t = u'(g_t) - \beta h\mathbb{E}_t u'(g_{t+1})$, and the Euler equation is $\mu_t = \beta\mathbb{E}_t[\mu_{t+1}R_{t+1}]$. At the steady state, $g^* = (1-h)c^*$ and $\mu^* = (1-\beta h)u'(g^*) > 0$. Cancelling μ^* again yields $1 = \beta R^*$. □

The proposition concerns deterministic steady-state levels, not identification from a stochastic observation. Its role is to clarify the core centering convention: at fixed technology and discounting, subtracting these levels does not itself reveal the gap or curvature parameters. It does not establish that history is necessary once the calibration and full state are known, or that a finite demonstration uniquely identifies a calibration. Those are distinct questions from the empirical value of context for the trained policy.

H FULL CORE RESULTS

This appendix gives further controls for the core information comparison in Table 4. All numbers use the corrected evaluator of Appendix F and the unified protocol: 5 seeds per variant (15 runs for the three comparators of Table 4, plus 5 for the θ-only variant below), 200 test episodes per economy, rolled reference, median across seeds of per-seed episode medians.

The information factorial. A fourth core variant (θ-only: calibration token given, history truncated) completes a 2×2 factorial in history and calibration, scored on the same episodes. On the welfare-sensitive economies, with-context scores 1.9 ppm against 33.3 for θ-only under habits (paired hierarchical difference -32 ppm, CI $[-47, -20]$) and 6.3 against 17.7 under subsistence. This comparison concerns how the trained network uses different conditioning representations. It does not imply that demonstrations contain more information than the calibration and full Markov state.

Tail behaviour. Medians could hide rare failures. Pooling episodes across seeds, context cuts the habit-economy CVaR_{95} (mean of the worst 5%) from 3,833 to 299 ppm and the worst episode from 13,781 to 1,331 ppm; on subsistence CVaR_{95} falls from 2,827 to 741 ppm. In paired comparisons on shared episodes the with-context model beats no-context on 87% of habit episodes and 67% of subsistence episodes. No episode diverges or is dropped in any variant.

Horizon. Re-scoring the welfare-sensitive economies at $H \in \{20, 80, 128\}$ leaves the contrasts intact; the gap is fully present at $H{=}20$ and per-period regret is roughly constant in H for every variant. The context-using model remains within single-digit median ppm of the reference through $H{=}128$.

Reward and normalization controls. Zeroing every reward token at evaluation changes little ($1.9 \to 0.4$ ppm on habits, $6.3 \to 5.6$ on subsistence), and five seeds retrained with rewards permanently zeroed reproduce the main results (1.2, 2.6 ppm): reward tokens are not required for the measured context benefit. A pilot retrained with references computed causally from the episode's opening, with no calibration-derived input, pays a few ppm (9.2 vs. 2.6 on habits, 6.5 vs. 3.9 on subsistence, matched $\lambda{=}0$ seeds) while the reported context advantage survives. This reduces concern that the effect depends entirely on calibration-derived centering.

Context corruption in welfare units. On the frozen core checkpoints, shuffling the context in time, substituting a context from a different calibration, or replicating a single transition destroy the habit-economy gain (154, 194, 44 ppm against 1.9 with the true context), while the gain saturates by $k \approx 32$ transitions. Removing only states or only actions also degrades performance; such inputs are outside the training format, so this is a robustness diagnostic rather than a channel-identification result.

Markov-state audit. The no-context comparator's meaning rests on its input being the complete Markov state. The specifications declare the state explicitly, lags included; for the habit economy it is capital at $t{-}1$, consumption at $t{-}1$, and log productivity at t. The harness independently derives each economy's state as its predetermined stocks plus exogenous processes, raising an error unless that set equals the declared state columns. Every scored number passed this check by construction.

I CONTINUATION AND REPLAY ABLATIONS

Balanced sampling from scratch. Training the wide corpus from random initialization for 30k steps with the two-level sampler, two seeds per mixture, gives on internal habits median regrets of 79.2, 57.5, and 56.6 ppm at core shares $p_{\text{old}} = 0.25, 0.5, 0.75$, and on subsistence 64.4, 40.7, 35.1 ppm, with panel skills 0.015, 0.033, 0.044. More replay helps precision, but from-scratch joint training still incurs three to thirteen times the median regret of the continued checkpoints on these economies while transferring worse. The per-run values behind these medians span a noticeable range (e.g. habit 40.6–72.7 ppm at $p_{\text{old}} = 0.75$), so the two-run medians locate the recipe's scale, not a precise level.

Table 18: **Training recipes and the measured joint result.** Broad from-scratch training and the continuation recipes use the same policy architecture but differ in initialization and total training budget as shown. Continuation retains substantially lower planner regret while improving held-out skill in the tested configurations. Habits and subsistence in ppm; skill is the development-panel whole-window mean (Appendix D.5).

Recipe	Init	Steps	p_{old}	Habits	Subsist.	Skill	Runs
Broad joint, from scratch	random	30k	0.75	56.6	35.1	0.044	2
Continuation + replay	core donor	30k + 30k	0.50	6.7	12.9	0.134	5
Continuation + replay	core donor	30k + 30k	0.75	4.2	11.4	0.133	5

Continuation. The continued models (Table 18) reach 6.7 ppm (habits, $p_{\text{old}} = 0.5$) and 4.2 ppm ($p_{\text{old}} = 0.75$) with panel skills 0.134 and 0.133. The comparison to from-scratch training is between complete recipes: continuation changes initialization, learning rate, total budget (60k vs. 30k steps), and training order simultaneously. A budget-matched from-scratch control and an exposure-matched shuffled-order control have not been run; the paper accordingly claims a useful training path, not an isolated causal mechanism.

Core-continued as the narrow reference. The proper narrow reference for the breadth curve is Core-continued (30k + 30k on the core corpus alone), which improves markedly over Core-original on the planner suite (e.g. habits 0.9 vs. 1.9 ppm) while remaining weak on the panel (0.027 mean skill). Comparing added-breadth conditions to Core-original would conflate breadth with the additional optimization budget.

J External transfer to EconGym

Task and information. EconGym (Mi et al., 2025) is an independently developed economic simulator. We use its `optimal_monetary` scene, where a central bank chooses the policy rate. Nothing from this simulator enters the policy-training, normalization, or prior-fitting supports. The source is a noisy interest-rate rule responding to inflation and growth. Other agents remain rule-based during the transfer evaluation, and the reserve ratio is fixed at 0.08.

The open-loop dataset contains 48 source-rule episodes (seeds 1000–1047), split into 24 development and 24 test episodes. The local ARX penalty is selected on development only. The role interface maps the policy rate and inflation to gross-rate roles, output to log GDP, and wages and prices to their log-index roles. Lending and deposit rates are excluded because clipping to the policy rate could duplicate the target. Reference deviations are estimated causally at each forecast origin.

Prediction. Test origins begin at step 33 with stride two. We evaluate 8–32 visible positions because the episodes are short. Under the source rule, rates spend long spells at the lower bound; both all-origins skill and a predeclared subset with the previous rate above the bound are reported. These aggregates answer different questions and are not interchangeable.

Core+8-C improves on persistence at every tested length, but local ARX is stronger on all origins. At $L = 16$, their skills are 0.189 and 0.618. Away from the lower bound, Core+8-C reaches 0.748–0.791 over $L = 8, 16, 32$, and the ARX advantage is less consistent. This conditional result does not reverse the unfiltered comparison.

Closed-loop takeover. Each method takes over after the same 32-step source-rule warm-up on 24 test episode seeds. The policy predicts a deviation, converts it to a level using the causal reference, and clips to the allowed action bounds. The local ARX comparator is fitted once to the 31 labelled warm-up transitions, then frozen. The added noise-free-source comparator removes the source rule's action noise without changing its deterministic rule.

Table 5 in the main text reports the mean reward per surviving step, total post-takeover return, and a fixed 20-step return with zero reward after termination. Total and fixed-horizon returns avoid normalizing away different survival times. The network row uses the original continuation seed

Table 19: **EconGym open-loop prediction.** Test-set skill relative to persistence from pooled squared errors. Local ARX is fitted on the last $L - 1$ labelled transitions; AR and BVAR use all available history. The all-origins column is the unfiltered aggregate.

Method	All origins	Away from the lower bound
Random walk	+0.000	+0.000
AR (BIC order)	-0.186	+0.431
Minnesota BVAR	-0.119	+0.456
Local ARX, L=8	+0.586	+0.821
Local ARX, L=16	+0.618	+0.821
Local ARX, L=32	+0.447	+0.638
Core+8-C, L=8	+0.155	+0.767
Core+8-C, L=16	+0.189	+0.791
Core+8-C, L=32	+0.150	+0.748

Table 20: **Paired EconGym differences.** Core+8-C minus each alternative, with 95% episode-bootstrap intervals. The advantage over hold-last in reward/step is not resolved in total return.

Comparator	Reward/step difference	Total-return difference
Noise-free source rule	$+0.022$ $[+0.015, +0.028]$	$+0.36$ $[+0.24, +0.46]$
Local ARX	-0.020 $[-0.028, -0.012]$	-0.30 $[-0.44, -0.16]$
Hold last rate	$+0.015$ $[+0.004, +0.025]$	$+0.03$ $[-0.11, +0.16]$

0; seeds 1–4 yield reward/step 0.403–0.408 and total return 8.25–8.39. The intervals in Table 20 resample the 24 paired episode seeds and refer to network seed 0.

The network improves over both source-rule variants, so removing the teacher's action noise alone does not explain the gain. However, ARX achieves higher return, the network has no resolved total-return advantage over hold-last, and the random rule draw is within 0.05 of its total return. These comparisons delimit the external result: the pretrained policy is usable without scene-specific training, but a simple fit to the same demonstration is more effective.

Scene-trained PPO and action contract. PPO reaches reward/step 0.540, total return 19.08, and mean survival 67.3 steps, compared with about 52 steps for the frozen network. Survival is counted from the start of the episode and includes the 32-step warm-up, so the frozen network survives about 20 post-takeover steps and PPO about 35; total return equals reward per step times post-takeover survival. The original PPO interface includes a reserve-ratio output as well as the policy rate. Repeating training with reserve-ratio bounds pinned at 0.08 reproduced the checkpoint weights and evaluation results; the associated loan cap was inactive in these runs. Thus, restricting this action did not change the reported reference. This is specific to the tested scene and agents, not a general equivalence of the action spaces.

The short episodes and substantial distribution shift limit the scope of this test. A reward gain in this scene is not evidence of equilibrium validity under arbitrary interventions, long-horizon stability, or effective policy for historical economies.

K Evidence organization and reproducibility

Reporting sources. The paper separates query-only prediction, whole-window prediction, original planner retention, prefix-wrapper control, and external-simulator return. Each display is tied to its own evaluation cohort, and the new tables and the prospective plot are generated from one structured numerical snapshot, so their values can be checked against the same source entries.

Selection history. The development panel selected corpus C and the wide replay share. The prospective-target draw and the replacement generators were fixed before their respective scoring runs. The complete selection history, including earlier pilot use of three prospective targets, is in

Appendix B. No claim depends on treating the development set as untouched or on counting five shared-donor continuations as five independent pretraining pipelines.

Compute. The original experiments use 30k-step stages of a roughly five-million-parameter model. Planner evaluation runs on CPU with cached Octave/Dynare solves. In the additional experiments, a prospective continuation took approximately 53 minutes on one A100; the six non-structural continuation runs took roughly 50 minutes each. The TabPFN comparison required 14,680 fits, about 45 minutes on each of two A100s. The short-prefix planner evaluation took about two hours with 12 CPU processes, and the rate-constrained PPO retraining took 3.9 CPU hours. These are measured run costs for the reported configurations, not a hardware-normalized efficiency comparison.

Scope of reproduction. The appendices document the corpus, interfaces, splits, training recipes, and evaluation protocols needed to reimplement the experiments. Reproducing training additionally requires the packed trajectories, source-model specifications, training configurations, and experiment code; reproducing intervals requires the paired predictions and episode identifiers. Import tooling and source lists are needed where model-file licenses do not permit redistribution.

A compute-matched wide from-scratch control, an exposure-matched shuffled-order control, and a matched no-context continuation are not part of the selected evidence. The paper also does not evaluate fixed-position history masking, exhaustive non-structural generators, or deployment on historical economies. These exclusions are reflected in the scope of the claims rather than treated as implicit positive results.