BEYOND PREDICTION: STRUCTURAL TRANSFER AND PARTIAL IDENTIFICATION IN SYMBOLIC DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational models of human learning and memory aim to discover concise and interpretable relationships among practice history, temporal intervals, and subsequent recall outcomes. Symbolic regression provides a direct approach toward this goal; however, in heterogeneous learning environments, good fit on a single dataset or recurring expressions across multiple environments do not necessarily indicate transferable structures, nor do they reveal which structural components truly receive cross-environment evidential support. To address this issue, we propose Cross-Fitted Domain-Robust Symbolic Regression (CF-DRSR), an evidence-calibrated framework for cross-environment symbolic modeling. CF-DRSR separates symbolic discovery, transfer validation, and structural identification: it constructs comparable candidate structural families through symbolic search and structural alignment, evaluates the transfer support of candidate structures on data not involved in discovery through source-held-out predictive qualification, and characterizes identifiable structural content under multiple compatible explanations through set-valued structural identification. Experiments on multiple heterogeneous learning datasets demonstrate that CF-DRSR achieves effective prediction on unseen domains while revealing the environment-dependent transfer scope and boundaries of candidate structures. Further analysis shows that recurring predictive structures do not necessarily correspond to uniquely identifiable laws, as multiple structural explanations may remain consistent with the available evidence. These findings suggest that cross-environment symbolic modeling should distinguish among three levels: predictive generalization, structural transfer, and structural identification.

1 INTRODUCTION

A long-standing goal of memory modeling is to extract concise and interpretable relations among learners' practice histories, temporal spacing, and subsequent recall. Prior work shows that the timing of practice and retrieval practice substantially affects retention and transfer (Smolen et al., 2016; Carpenter et al., 2022). Symbolic regression offers a direct route toward this goal: it searches behavioral data for explicit mathematical expressions, turning a predictor into an inspectable structural hypothesis (Schmidt & Lipson, 2009; Udrescu & Tegmark, 2020). Real learning data, however, arise from different tasks and learner populations, so an expression fitted and interpreted within one dataset may not extend to another.

Multi-dataset symbolic modeling exploits common information through shared functional structures, joint equation discovery, or dataset-specific parameters (Sen Fong & Motani, 2024; Ciocanel et al., 2026). These approaches can reveal potential commonalities among observed datasets, but shared fit within those datasets does not establish that a structure extends to a dataset excluded from discovery.

Recent work combines symbolic regression with domain generalization to learn explicit equations that generalize from multiple source domains to an unseen domain (Alshehri & Pishnamazi, 2026; Alhowyan et al., 2026). Such methods typically learn and evaluate complete equations and rely on sufficiently stable transferable relations across the domains. In real learning data, domains differ in target language, task design, learner population, and observation process. They may therefore

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.

share only parts of a candidate structure and need not admit a single equation that describes them all. Good unseen-domain prediction by a complete equation neither establishes that all domains obey one regularity nor identifies which constituent structures receive stable cross-dataset support.

We therefore treat cross-domain sharing as an object of evaluation rather than a modeling premise: can a candidate structure discovered from source data retain predictive value in a dataset excluded from its construction, and across which datasets does the available evidence persist?

To answer these questions, we propose **CF-DRSR**, a cross-fitted symbolic domain-generalization framework consisting of DRSR and cross-environment transfer certification. DRSR constructs candidate expressions through multiple discovery routes and forms a zero-shot predictor. Transfer certification evaluates MIBD-aligned candidate families on datasets excluded from candidate construction and characterizes their evidence scope and degree of identification. Figure 1 contrasts the standard symbolic-DG inference—good zero-shot prediction is taken as the endpoint—with CF-DRSR, which uses cross-fitted held-out evidence to ask separately whether a symbolic structure transfers and what structural content is identified.

Our contributions are threefold:

1. **Cross-fitted symbolic domain generalization.** CF-DRSR separates out-of-domain generalization of the complete predictor from cross-dataset transfer evaluation of candidate families under strict source-only constraints.
2. **Candidate-family transfer certification and set-valued identification.** MIBD and a dual-standard identity rule align candidate families, whose predictive value is evaluated in datasets excluded from their construction. The resulting evidence characterizes both support scope and identified content relative to a frozen candidate space.
3. **Shared candidate structures have scope and boundaries.** Across four real learning datasets, CF-DRSR attains competitive zero-shot prediction while retaining multiple structural explanations compatible with the evidence. Good cross-domain prediction need not identify a unique shared regularity.

2 Related Work

Symbolic regression and cross-domain generalization. Symbolic regression searches for explicit mathematical expressions that jointly support prediction and structural interpretation. Representative approaches include AI Feynman, Deep Symbolic Regression, and PySR (Udrescu & Tegmark, 2020; Petersen et al., 2021; Cranmer, 2023). In multi-dataset settings, Multi-Level Symbolic Regression, parametric symbolic regression, and multi-experiment equation learning exploit shared structures, condition-dependent parameters, or joint discovery (Sen Fong & Motani, 2024; Zhang et al., 2023; Ciocanel et al., 2026).

Recent studies extend symbolic modeling to multiple environments and unseen domains. DIF extracts invariant functions from multi-environment dynamical systems (Gui et al., 2025). In drug-solubility prediction, related methods evaluate explicit equations in unseen drug domains under leave-one-domain-out designs and combine symbolic regression with mechanisms such as IRM and DANN (Alshehri & Pishnamazi, 2026; Alhowyan et al., 2026). ROKAN combines domain-robust learning with readable symbolic expressions for memory behavior in heterogeneous learning domains (Shen et al., 2025). These approaches extend cross-domain symbolic modeling to dynamical systems, scientific prediction, and learning behavior, but typically evaluate complete equations or predictors. We instead evaluate candidate families, align them through MIBD and an identity rule, and characterize their cross-dataset evidence scope.

Domain generalization and source-domain model selection. Domain generalization learns predictors for unseen distributions using only source environments. Representative methods include IRM, GroupDRO, and VREx (Arjovsky et al., 2019; Sagawa et al., 2020; Krueger et al., 2021; Zhou et al., 2022). DomainBed further demonstrates that source-only model selection materially affects domain-generalization conclusions (Gulrajani & Lopez-Paz, 2020). Existing work usually evaluates complete predictors or latent representations, yet similar predictive performance need not imply a unique structural explanation (Fisher et al., 2019). CF-DRSR therefore evaluates candidate-

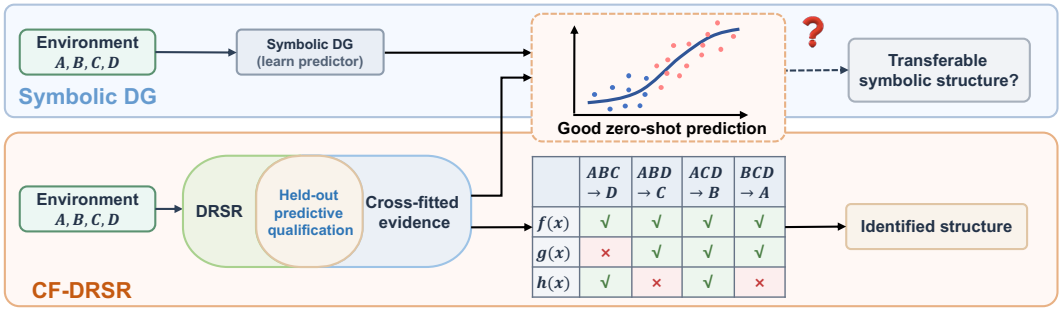

Figure 1: **Prediction does not by itself certify transferable structure.** Conventional symbolic domain generalization learns a predictor and evaluates zero-shot performance, leaving structural transfer unresolved. CF-DRSR adds held-out predictive qualification and cross-fitted evidence across outer routes, separating frozen zero-shot prediction from the identification of structural content supported across environments.

family transfer evidence separately from out-of-domain performance of the final predictor and retains explanations that current evidence cannot distinguish.

Memory prediction across datasets. Classical memory models describe subsequent recall through forgetting, activation, half-life, and skill practice, including Wickelgren, ACT-R, HLR, and DAS3H (Wickelgren, 1974; Anderson & Lebiere, 1998; Settles & Meeder, 2016; Choffin et al., 2019). Duolingo SLAM provides a common predictive task based on real language-learning logs (Settles et al., 2018). More recent methods explicitly address heterogeneity across learning environments: I-Mem learns environment-invariant representations (Shen et al., 2026), while ROKAN combines domain-robust objectives with readable expressions (Shen et al., 2025). Our focus is instead on determining across which learning datasets machine-discovered candidate structures retain predictive value, without interpreting them as universal memory laws.

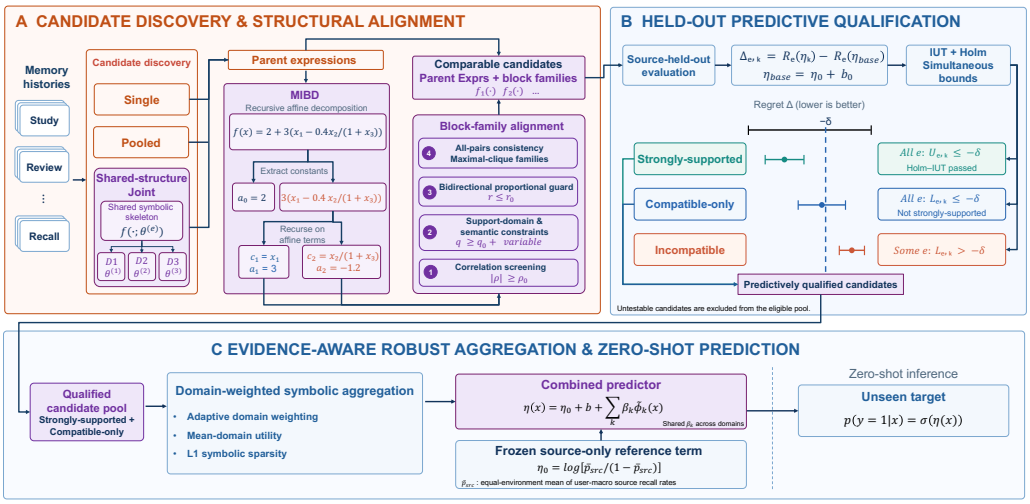

Figure 2: **CF-DRSR framework.** (A) Single, Pooled, and shared-structure Joint discovery produce parent expressions; MIBD and dual-standard alignment form comparable candidate families. (B) Source-held-out comparisons, simultaneous bounds, and multiplicity control assign transfer-evidence states. (C) Strong and Compatible families enter source-only domain-robust aggregation, after which the predictor is frozen for zero-shot evaluation on the unseen target. Structural evidence states are not revised using target outcomes.

3 CF-DRSR

CF-DRSR separates cross-environment symbolic modeling into two related problems: DRSR constructs a symbolic predictor that generalizes to an unseen environment, while cross-fitted evidence analysis determines whether candidate structures receive cross-environment support and characterizes their evidence scope and identifiability. Figure 2 shows the DRSR prediction process for a fixed target environment; cross-configuration certification and set-valued structural identification are defined in Section 3.5.

3.1 Problem Formulation

Let $\mathcal{E} = \{e_1, \ldots, e_M\}$ denote the environments under study. Environment e provides the observations $\mathcal{D}_e = \{(x_i^{(e)}, y_i^{(e)})\}_{i=1}^{n_e}$. Here, $x_i^{(e)}$ contains learning-history variables and $y_i^{(e)}$ is the corresponding recall outcome. Environment distributions may differ and need not obey the same latent symbolic relation. Given a loss ℓ on the linear predictor—log loss with logits for our binary task—the weighted empirical risk of a linear predictor f in environment e is

$$R_e(f) = \sum_{i=1}^{n_e} w_i^{(e)} \ell\Big(y_i^{(e)}, f(x_i^{(e)})\Big), \qquad w_i^{(e)} \geq 0, \qquad \sum_{i=1}^{n_e} w_i^{(e)} = 1.$$

The normalized weights $w_i^{(e)}$ are determined by the evaluation unit.

For target environment $t \in \mathcal{E}$, let $\mathcal{E}_{-t}$ denote all remaining environments used as sources, and write $\mathcal{E}_s = \mathcal{E}_{-t}$ when t is fixed. DRSR uses only $\mathcal{E}_{-t}$ to construct a frozen symbolic predictor $\widehat{\eta}^{(-t)}$ and candidate-family set $\mathcal{F}^{(-t)}$. The target environment participates only in final zero-shot evaluation, not in candidate discovery, alignment, parameter estimation, or model selection.

CF-DRSR consequently returns two objects: $\widehat{\eta}^{(-t)}$ is the prediction object for the unseen environment, whereas context-conditioned structural identities derived from $\mathcal{F}^{(-t)}$ form the structural object for transfer certification and identification. The goal is not to recover a unique true equation, but to construct an unseen-environment predictor and identify transferable structural components and evidence scopes supported by the available cross-environment evidence.

3.2 Candidate Discovery and Structural Alignment

Candidate discovery generates structural hypotheses for subsequent held-out evaluation rather than directly selecting a transferable regularity. For a fixed target t, let $\mathcal{A} \subseteq \mathcal{E}_{-t}$ denote the environments used for a discovery run. We use three complementary regimes: Single searches each environment separately; Pooled searches their combined data; and Joint searches a shared symbolic skeleton with environment-specific numerical parameters (Sen Fong & Motani, 2024; Zhang et al., 2023; Ciocanel et al., 2026). The normalized and exactly deduplicated parent-expression bank is

$$\mathcal{B}(\mathcal{A}) = \mathrm{Canonicalize}(\mathcal{B}_{\mathrm{single}}(\mathcal{A}) \cup \mathcal{B}_{\mathrm{pooled}}(\mathcal{A}) \cup \mathcal{B}_{\mathrm{joint}}(\mathcal{A})). \tag{1}$$

The three terms are candidate sets from the corresponding regimes; $\mathcal{B}(\mathcal{A})$ is a bank rather than a single optimal expression. Search interfaces, the Joint objective, and canonicalization are deferred to the supplementary method details.

Maximal Interaction Block Decomposition. To avoid fragmenting evidence across algebraically equivalent representations, MIBD recursively separates additive terms and external constant multiples while retaining nonlinear interactions that cannot be further separated:

$$f(x) = b_f + \sum_{\mu=1}^{J_f} \alpha_\mu c_\mu(x), \qquad c_\mu \text{ is a maximal interaction block.} \tag{2}$$

Here J_f is the number of blocks, b_f is the intercept, and α_μ is an external coefficient. MIBD exactly reconstructs the parent expression but does not require a unique decomposition across all algebraically equivalent representations.

Dual-standard identity alignment. CF-DRSR aligns blocks through two identity standards. The parameterized structural standard compares normalized operator trees, variables, and parameter slots. The conservative functional standard uses common-valid coverage $q_{\mu\nu}$, Pearson output correlation $\rho_{\mu\nu}$, and bidirectional zero-intercept proportional residual $r_{\mu\nu}$, compared with pre-specified thresholds q_0, ρ_0, r_0. Blocks satisfying either standard are linked; pairwise-consistent maximal cliques form the candidate-family bank $\mathcal{F}(\mathcal{A}) = \{F_1, \ldots, F_K\}$, where K is the number of families. Exact statistics, thresholds, reference-support construction, and overlap handling are given in the supplementary method details.

When $\mathcal{A} = \mathcal{E}_{-t}$, write $\mathcal{F}^{(-t)} = \mathcal{F}(\mathcal{E}_{-t})$. When source environment v is held out for qualification, let $\mathcal{A}_{-v}^{(-t)}$ denote the remaining source environments and write $\mathcal{F}^{(-t,-v)} = \mathcal{F}(\mathcal{A}_{-v}^{(-t)})$. Discovery and alignment are frozen before v is read.

3.3 Held-Out Predictive Qualification

For a fixed target t, each source environment $v \in \mathcal{E}_{-t}$ is held out in turn for qualification. Discovery, alignment, and parameter estimation use only $\mathcal{A}_{-v}^{(-t)}$; v evaluates comparisons that were frozen in advance.

To compare candidate structures fairly, we hold the remainder of the expression fixed and replace only a candidate family in the same structural position. Let C be the fixed context and a the focal family instance; the resulting complete alternative is $L = C \oplus a$. For context-matched alternatives L, L', let $\eta_L^{(-t,-v)}$ and $\eta_{L'}^{(-t,-v)}$ be predictors fitted only on $\mathcal{A}_{-v}^{(-t)}$. Their risk difference in v is $D_v(L, L') = R_v\left(\eta_L^{(-t,-v)}\right) - R_v\left(\eta_{L'}^{(-t,-v)}\right)$. Thus $D_v(L, L') < 0$ favors L. A pre-specified practical margin $\tau > 0$ and simultaneous cross-environment bounds determine the qualification state $\mathrm{state}_t(L)$, which is Strong, Compatible, Incompatible, or Untestable. Strong means stable support against all legal competitors; Compatible means no competitor excludes the alternative but at least one comparison remains unresolved; Incompatible means at least one competitor is stably better; and Untestable denotes incomplete legal comparisons. States belong to complete context-conditioned alternatives, not to isolated families. Detailed competitor construction and inference are provided in the supplementary method details.

3.4 Evidence-Aware Robust Aggregation and Zero-Shot Prediction

Qualification determines which alternatives may enter aggregation. For target t, let $\mathcal{Q}_t$ denote the pool of alternatives labeled Strong or Compatible. For each $L \in \mathcal{Q}_t$, let $\widetilde{\phi}_L(x)$ be its symbolic basis reconstructed by a frozen source-only rule.

DRSR fits a source-only reference linear predictor η_{ref}, an unpenalized intercept correction β_0, and external coefficients β_L: $\eta(x; \beta_0, \beta) = \eta_{\mathrm{ref}}(x) + \beta_0 + \sum_{L \in \mathcal{Q}_t} \beta_L \widetilde{\phi}_L(x), \qquad \beta = (\beta_L)_{L \in \mathcal{Q}_t}$. Define the environment-wise risk difference and its source-environment mean by $\Delta_e(\beta_0, \beta) = R_e(\eta(\cdot; \beta_0, \beta)) - R_e(\eta_{\mathrm{ref}}), \qquad \overline{\Delta} = |\mathcal{E}_s|^{-1} \sum_{e \in \mathcal{E}_s} \Delta_e$. The aggregation objective controls worst-environment risk, average risk change, and structural sparsity: $\mathcal{L}(\beta_0, \beta) = \mathrm{smax}_{\kappa, e \in \mathcal{E}_s} \Delta_e(\beta_0, \beta) + \overline{\Delta}(\beta_0, \beta) + \lambda_1 \|\beta\|_1$, where smax_κ is a smooth maximum with smoothing parameter κ, and λ_1 is the L_1 regularization strength.

After source-only configuration selection, DRSR refits and freezes all parameters on the complete source set:

$$\begin{aligned} \widehat{\eta}^{(-t)}(x) &= \widehat{\eta}_{\mathrm{ref}}^{(-t)}(x) + \widehat{\beta}_0 + \sum_{L \in \mathcal{Q}_t} \widehat{\beta}_L \widetilde{\phi}_L(x), \\ \widehat{p}(y = 1 \mid x) &= \sigma\left(\widehat{\eta}^{(-t)}(x)\right). \end{aligned} \tag{3}$$

The target environment is excluded from all model construction and selection.

3.5 Cross-Environment Certification and Structural Identification

Zero-shot utility of a frozen predictor does not establish stable support for each constituent symbolic structure. After repeating the procedure across outer-target configurations, we represent a context-

conditioned canonical identity as $h = (a \mid C)$ and map the state of alternative $L = C \oplus a$ to the corresponding h. Only matching canonical identities are aggregated across configurations.

Let $s_t(h)$ denote the associated Strong, Compatible, Incompatible, or Untestable state in outer-target configuration t, and define $\mathbf{s}(h) = (s_t(h))_{t \in \mathcal{E}}$. We use $\mathrm{Scope}(h)$ for the outer-target configurations in which h is Strong or Compatible. It is therefore a cross-fitted evidence scope rather than direct validation of h by the target environment.

For each configuration t, let $\mathcal{H}_t^{\mathrm{cert}}$ be the set of structural explanations satisfying the certification criterion and $\mathcal{H}_t^{\mathrm{comp}}$ the set not excluded by current evidence, with $\mathcal{H}_t^{\mathrm{cert}} \subseteq \mathcal{H}_t^{\mathrm{comp}}$. Each $S \in \mathcal{H}_t^{\mathrm{comp}}$ is a set of canonical identities h. The identified content is $\mathcal{I}_t = \bigcap_{S \in \mathcal{H}_t^{\mathrm{comp}}} S$. Thus $\mathcal{I}_t$ records the structure jointly required by current evidence rather than selecting one complete explanation. The structural output of CF-DRSR consists of the identified contents $\mathcal{I}_t$ from all outer-target configurations; no additional global intersection is imposed across routes. All certification and identification claims are conditional on the frozen candidate space, identity rules, and evaluated environments.

4 Experiments

We evaluate CF-DRSR along three dimensions: prediction in unseen domains, cross-environment structural support, and the extent to which supported structure is identified. We first validate the identity and set-valued inference procedures in controlled settings, then organize the real-data evaluation around three questions:

RQ1 — Predictive Transfer. *Can we generalize to unseen domains?*

RQ2 — Shared Structure and Transfer Boundaries. *What transfers, and where?*

RQ3 — Interpretation and Identification. *What structural content is identified?*

We conclude with ablations and robustness analyses of the main design choices and evidence boundaries.

4.1 Experimental Setup

Datasets and protocol. We use four datasets from three learning platforms: MaiMemo, Duolingo En2Es, Duolingo En2Fr, and Anki. En2Es and En2Fr involve different target languages and are treated as separate data domains, but not as independent platform-level replications. The main experiments use five variables—practice count, historical success rate, item difficulty, learner proficiency, and elapsed time—constructed only from records preceding the current event. We use nested leave-one-dataset-out evaluation. In each outer route, one dataset is the unseen target and the other three are sources; leave-one-environment-out cross-fitting is then performed within the sources. The outer target is used only for final zero-shot evaluation, and each inner pseudo-target is used only for qualification in its fold. Candidate discovery, alignment, parameter estimation, threshold setting, and model selection use only the remaining source data. Dataset statistics, feature construction, preprocessing, and the complete nested protocol are provided in the supplementary material.

Baselines and metrics. We compare against HLR, Wickelgren, ACT-R, and DAS3H; PySR and DSO; GroupDRO, VREx, IRM, ASGDRO, and I-Mem; and ROKAN. Standard ERM is also reported. Raw-feature logistic and raw-feature robust models serve as controlled baselines in Section 4.6. Every method is trained and selected using source datasets only and follows the same outer protocol. We report user-macro NLL, obtained by averaging NLL within each user and then weighting users equally, and pooled AUC over all target-dataset predictions. The main setting uses qualification margin $\tau_{\mathrm{S5}} = .005$, structural tolerance $\epsilon_{\mathrm{RQ3}} = .005$, and test level $\alpha = .05$.

4.2 Controlled Validation of CF-DRSR

Before the real-data evaluation, we construct three experiments with known ground truth to isolate structural identity alignment, held-out predictive qualification, and set-valued structural identification. Fixing the candidate space or generating laws separates downstream inference from symbolic-search error.

4.2.1 IDENTITY RECOVERY

We compare exact-syntax matching, algebraic-equivalence matching, and our dual-standard identity rule on 896 independent test pairs. The dual-standard rule attains precision, recall, and F1 of 1.000 with no false merges or misses, whereas the alternatives fail to preserve both cross-representation equivalence and semantic constraints. Construction details and class-balanced results are provided in Appendix B and Table 5.

4.2.2 CONDITIONAL HELD-OUT PREDICTIVE QUALIFICATION

The second experiment compares complete candidate laws under a shared structural context. In four environments, $f_e = A\exp(-\theta_e t) + 4h, \qquad g_e = A\exp(-\theta_e t) + 4th$, where $t \sim U(0,1)$, h is a truncated standard-normal variable, $\theta_e \in \{1,2,3,4\}$, and $A \in \{8,4,2\}$. The law is used as the logit for a binary outcome. Both share exponential decay, but f has an additive history term and g a time–history interaction. Each route uses three environments as sources and one as an unseen target. We consider a homogeneous baseline $\{f,f,f\} \to f$, mixed sources $\{f,f,g\} \to f$, and target reversal $\{f,f,f\} \to g$. The conditional test compares complete f and g laws under the same training environments, fitting rule, and shared context; the held-out environment's generating label is not used to construct competitors.

The three configurations yield 1,440 experimental units. Under mixed sources, every held-out discrimination favors the law matching that environment's generating mechanism. Under target reversal, every boundary test detects the $f \to g$ reversal. Both pass the simultaneously adjusted criterion. Oracle candidates fix competitors in advance, isolating qualification itself. Target outcomes are accessed only after source evidence is frozen.

4.2.3 SET-VALUED STRUCTURAL IDENTIFICATION

The third experiment tests certified explanations, compatible explanations, and their identified core. Across three environments, two atomic structures A and B define candidates $\varnothing$, $\{A\}$, $\{B\}$, and $\{A,B\}$; the complete explanation is the zero-loss reference. Population-sufficient candidates define the ground-truth sufficient set and its intersection the ground-truth core. We construct: (i) *required structure*, where every sufficient explanation contains A; (ii) *substitutable structure*, where either A or B is sufficient; and (iii) *low information*, whose core remains $\{A\}$ but whose smaller sample and higher noise make alternatives difficult to exclude. Each regime is repeated 200 times.

Table 1: Set-valued structural identification (200 repetitions per regime). Agreement compares estimated and ground-truth cores; expansion means that the compatible set strictly contains the certified set.

Regime	Core agreement	Empty estimated core	Compatible-set expansion	Certified / compatible count
Required structure	1.000	.000	.005	1.995 / 2.000
Substitutable structure	1.000	1.000	.010	2.990 / 3.000
Low information	.005	.995	1.000	1.000 / 3.865

The first two regimes recover $\{A\}$ and the empty core in 100% of repetitions. Under low information, the compatible set expands and the estimated core is empty in 99.5% of repetitions. This means the sample cannot establish a structure as necessary across every compatible explanation; it does not mean that the procedure declares the true core empty. Weak evidence thus expands the compatible set instead of forcing a unique answer.

4.3 RQ1 — PREDICTIVE TRANSFER: CAN WE GENERALIZE TO UNSEEN DOMAINS?

RQ1 asks whether a DRSR predictor discovered, qualified, aggregated, and frozen using source domains alone transfers to an unseen target. Table 2 reports pooled AUC and user-macro NLL.

DRSR obtains the highest average AUC (.7006) and lowest average NLL (.3601). It ranks first on both metrics for En2Es, En2Fr, and Anki; on MaiMemo it attains the lowest NLL but trails I-Mem and ROKAN in AUC. Averaging first within Duolingo and then equally across the three platforms gives DRSR AUC/NLL of $.7089/.3731$.

Table 2: Frozen prediction on four unseen target domains. Cells report pooled AUC / user-macro NLL. Avg. is the arithmetic mean over data domains; En2Es and En2Fr are both Duolingo datasets and are not independent platform replications. Best and second-best values are bold and underlined.

Method	Explicitly interpretable	DG-style method	MaiMemo	En2Es	En2Fr	Anki	Avg.
HLR (2016)	√	–	.5029/.6412	.5155/.6186	.5007/.8081	.5021/1.9104	.5053/.9946
Wickelgren	√	–	.4922/.4338	.5145/.3322	.4998/.3685	.4969/.5588	.5008/.4233
ACT-R	√	–	.4033/.3495	.5158/.3090	.4977/.3729	.5744/.5194	.4978/.3877
DAS3H	√	–	.4929/<u>.3493</u>	.5441/.3045	.5328/.3567	.5863/1.2028	.5390/.5533
ERM	–	–	.6649/.3817	.6410/<u>.2980</u>	.6350/.3460	.6342/.6156	.6438/.4101
GroupDRO	–	√	.6660/.3635	.6157/.2983	.6085/.3543	.6535/.5531	.6359/.3923
VREx	–	√	.6793/.3761	.6411/.2987	.6351/.3456	.6408/.6071	.6491/.4060
ASGDRO	–	√	.6696/.3624	.6053/.3014	.6144/.3656	.6471/.5617	.6341/.3978
IRM	–	√	.6698/.3675	<u>.6476</u>/.2982	<u>.6451</u>/<u>.3449</u>	.6646/.5151	<u>.6568</u>/<u>.3814</u>
DSO	√	–	.6631/.3685	.5995/.3110	.5989/.3513	.5152/.6087	.5942/.4099
PySR	√	–	.6601/.3822	.6163/.3016	.6334/.3458	<u>.6935</u>/<u>.4984</u>	.6509/.3820
I-Mem	–	√	**.7051**/.4231	.6159/.3106	.5955/.3661	.6464/.5312	.6407/.4078
ROKAN	√	√	<u>.6848</u>/.3816	.6171/.2992	.6249/.3510	.6127/.5445	.6349/.3941
DRSR	√	√	.6739/**.3308***	**.6650***/**.2976***	**.6868***/**.3446***	**.7769***/**.4673***	**.7006***/**.3601***

√ denotes category membership. Explicit interpretability requires an inspectable predictive equation or cognitively parameterized structure. DG-style methods include explicit cross-environment invariance, robust optimization, or source-domain transfer validation. * denotes p-value < 0.05 compared to the best-performing baseline.

4.4 RQ2 — Shared Structure and Transfer Boundaries

RQ2 asks which structures repeatedly receive support and whether that support corresponds to unseen-target utility.

Conditional qualification. Requiring a complete context-matched competitor in all three inner folds leaves one full-coverage pair on each of the MaiMemo and Anki routes. Neither comparison certifies a direction, so each retains two Compatible alternatives. En2Es and En2Fr have no full-coverage pair; remaining alternatives are Untestable. No route produces a Strong or Incompatible alternative. Conditional qualification is therefore limited mainly by cross-fold competitor coverage.

Single-structure held-out behavior and unseen-target utility. Among candidates with complete source-held-out comparisons, nine improve consistently, four remain unresolved, and two fail to reach the pre-specified margin. All consistently improving candidates occur on En2Es, En2Fr, or Anki; MaiMemo contains all unresolved and below-margin candidates. Across aligned routes, `success_rate` improves consistently on three routes but is unresolved for MaiMemo, while `user_proficiency` improves consistently elsewhere and falls below the margin for MaiMemo.

For candidate k, define unseen-target gain $G^{\text{single}}_{t,k} = R_t(\eta_{\text{base}}) - R_t(\eta_k)$, where η_k adds k with source-only parameters. All nine consistently improving candidates have positive gains with paired user-level 95% bootstrap intervals above zero. Mean gain is negative only for MaiMemo and positive for the other routes. This frozen candidate space therefore shows a descriptive correspondence between source-held-out behavior and unseen-target utility, with MaiMemo as the main boundary.

Table 3: Single-structure source-held-out behavior and unseen-target utility for the Anki route.

Candidate structure	Representation	Source-held-out behavior	G^{single}	95% interval
x_{prof}	MIBD	Consistent improvement	$+.013913$	$[+.008436, +.019466]$
x_{succ}	MIBD	Consistent improvement	$+.022468$	$[+.017942, +.027322]$
$\log(\lvert x_{\text{prof}}\rvert + 1)$	MIBD	Consistent improvement	$+.014350$	$[+.010667, +.017990]$
$\sqrt{\lvert - x_{\text{diff}} + x_{\text{succ}} + x_{\text{prof}}\rvert}$	Parent	Consistent improvement	$+.035763$	$[+.031686, +.039717]$
$\lvert x_{\text{diff}}\rvert$	Parent	Consistent improvement	$+.011900$	$[+.008016, +.015342]$

Anki case study: from discovery to a frozen predictor. The Single, Pooled, and Joint searches produce 149 complete expressions, 127 fully decomposable by MIBD. Alignment and held-out

evaluation retain five cross-fold-testable identities for aggregation. The frozen predictor is

$$\eta_{\text{Anki}}(x) = \beta_0 + \beta_1 x_{\text{prof}} + \beta_2 x_{\text{succ}} + \beta_3 \log(1+|x_{\text{prof}}|) + \beta_4 \sqrt{|-x_{\text{diff}} + x_{\text{succ}} + x_{\text{prof}}|} + \beta_5 |x_{\text{diff}}|.$$

It attains AUC .7769 and NLL .4673. The proficiency and history terms echo ACT-R's accumulation of experience and DAS3H's learner–skill history (Anderson & Lebiere, 1998; Choffin et al., 2019). The joint term is consistent with the broad idea that the value of difficulty depends on learning conditions (Bjork & Bjork, 2020), but is not interpreted as an identified psychological mechanism. En2Es and En2Fr retain proficiency and success history; MaiMemo retains item difficulty and a joint nonlinear structure.

4.5 RQ3 — Interpretation and Identification

RQ3 asks which content is required by every compatible explanation in the frozen Parent+MIBD candidate space.

Set-valued partial identification: no unique core. MaiMemo, En2Es, En2Fr, and Anki contain 40, 1, 1, and 13 certified explanations, and 54, 3, 3, and 31 compatible explanations. The compatible-set intersection is empty on all four routes, so no atomic structure is established as necessary across every non-rejected explanation. The reference-only control is rejected on every route. The empty core therefore reflects substitution among nonempty explanations, not an absence of useful symbolic structure; nor does it prove that every compatible explanation is population-sufficient.

Recurring themes beyond the strict core. `user_proficiency_online` occurs in a minimum-size certified explanation for En2Es, En2Fr, and Anki; `success_rate_online` occurs for both Duolingo datasets; and $\sqrt{|-x_{\text{diff}} + x_{\text{succ}} + x_{\text{prof}}|}$ occurs for Anki and MaiMemo. These recurring themes do not constitute a unique identified core. All conclusions remain conditional on the frozen search realization and candidate space.

4.6 Ablations and Robustness Analyses

Predictive mechanisms and sensitivity analyses. Applying the domain-robust objective to raw features improves average AUC/NLL from .6176/.6993 to .6362/.5120, but remains below the full DRSR result of .7006/.3601. Removing predictive qualification or robust aggregation reduces the average to .6925/.3749 and .6916/.3745, respectively, supporting their roles in candidate admission and parameter estimation. Candidate-space sensitivity further shows that repeating RQ3 in Parent-only, MIBD-only, and joint Parent+MIBD spaces changes the MaiMemo core: the joint space becomes empty, whereas restricted spaces retain two atoms or one parent law. This indicates that identification depends on substitution among admissible representations. Detailed ablation and sensitivity results are provided in Appendix D.

Discovery stability. Expanding the variable space increases candidate count but yields no additional stable cross-environment structure, and conclusions remain unchanged across examined evidence thresholds. Repeating source-only discovery with three independent seeds without candidate pooling gives aligned-bank Jaccard similarity of .123–.175, increasing to .494–.778 among fully testable sets after qualification. Both user proficiency online and success rate online are fully testable in all 12/12 route–seed combinations. The compatible-set intersection is empty in 11/12 cases, with one En2Es seed retaining a two-atom core, indicating that non-identification depends on the realized candidate space rather than a single stochastic search. Detailed stability results are provided in Appendix D.

5 Conclusion

CF-DRSR separates symbolic domain generalization into unseen-domain prediction, cross-environment structural evaluation, and set-valued identification. A source-only frozen DRSR predictor provides competitive zero-shot performance. Single-structure held-out behavior corresponds descriptively to unseen-target utility, but support varies by configuration and MaiMemo is the principal boundary. Every route admits nonempty certified explanations, yet compatible explanations have an empty intersection. Competitive predictive transfer can therefore coexist with bounded structural transfer and non-unique explanations.

AI USE STATEMENT

During the research process, we used generative artificial intelligence tools to assist with certain non-core tasks, including language refinement, code debugging assistance, and research workflow discussions. Generative AI tools were not used to generate experimental data, design the core algorithm, replace theoretical analysis, or produce research conclusions without author verification. All content assisted by AI tools was carefully reviewed, validated, and revised by the authors. The authors take full responsibility for the final content of this paper, including the methodology, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

We provide an anonymous code repository to facilitate the reproducibility of our results:

> Anonymous repository: `https://anonymous.4open.science/r/CF-DRSR-8B7B/`.

The repository contains the implementation of CF-DRSR, experimental configurations, model construction procedures, and the code required to reproduce the main experimental results. The main paper and appendix further describe the symbolic candidate discovery process, cross-fitting protocol, model construction procedure, evaluation metrics, and data preprocessing pipeline. For publicly available datasets, we follow the original data release protocols and provide complete descriptions of the data processing procedures. Through the released implementation and detailed experimental protocols, we aim to enable independent researchers to verify the effectiveness of the proposed method and facilitate further research.

REFERENCES

Adel A. Alhowyan, Ahmad J. Obaidullah, and Wael A. Mahdi. Physics-consistent equation discovery for drug solubility in supercritical CO2 through symbolic regression and domain generalization. *Frontiers in Medicine*, 13, 2026.

Sameer Alshehri and Mahboubeh Pishnamazi. Symbolic and domain-generalized machine learning for interpretable solubility modeling in supercritical CO2. *Scientific Reports*, 2026.

John R. Anderson and Christian J. Lebiere. *The Atomic Components of Thought*. Psychology Press, 1998.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Robert A. Bjork and Elizabeth L. Bjork. Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4):475, 2020.

Shana K. Carpenter, Steven C. Pan, and Andrew C. Butler. The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, 1(9):496–511, 2022.

Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie. DAS3H: Modeling student learning and forgetting for optimally scheduling distributed practice of skills. In *Proceedings of the 12th International Conference on Educational Data Mining*, pp. 29–38, 2019.

Maria-Veronica Ciocanel, John T. Nardini, Kevin B. Flores, Erica M. Rutter, Suzanne S. Sindi, and Alexandria Volkening. Enhancing generalizability of model discovery across parameter space with multi-experiment equation learning for biological systems. *PLOS Computational Biology*, 22(4): e1014161, 2026.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Shurui Gui, Xiner Li, and Shuiwang Ji. Discovering physics laws of dynamical systems via invariant function learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 20662–20693, 2025.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *International Conference on Machine Learning*, pp. 5815–5826, 2021.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

Kei Sen Fong and Mehul Motani. Multi-level symbolic regression: Function structure learning for multi-level data. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 2890–2898, 2024.

Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1848–1858, 2016.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 56–65, 2018.

Xiaoxuan Shen, Zhihai Hu, Di Chen, Jianwen Sun, and Shengyingjie Liu. ROKAN: Toward interpretable and domain-robust memory behavior modeling. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 2588–2598, 2025.

Xiaoxuan Shen, Zhihai Hu, Fuqing Li, Shengyingjie Liu, and Jianwen Sun. Invariant representation learning for memory behavior modeling via adaptive environment separation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):2047–2055, 2026.

Paul Smolen, Yili Zhang, and John H. Byrne. The right time to learn: Mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, 17(2):77–88, 2016.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Wayne A. Wickelgren. Single-trace fragility theory of memory dynamics. *Memory & Cognition*, 2 (4):775–780, 1974.

Michael Zhang, Samuel Kim, Peter Y. Lu, and Marin Soljačić. Deep learning and symbolic regression for discovering parametric equations. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16775–16787, 2023.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

A METHOD AND EXPERIMENTAL DETAILS

A.1 ALGORITHM AND NUMERICAL IMPLEMENTATION

Complete procedure. Given environments $\{(e, \mathcal{D}_e) : e \in \mathcal{E}\}$, candidate space $\mathcal{G}$, and frozen identity, qualification, and aggregation rules, CF-DRSR proceeds as follows.

1. For each outer target t, set $\mathcal{E}_{-t} = \mathcal{E} \setminus \{t\}$. Before the predictor is frozen, t is excluded from discovery, alignment, parameter estimation, threshold selection, and model selection.
2. For each inner pseudo-target $v \in \mathcal{E}_{-t}$, run Single, Pooled, and Joint discovery on $\mathcal{E} \setminus \{t, v\}$; canonicalize the parent bank; apply MIBD and dual-standard alignment; and freeze complete context-matched competitors before evaluating them on v.
3. Aggregate legal inner comparisons into $\mathrm{state}_t(L)$. Strong and Compatible alternatives form $\mathcal{Q}_t$. Select the aggregation configuration and refit on all sources, freeze $\widehat{\eta}^{(-t)}$, and evaluate it once on t. Target results never revise candidates or states.
4. Across outer configurations, aggregate only matching canonical identities $h = \pi(L) = (a \mid C)$ to obtain $\mathbf{s}(h)$ and $\mathrm{Scope}(h)$.
5. Within each frozen route-level candidate space, construct $\mathcal{H}_t^{\mathrm{cert}} \subseteq \mathcal{H}_t^{\mathrm{comp}}$ and return $\mathcal{I}_t = \bigcap_{S \in \mathcal{H}_t^{\mathrm{comp}}} S$. No additional global intersection is imposed.

Candidate discovery. Let $\mathrm{SR}(\mathcal{D}; \mathcal{G})$ return a candidate set. Single applies it separately to each environment, Pooled to their union, and Joint returns a shared skeleton with environment-specific parameter instances. A Joint candidate solves

$$(\widehat{f}, \widehat{\Theta}) \in \arg\min_{f \in \mathcal{G}, \Theta} \left\{ |\mathcal{A}|^{-1} \sum_{e \in \mathcal{A}} R_e(f(\cdot; \theta_e)) + \lambda_{\mathrm{par}} \Omega(f) \right\}. \tag{4}$$

Canonicalization normalizes expressions and numeric constants and removes exact duplicate representations. Neither search configuration nor ranking reads a held-out environment.

MIBD and family alignment. MIBD fails closed on unbound parameters, unknown variables, and unsupported operators. It recursively separates addition, numeric constants, and purely numeric multiplicative factors, while preserving variable-dependent products, ratios, powers, and nonlinear compositions as maximal blocks. Duplicate blocks within a parent are merged, coefficients are added exactly, zero blocks are removed, and AST-path provenance is retained. Reconstruction must be symbolically exact or have maximum absolute error at most 10^{-10} on the fold's source-only support.

For blocks μ, ν with raw output vectors v_μ, v_ν, define

$$a_{\nu \leftarrow \mu} = \frac{v_\mu^\top v_\nu}{v_\mu^\top v_\mu}, \quad r_{\nu \leftarrow \mu} = \frac{\|v_\nu - a_{\nu \leftarrow \mu} v_\mu\|_2}{\|v_\nu\|_2 + \epsilon_{\mathrm{num}}}, \quad r_{\mu\nu} = \max\{r_{\nu \leftarrow \mu}, r_{\mu \leftarrow \nu}\}. \tag{5}$$

The conservative functional path requires identical variable sets, nondegenerate outputs, at least three common valid points, and

$$|\rho_{\mu\nu}| \geq .8, \qquad q_{\mu\nu} \geq .8, \qquad r_{\mu\nu} \leq .02. \tag{6}$$

The parameterized structural path instead requires a common provenance-backed skeleton κ,

$$\kappa(c_\mu) = \kappa(c_\nu), \qquad \mathrm{vars}(c_\mu) = \mathrm{vars}(c_\nu), \qquad q_{\mu\nu} \geq .8. \tag{7}$$

Maximal cliques in the identity graph are family proposals. Overlapping bridge members are marked ambiguous rather than forced into a family. A validation-ready family must be evaluable, unambiguous, and cover at least two legal source environments.

Context-matched qualification. Alternatives $L = C \oplus a$ and $L' = C \oplus b$ are comparable only when they share training environments, context C, typed slot, fitting rule, and complexity tier. Competitors are generated mechanically from the fold-local bank and frozen before reading the pseudo-target. With practical margin $\tau_{\mathrm{S5}} > 0$, write $L \succ L'$ only when comparisons have full required coverage,

the multiplicity-adjusted intersection–union test rejects no advantage, and every simultaneous upper bound is at most $-\tau_{\mathrm{S5}}$. Then

$$\mathrm{state}(L) = \begin{cases} \text{Strong}, & \forall L' \in \mathcal{C}(L),\ L \succ L', \\ \text{Compatible}, & \nexists L' : L' \succ L \text{ and } \exists L' : L \not\succ L', \\ \text{Incompatible}, & \exists L' : L' \succ L, \\ \text{Untestable}, & \mathcal{C}(L) = \varnothing \text{ or required coverage is incomplete.} \end{cases} \tag{8}$$

Candidate-reference diagnostic and inference. This auxiliary diagnostic does not define the four qualification states. For pseudo-target user u, candidate k, and trial masses m_i, define

$$\Delta_{u,e,k} = \frac{\sum_{i\in(u,e)} m_i[\ell\{y_i, \eta_k(x_i)\} - \ell\{y_i, \eta_{\mathrm{base}}(x_i)\}]}{\sum_{i\in(u,e)} m_i}, \qquad \widehat{\Delta}_{e,k} = |\mathcal{U}_e|^{-1} \sum_u \Delta_{u,e,k}. \tag{9}$$

Whole users are resampled. Candidates within an environment share Rademacher multipliers; the maximum absolute studentized deviation over candidate–environment contrasts supplies simultaneous bounds, followed by Holm correction.

Candidate-set refitting and identification. For route t, enumerate nonempty $F \subseteq K_t$. In each pseudo-target fold, both F and the full reference K_t are refitted from scratch on legal sources. If $L_{u,e}(F)$ is user-level weighted NLL, define

$$\Gamma_{u,e}(F) = L_{u,e}(F) - L_{u,e}(K_t), \qquad \widehat{\Gamma}_e(F) = |\mathcal{U}_e|^{-1} \sum_u \Gamma_{u,e}(F). \tag{10}$$

All $F \times e$ contrasts within a route share $B = 2000$ Rademacher multiplier draws. With simultaneous bounds $L_e(F), U_e(F)$, $\epsilon_{\mathrm{RQ3}} = .005$, and $\alpha = .05$,

$$\max_e U_e(F) \le \epsilon_{\mathrm{RQ3}} \Rightarrow F \in \mathcal{H}_t^{\mathrm{cert}}, \qquad \max_e L_e(F) \le \epsilon_{\mathrm{RQ3}} \Rightarrow F \in \mathcal{H}_t^{\mathrm{comp}}. \tag{11}$$

Failed or missing required-fold fits are insufficiently testable; otherwise a set is rejected when its maximum lower bound exceeds the tolerance.

Source-only aggregation. Each fold reconstructs and standardizes bases using legal sources only. The convex frozen-pool problem is optimized by proximal gradient with backtracking and monotone/adaptive restart. The intercept is unpenalized and symbolic coefficients use ordinary L_1 soft thresholding:

$$\beta^{(m+1)} = \mathrm{SoftThreshold}_{a_m\lambda_1}\left[\beta^{(m)} - a_m \nabla_\beta \mathcal{L}_{\mathrm{smooth}}(b^{(m)}, \beta^{(m)})\right]. \tag{12}$$

Safe configurations satisfy $\Delta_e - \mathrm{SE}_e \le 0$ in every fold; this is a permissive selection state, not certification. Ties favor fewer nonzero structures, lower AST burden, lower regret, and frozen grid order.

A.2 Datasets, Features, and Temporal Preprocessing

Table 4: Real datasets and frozen evaluation sizes. En2Es and En2Fr are evaluated separately but both belong to Duolingo.

Dataset	Platform/task	All records	Test records	Test users
MaiMemo	MaiMemo	400,735	25,900	1,045
En2Es	Duolingo Spanish	399,998	83,968	6,030
En2Fr	Duolingo French	399,990	84,707	3,728
Anki	Anki	722,753	139,731	1,000

All datasets are converted to a common event schema containing domain, split, user, item, timestamp, recall rate, and trial mass. Features use only strict history before query i: log historical user–item observation count; Beta$(1,1)$-smoothed user–item success rate; item difficulty as negative smoothed logit; learner proficiency as smoothed logit; and $\log(1+\Delta t)$. Duplicate domain–user–item–timestamp rows are aggregated. Records sharing a user timestamp are queried before any of them updates history. Anki within-day jitter that only encodes original row order is removed. Whole timestamp batches are split chronologically into approximately 70/10/20 percent train/validation/test sets. Validation and test queries may use earlier events available online, never current or future responses.

A.3 Nested Evaluation and Comparison Settings

Each outer route leaves one dataset completely unseen. The three sources rotate as inner pseudo-targets; a pseudo-target is used only for held-out evaluation, never candidate construction, alignment, or parameter fitting. After source-side qualification and configuration selection, the model is refitted on all sources, frozen, and evaluated once on the outer target. Single, Pooled, and Joint use the same grammar, variables, complexity ceiling, and matched search budgets. Raw-feature logistic minimizes regularized mean risk; raw-feature robust uses the same robust-plus-mean objective as DRSR. All baseline hyperparameters are source-selected.

B Supplementary Controlled-Validation Design

Identity recovery. Matching rules are calibrated on 224 pairs and evaluated on a disjoint 896-pair test set. Seven allowed transformations and seven hard-negative classes yield 448 positive and 448 negative pairs, spanning cross-representation equivalence and conflicts in semantics, parameter slots, or structural operations.

Table 5: Structural identity recovery on the synthetic test set of 896 pairs. Macro-F1 weights the 14 construction classes equally.

Matching rule	Precision	Recall	F1	Macro-F1	False merges	Misses	Hard-negative FPR
Exact syntax	.000	.000	.000	.357	128	448	.286
Algebraic equivalence	.778	1.000	.875	.857	128	0	.286
Dual-standard identity	1.000	1.000	1.000	1.000	0	0	.000

Conditional qualification. Each environment contains 500 users with four observations each. Three generating configurations are crossed with two input granularities, three signal strengths, four target configurations, and 20 seeds, giving 1,440 units. All 18 mixed-source held-out discrimination tasks pass the simultaneously corrected .005 threshold; the minimum simultaneous lower bound is .02827. All 18 target-reversal boundary tests detect the prescribed $f \to g$ reversal; the minimum simultaneous lower bound is .06407.

Set-valued identification. For the candidates $\varnothing, \{A\}, \{B\}, \{A, B\}$, the required-structure mean contrasts are $.030, 0, .030$, the true-alternatives contrasts are $.030, 0, 0$, and both use 300 users per environment with noise SD .015. The low-information regime uses $.012, 0, .012$, 20 users, and noise SD .080; its population core remains $\{A\}$. Each regime has 200 Monte Carlo repetitions and 1,000 multiplier max-t bootstrap draws.

C Supplementary RQ2 and RQ3 Results

C.1 All Single-Structure Results

C.2 Frozen Route Equations

The full standardized Anki predictor is

$$\begin{aligned}\eta_{\mathrm{Anki}}(x) = 1.6480 &+ 0.4491\frac{x_{\mathrm{prof}} - 1.0980}{1.1617} + 0.0133\frac{x_{\mathrm{succ}} - 0.6165}{0.1803} \\ &- 0.2622\frac{\log(1 + |x_{\mathrm{prof}}|) - 0.7275}{0.4485} + 0.5257\frac{\sqrt{|-x_{\mathrm{diff}} + x_{\mathrm{succ}} + x_{\mathrm{prof}}|} - 1.7289}{0.6103} \\ &+ 0.0640\frac{|x_{\mathrm{diff}}| - 1.5818}{0.9088}. \end{aligned} \tag{13}$$

The Duolingo predictors are

$$\eta_{\mathrm{En2Es}}(x) = 1.5317 + 0.7383\frac{x_{\mathrm{prof}} - 1.1652}{1.2516} + 0.1484\frac{x_{\mathrm{succ}} - 0.6072}{0.1757}, \tag{14}$$

$$\eta_{\mathrm{En2Fr}}(x) = 1.5552 + 0.7270\frac{x_{\mathrm{prof}} - 1.1710}{1.2585} + 0.1625\frac{x_{\mathrm{succ}} - 0.6062}{0.1754}. \tag{15}$$

Table 6: Source-held-out behavior and unseen-target gain for all 15 fully testable candidates.

Target	Candidate	Held-out behavior	G^{single}	95% interval
MaiMemo	x_{diff}	Unresolved	$-.019934$	$[-.024958, -.014783]$
MaiMemo	x_{prof}	Below margin	$+.000736$	$[-.003615, +.005533]$
MaiMemo	x_{succ}	Unresolved	$+.012036$	$[+.008561, +.015857]$
MaiMemo	$\sqrt{\lvert -x_{\text{diff}} + x_{\text{succ}} + x_{\text{prof}}\rvert}$	Unresolved	$-.027126$	$[-.034952, -.018983]$
MaiMemo	$\lvert x_{\text{diff}}\rvert$	Unresolved	$-.016435$	$[-.019649, -.013199]$
MaiMemo	$x_{\text{succ}} + \sqrt{\lvert x_{\text{diff}}\rvert}$	Below margin	$-.004436$	$[-.007077, -.001340]$
En2Es	x_{prof}	Consistent improvement	$+.047550$	$[+.045601, +.049428]$
En2Es	x_{succ}	Consistent improvement	$+.055372$	$[+.052427, +.058291]$
En2Fr	x_{prof}	Consistent improvement	$+.039395$	$[+.036380, +.042483]$
En2Fr	x_{succ}	Consistent improvement	$+.036554$	$[+.032723, +.040187]$
Anki	x_{prof}	Consistent improvement	$+.013913$	$[+.008436, +.019466]$
Anki	x_{succ}	Consistent improvement	$+.022468$	$[+.017942, +.027322]$
Anki	$\log(\lvert x_{\text{prof}}\rvert + 1)$	Consistent improvement	$+.014350$	$[+.010667, +.017990]$
Anki	$\sqrt{\lvert -x_{\text{diff}} + x_{\text{succ}} + x_{\text{prof}}\rvert}$	Consistent improvement	$+.035763$	$[+.031686, +.039717]$
Anki	$\lvert x_{\text{diff}}\rvert$	Consistent improvement	$+.011900$	$[+.008016, +.015342]$

They obtain AUC/NLL $.6650/.2976$ and $.6868/.3446$, respectively. For MaiMemo, the main artifact fixes the structure but does not expose directly reportable coefficients. With source-frozen centering and scaling,

$$\eta_{\text{MaiMemo}}(x) = 1.9075 + \beta_{\text{d}} z_{x_{\text{diff}}}(x) + \beta_{\text{n}} z_{\sqrt{\lvert -x_{\text{diff}} + x_{\text{succ}} + x_{\text{prof}}\rvert}}(x), \tag{16}$$

where both coefficients are source-estimated and frozen. This predictor obtains $.6739/.3308$. A separate fixed four-candidate-pool sensitivity yields

$$\eta^{\text{fix}}_{\text{MaiMemo}}(x) = 1.9435 - .0649 z_{\text{d}}(x) + .4456 z_{\text{n}}(x), \tag{17}$$

with AUC/NLL $.6515/.3579$; it is supplementary and must not replace the main result.

C.3 Candidate-Representation Sensitivity

Table 7: Candidate-representation sensitivity in the main variable space.

Outer	Space	Atoms	Sets	Testable	Certified	Compatible	Rejected	Core size
MaiMemo	Parent+MIBD	6	63	63	40	54	9	0
MaiMemo	MIBD-only	3	7	7	2	2	5	2
MaiMemo	Parent-only	3	7	7	3	4	3	1
En2Es	Parent+MIBD	2	3	3	1	3	0	0
En2Es	MIBD-only	2	3	3	1	3	0	0
En2Fr	Parent+MIBD	2	3	3	1	3	0	0
En2Fr	MIBD-only	2	3	3	1	3	0	0
Anki	Parent+MIBD	5	31	31	13	31	0	0
Anki	MIBD-only	3	7	7	2	7	0	0
Anki	Parent-only	2	3	3	2	3	0	0

The reference-only control is rejected on every route; simultaneous lower bounds are $.0339$, $.0558$, $.0668$, and $.0469$ for MaiMemo, En2Es, En2Fr, and Anki, all exceeding $.005$.

D Supplementary Ablations and Robustness

Predictive mechanisms: controlled comparisons and core ablations. Table 8 reports the per-domain results underlying the main-text summary.

Conditional contribution. Zeroing one structure at a time in the frozen joint predictors yields seven positive and three negative conditional contributions. En2Fr proficiency and success-rate contributions are negative jointly despite positive standalone target gains, illustrating redundancy or suppression.

Table 8: Controlled comparisons and core ablations of predictive mechanisms. Each cell reports pooled AUC / user-macro NLL; higher AUC and lower NLL indicate better performance. The best value in each column is highlighted in bold.

Method	MaiMemo	En2Es	En2Fr	Anki	Avg.
Raw-feature logistic	.6606/.3539	.6489/.3071	.6576/.3492	.5034/1.7870	.6176/.6993
Raw-feature robust	.6641/.3614	.6609/.2984	.6717/.3474	.5482/1.0408	.6362/.5120
w/o predictive qualification	.6546/.3658	.6634/.2996	.6852/.3467	.7669/.4873	.6925/.3749
w/o robust aggregation	.6546/.3578	.6573/.3187	.6860/**.3444**	.7684/.4769	.6916/.3745
Full	**.6739**/**.3308**	**.6650**/**.2976**	**.6868**/.3446	**.7769**/**.4673**	**.7006**/**.3601**

Expanded variable space. Adding six I-Mem-guided history variables increases candidate availability but leaves only 1–3 testable candidates per route and produces no Strong candidate. Except for one Incompatible En2Es candidate, testable candidates are Compatible. A richer interface expands discovery without strengthening cross-environment qualification.

Threshold sensitivity. For $\tau_{\mathrm{S5}} \in \{0, .0025, .005, .0075, .01\}$, the nine Strong candidates on En2Es, En2Fr, and Anki are unchanged; MaiMemo Compatible counts decrease from six to one as the margin grows. Increasing ϵ_{RQ3} from 0 to .01 increases total certified/compatible explanation counts from $9/61$ to $68/97$ and can shrink the identified core by admitting more alternatives. For $\alpha \in \{.025, .05, .10\}$, every route retains an empty main-space core.

Search-seed stability. Across individual seed pairs among seeds 9201, 9202, and 9203, route-level aligned-bank Jaccards range from .100 to .200 and testable-set Jaccards from .333 to .833. MaiMemo, En2Fr, and Anki have empty cores under every seed; En2Es is empty for seeds 9202 and 9203, while seed 9201 retains the two-atom core {`item_difficulty_online`, `user_proficiency_online`}.