COMPLEMENTARY PAVE SPECIFICATIONS FOR CHARACTERIZING AND IDENTIFYING LLM CAPABILITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter Vector (PAVE) *specifications* encode each candidate large language model (LLM) and each task as fine-tuning updates of a common auxiliary model, enabling LLM selection without evaluating candidates on the target task. Extending PAVE beyond models derived from a common base, we find that auxiliary models provide complementary views of model–task compatibility: the same behavioral observations can yield accurate model comparisons in one parameter space but misleading comparisons in another. To connect these views to actual model performance, we prove that, for a fixed auxiliary model and under explicit assumptions, the probability that two tasks agree on a model pair's true performance ordering is non-decreasing in their nonnegative PAVE cosine similarity, up to a quantified approximation error. This result motivates Reference-Guided Multi-Auxiliary Voting (RG-MAV), which uses evaluated reference tasks that are most informative for the target in each auxiliary model's parameter space to weight and combine model comparisons separately for each candidate pair. Extensive experiments on heterogeneous LLMs and target tasks from six domains support the benefit of reference locality and show robustness to specification-training seeds and individual auxiliary-model removal. RG-MAV achieves an average Kendall's τ_b of 0.677, which exceeds the 0.517 of a fixed best auxiliary model selected on reference tasks and the 0.600 of EmbedLLM, and reaches over 96% of the 0.704 attained by a post-hoc task-wise oracle. These results demonstrate that RG-MAV can effectively select and combine accurate views, and that reusable parameter-space characterizations, guided by reference-task evidence, can support effective model selection before any candidate is run on the target task.

1 INTRODUCTION

Selecting a large language model (LLM) for a new task requires identifying which candidate's capabilities match the task requirements, rather than simply choosing the model with the highest overall benchmark score (Ethayarajh & Jurafsky, 2021; Burnell et al., 2023). Independently developed LLMs differ in training and specialization, and models with similar aggregate performance can still exhibit different strengths and errors (Kim et al., 2025). Evaluating every candidate on target-task data provides direct evidence, but requires running the entire candidate pool whenever a new task arrives and may also require exposing target data to remotely accessed models. An alternative is to characterize model behavior in advance and reuse these characterizations for new tasks. The central challenge is whether such characterizations preserve the behavioral distinctions that determine relative model performance without observing candidate behavior on the target task.

The *learnware* paradigm (Zhou, 2016; Zhou & Tan, 2024; Zhou, 2026) approaches this problem by associating trained models with reusable specifications. Parameter Vector (PAVE) specifications (Tan et al., 2025; Shi et al., 2026) construct a common space for model behavior and task requirements through a shared auxiliary model[1]. A model PAVE is obtained by fine-tuning the auxiliary on that model's responses to shared probe inputs, while a task PAVE is obtained from labeled task examples. Cosine similarities between these vectors provide model–task compatibility scores, allowing cached model PAVEs to be reused without running candidates on the target task. Prior studies

[1]For brevity, we sometimes refer to the auxiliary model simply as the auxiliary.

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.

have demonstrated the promise of this approach mainly within collections derived from shared base models. Extending it to independently developed LLMs raises a fundamental question: when does similarity in an auxiliary model's parameter space preserve the behavioral distinctions that determine which candidate performs better on a new task?

We investigate this question by extending PAVE to heterogeneous LLMs and varying the auxiliary model while keeping the candidate probe responses and task data unchanged. Across 13 auxiliary models and 22 tasks, we observe complementary ranking behavior: a parameter space that yields inaccurate rankings on some tasks can perform well on others, where a different space is less effective (Section 2). Thus, the auxiliary model is noninterchangeable: it induces a particular *characterization view*, making some behaviorally important distinctions easier to recover than others. The key question is therefore how to use a shared set of characterization spaces to make task-specific model comparisons, without using target-task evaluations to determine which space to trust.

Under a linearized approximation of fine-tuning, we can intuitively show that the auxiliary model's gradients determine how behavioral differences are weighted in parameter space. A natural approach would be to define the characterization space directly as a function of these gradients. In practice, however, this is infeasible: the gradients depend on many interacting factors, which makes them difficult to analyze. We address this problem from another angle by aiming at a more essential problem first, i.e. when a reference task, on which candidate performances are already known, provides evidence about a new task. Therefore for a fixed auxiliary model and candidate pair, we study the probability that two tasks agree on the pair's true performance given their task-PAVE cosine similarity. Under explicit population assumptions, we derive a controlled approximation showing that this probability is approximately nondecreasing over nonnegative similarities, with a quantified approximation error (Section 3). The result connects task similarity in parameter space to the transfer of pairwise ranking evidence, motivating reference-task neighborhoods defined separately in each auxiliary model's own space.

Building on this insight, we propose *reference-guided multi-auxiliary voting* (RG-MAV). The system maintains model PAVEs, reference-task PAVEs, and candidate evaluation scores on the reference tasks. For each candidate model pair on a new task, RG-MAV retrieves relevant references in each auxiliary space, taking into account both task similarity and the orientation of the pairwise comparison. The observed reference task orderings then guide the weighting of each auxiliary model's pairwise decision, and the weighted decisions are aggregated into a complete ranking using Copeland scoring (Section 4). Importantly, the same auxiliary pool and cached model PAVEs are reused across tasks; what changes is how their evidence is combined for each comparison.

We evaluate RG-MAV with 16 heterogeneous LLMs on 11 target tasks spanning six domains, using 11 disjoint tasks as references. Its average Kendall's τ_b reaches 0.677, compared with 0.517 for a fixed best auxiliary model selected on reference tasks, 0.590 for uniform voting, and 0.600 for EmbedLLM. On five target tasks, RG-MAV even surpasses the post-hoc task-wise oracle, demonstrating that our method can provide benefits beyond task-wise auxiliary selection. Its mean Top-1 regret is only 0.0158, lower than all deployable baselines, showing that RG-MAV remains competitive even when evaluation focuses on selecting the best candidate rather than recovering the full ranking. Component ablations on RG-MAV show that reference locality, orientation factor and pair-specific weighting are all indispensable, while additional experiments examine robustness to training seeds and individual auxiliary-model removal. Together, these results show that complementary parameter-space characterizations, guided by reference-task evidence, can support robust model selection without access to candidate behavior on the target task.

Our contributions are threefold:

- We investigate PAVE-based capability characterization for heterogeneous LLMs and identify complementary ranking information across auxiliary-model parameter spaces under controlled behavioral observations (Section 2).
- We establish a conditional theoretical relationship between task-PAVE similarity and agreement on true pairwise performance orderings, with a spectral characterization and controlled approximation bounds that motivate local reference-task evidence (Section 3).
- We develop RG-MAV to exploit this evidence separately for each candidate pair, improving both model-ranking quality and top-1 selection regret without evaluating candidate models on the target task during selection (Section 4).

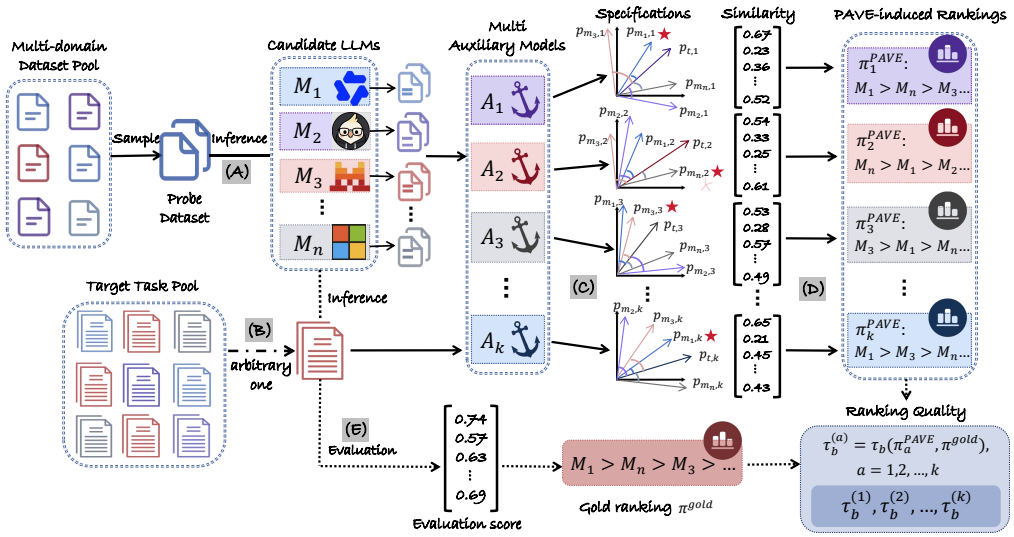


Figure 1: **Identifying learnwares with PAVE by multi auxiliaries.** **(A)** LLMs infer on the probe dataset to produce model-specific pseudo datasets. **(B)** A target task is randomly selected for illustration. **(C)** Fine-tuning each auxiliary on pseudo datasets and the target task yields the corresponding model and task specifications. **(D)** Cosine similarities induce a model ranking per auxiliary. **(E)** Evaluating LLMs on the target task is used only to assess the quality of PAVE-induced rankings.

2 Discovering Complementarity

In this section, we show through extensive experiments that auxiliary models exhibit complementary ranking behavior, and that simple properties such as auxiliary model scale or post-training domain do not reliably predict the tasks on which an auxiliary performs well.

2.1 Experiment Setup

Ranking Evaluation Metric. We use Kendall's τ_b (Kendall, 1945; 1938) to measure learnware identification performance. Given two rankings over n items, τ_b is defined as

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where n_c and n_d are the numbers of concordant and discordant pairs, $n_0 = \binom{n}{2}$, n_1 and n_2 are the numbers of pairs tied in the first and second ranking, respectively. It ranges from -1 (perfect disagreement) to 1 (perfect agreement), with 0 indicating no association. We see this as an improvement of the performance measure as it aggregates over all candidate pairs rather than the top one alone, capturing whether the ranking remains reliable throughout. It is also a strict standard, as $\tau_b = 0$ indicates that concordant and discordant pairs are equally frequent among non-tied comparisons.

Datasets. We empirically study complementarity on 22 multiple-choice tasks across finance, healthcare, math, legal, code, and science. The probe dataset for generating model PAVE is sampled from datasets spanning six domains and is disjoint from the data of previous 22 tasks. For brevity, the full dataset list and descriptions are deferred to Appendix B.2.

Models. Unlike previous PAVE studies (Shi et al., 2026; Tan et al., 2025), we experiment directly on open-weight, heterogeneous LLMs from Hugging Face without any further fine-tuning, and obtain PAVEs on the filtered generations of those LLMs. The pool of 16 LLMs includes general-purpose models and specialists in healthcare, code, and legal tasks, spanning dense and mixture-of-experts architectures and ranging from 7B to 35B parameters. For brevity, the full model list and descriptions are deferred to Appendix B.3. To ensure fair comparison , we control inference hyperparameters in detail, as listed in Appendix B.6. We observe a drop in ranking performance when LLMs produce messy or explanatory responses (Appendix C), so we filter such outputs by sending the generation to DeepSeek-v4-Flash-0731 (DeepSeek-AI, 2026) to extract a clean choice (Yu et al., 2025).

Table 1: Kendall's τ_b between the evaluation-induced and PAVE-induced model rankings.

Benchmark	Q3 1.7B	Q3 8B	Q3.5 2B	Q3.5 9B	OLMo3 Inst.	OLMo3 Math	L3.2 3B	L3.1 8B	L3 8B	Medi Phi	Phi 3.5	Granite 4.2	Ornith 1.5	Max.
AQuA-RAT	0.80	0.68	0.60	0.78	0.58	0.65	0.68	0.77	0.77	0.57	0.50	**0.83**	**0.83**	0.83
Bhasha-Fin	0.58	0.52	0.55	0.68	0.48	0.55	0.73	0.45	0.70	0.72	0.73	0.73	**0.75**	0.75
Bhasha-Legal	0.03	0.70	0.58	0.62	0.63	0.67	0.60	**0.77**	0.52	0.67	0.53	**0.77**	0.65	0.77
CaseHOLD	0.53	0.77	0.73	0.83	0.78	0.70	0.80	0.77	0.72	0.75	0.57	**0.88**	0.68	0.88
CodeMMLU-Exec	0.63	0.25	0.65	0.42	**0.78**	0.73	0.67	0.55	0.57	0.75	0.75	0.32	0.45	0.78
CodeMMLU-Syntax	0.38	0.43	0.62	**0.77**	0.48	0.65	0.53	0.60	0.52	0.60	0.53	**0.77**	0.52	0.77
ContractNLI-A	0.57	0.52	0.73	0.55	0.77	0.77	0.63	0.68	0.45	**0.83**	0.73	0.63	0.68	0.83
FLARE-Causal20	0.13	0.30	0.13	0.42	0.38	0.28	0.43	0.05	0.18	**0.62**	0.42	0.12	0.27	0.62
HEAD-QA v2	0.50	0.70	0.58	0.62	0.42	0.62	0.62	0.73	0.60	0.63	0.72	**0.77**	**0.77**	0.77
MathQA	0.77	0.65	**0.80**	0.73	0.48	0.65	0.72	0.72	0.67	0.60	0.53	0.68	0.70	0.80
MedMCQA	0.50	0.62	0.63	0.68	0.62	0.55	0.67	0.70	0.57	0.65	0.63	**0.85**	0.80	0.85
PubMedQA	0.23	0.54	0.19	0.54	0.36	0.33	0.14	0.28	0.21	0.54	**0.66**	0.49	0.43	0.66
QASC	0.45	0.55	0.63	0.63	0.35	0.58	**0.70**	0.63	0.63	0.38	0.37	**0.70**	0.57	0.70
SciKnowEval	0.02	0.33	0.48	0.55	0.45	0.50	0.43	0.22	0.15	**0.63**	0.52	0.60	0.62	0.63
Twitter-FinSent	0.67	0.78	0.67	0.62	0.70	0.65	**0.82**	0.73	0.43	0.67	0.42	0.58	0.70	0.82
CyberMetric	-0.23	0.52	0.50	0.53	0.40	0.52	0.57	0.48	0.43	**0.62**	0.57	**0.62**	**0.62**	0.62
ElemMathQA	0.53	**0.58**	0.42	**0.58**	0.43	0.45	0.45	0.50	0.33	0.50	0.43	0.47	0.53	0.58
FLARE-Headlines	0.30	0.42	0.48	0.38	0.30	0.42	0.35	0.47	0.22	**0.60**	0.28	0.20	0.32	0.60
Indian-Legal	0.18	0.63	0.65	0.68	0.67	0.65	0.73	**0.80**	0.60	0.73	0.68	0.73	0.62	0.80
Math-MCQA	0.45	0.00	0.27	**0.57**	0.23	0.17	0.12	0.05	0.20	0.17	0.22	0.43	0.45	0.57
MedQA-USMLE	0.43	0.69	0.41	**0.71**	0.41	0.49	0.56	0.61	0.48	0.69	0.66	0.68	0.69	0.71
SciQ	0.60	0.70	0.48	0.48	0.33	0.70	**0.77**	0.72	0.55	0.62	0.55	0.68	0.70	0.77
Average	0.41	0.54	0.54	0.61	0.50	0.56	0.58	0.56	0.48	0.62	0.55	0.62	0.61	0.73

Note: Auxiliary and benchmark names are abbreviated for compactness. Q and L denote Qwen and Llama.

Auxiliary Models. We use 13 auxiliary models spanning the Llama, Qwen, OLMo, Phi, Granite, and Ornith families, with diverse scales, training stages, and post-training domains. The pool supports both within-family comparisons and contrasts between general-purpose and domain-specialized auxiliaries. Full descriptions are provided in Appendix B.1. For fine-tuning, we largely follow the configurations of LoRA Without Regret (Schulman & Thinking Machines Lab, 2025), without deliberate hyperparameter sweeps, as in Table 6.

2.2 Results

Table 1 summarizes the results from the complementarity experiment. We use **bold** to indicate the highest Kendall's τ_b on each task and underline the lowest of each auxiliary. We can see that every auxiliary performs catastrophically on some tasks where another may perform well. We call this phenomenon **PAVE complementarity** of auxiliaries: the tasks on which each auxiliary works well vary significantly, and the characterization spaces of the auxiliaries complement each other. This phenomenon holds when training seed changes, as discussed in Appendix D.1.

The results do not support simple heuristics based on auxiliary model scale or post-training domain specialization. Within Qwen3, scaling from 1.7B to 8.2B increases the average Kendall's τ_b from 0.41 to 0.54, yet the larger auxiliary performs substantially worse on several tasks, including AQuA-RAT (0.68 vs. 0.80), MathQA (0.65 vs. 0.77), and CodeMMLU Execution Prediction (0.25 vs. 0.63). The same pattern appears within Qwen3.5: the 9B auxiliary improves the average from 0.54 to 0.61, but trails the 2B variant on CodeMMLU Execution Prediction (0.42 vs. 0.65) and ContractNLI-A (0.55 vs. 0.73). Domain specialization is similarly inconsistent. Relative to OLMo-3-7B-Instruct, the math-specialized OLMo-3-7B-RL-Zero-Math improves on MathQA (0.65 vs. 0.48) but degrades on Math-MCQA (0.17 vs. 0.23). Likewise, MediPhi-Instruct outperforms its general-purpose Phi-3.5-mini-Instruct counterpart on MedMCQA (0.65 vs. 0.63) and MedQA-USMLE (0.69 vs. 0.66), while underperforming it on HEAD-QA v2 (0.63 vs. 0.72) and PubMedQA (0.54 vs. 0.66). From these comparisons, we conclude that auxiliary effectiveness is fundamentally task dependent.

The mean Kendall's τ_b of individual auxiliaries ranges from 0.41 to 0.62, whereas the average task-wise maximum reaches 0.73 (Table 1). This gap suggests the potential to select and combine useful views from the auxiliary models for task-specific model comparisons, yet with current PAVE theories we are still unable to predict which view is trustworthy, which the next section deals with.

3 THEORETICAL ANALYSIS

Unlike weight disentanglement in task arithmetic, the relationship between cosine similarity of PAVEs and model-task reuse performance has not been established. By introducing reference tasks, we first prove that the cosine similarity between PAVEs implicitly suggests the model-task performance rank under two population assumptions. We formalize a clear hypersphere mathematical model for PAVE, which then supports proving the probability that two tasks agree on the ordering of a model pair is approximately nondecreasing over nonnegative similarities, with a quantified approximation error. These theoretical results lay the foundation for developing RG-MAV.

3.1 PAVE HYPERSPHERES

Projection onto a hypersphere. PAVE identification ranks models by cosine similarity with a task PAVE, so only directions matter. For nonzero PAVEs $\boldsymbol{p}_i, \boldsymbol{p}_j$, normalization gives

$$\boldsymbol{q}_i = \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|}, \qquad \boldsymbol{q}_j = \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}, \qquad \cos(\boldsymbol{p}_i, \boldsymbol{p}_j) = \langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle.$$

In our problem, we fine-tune each auxiliary with LoRA, A matrices frozen, and flatten the B matrices across target modules. If this vector has dimension d_a for auxiliary a, its normalized PAVEs lie on a hypersphere $\mathbb{S}^{d_a-1}$. We denote normalized model and task PAVEs by $\boldsymbol{q}_{a,i}$[2] and $\boldsymbol{q}_{a,t}$, respectively.

Definition 3.1 (Model Difference Vector and Decision Boundary)**.** Fix an auxiliary a and n distinct normalized model PAVEs. For model indices $i < j$, define

$$\boldsymbol{d}_{a,i,j} = \boldsymbol{q}_{a,i} - \boldsymbol{q}_{a,j}, \qquad \mathbb{H}_{a,i,j} = \{x \in \mathbb{S}^{d_a-1} : \langle x, \boldsymbol{d}_{a,i,j} \rangle = 0\}.$$

In the following analysis, we fix the model PAVEs and treat the task PAVEs as variables. A model pair (i, j) determines a decision boundary $\mathbb{H}_{a,i,j}$ by their difference vector $\boldsymbol{d}_{a,i,j}$, in the sense that if the inner product of the normalized task PAVE and the difference vector is positive, then the auxiliary gives a rank $i > j$; if negative, the auxiliary gives $i < j$. That said, the two sides of the decision boundary represent two cases of model ranking, respectively.

A learning-dynamics view. Under linearized fine-tuning, a PAVE is the label residual projected through the auxiliary's gradients, $\boldsymbol{p}_{a,D} \approx \eta T |D|^{-1} \sum_{x \in D} \boldsymbol{J}_a(x)^\top (\boldsymbol{e}_{y(x)} - \pi_a(x))$, where π_a is the auxiliary's predictive distribution. Each pairwise decision is then effectively a vote over the probe instances on which the two candidates disagree, weighted by the auxiliary's tangent kernel $\boldsymbol{K}_a(z, x) = \boldsymbol{J}_a(z)\boldsymbol{J}_a(x)^\top$ with the target instances z. Different auxiliaries therefore amplify or wash out different behavioural differences between the candidates.

3.2 CROSS-TASK AGREEMENT ON PAIRWISE ORDERS

Setup. Fix an auxiliary a with $d = d_a \geq 3$ and a model pair (i, j), and suppress these indices. For a task t with normalized task PAVE $\boldsymbol{q}_t = \boldsymbol{q}_{a,t}$, the order label $\sigma(t) = \mathrm{sign}(s_t(i) - s_t(j)) \in \{-1, +1\}$ records which model has the higher evaluation score. We assume no ties. For two tasks t, t' drawn from a task population, we call $\rho = \langle \boldsymbol{q}_t, \boldsymbol{q}_{t'} \rangle$ their alignment and study the agreement probability

$$p(\rho) = \Pr[\sigma(t) = \sigma(t') \mid \rho].$$

Here the probability is defined for every $\rho \in [-1, 1]$ (Appendix F.3).

Assumption (I) (Isotropic exchangeable task geometry). For every $g \in O(d)$, $(g\boldsymbol{q}_t, g\boldsymbol{q}_{t'}) \overset{\mathrm{d}}{=} (\boldsymbol{q}_t, \boldsymbol{q}_{t'})$, where $\overset{\mathrm{d}}{=}$ denotes equality in distribution.

Assumption (CI) (Mediation and conditional independence). There exists a measurable $m\colon \mathbb{S}^{d-1} \to [-1, 1]$ such that, for all $u, v \in \{-1, +1\}$, $\Pr[\sigma(t) = u, \sigma(t') = v \mid \boldsymbol{q}_t, \boldsymbol{q}_{t'}] = \frac{1}{4}(1 + u\, m(\boldsymbol{q}_t))(1 + v\, m(\boldsymbol{q}_{t'}))$ almost surely.

[2] $\boldsymbol{q}_{a,j}$ also denotes model PAVE. The model indices i, j here are not to be confused with the i, j above, which represent arbitrary PAVEs.

Table 2: Estimated approximation error ε of Lemma 3.2 by auxiliary, in units of 10^{-9}.

	Q3 1.7B	Q3 8B	Q3.5 2B	Q3.5 9B	OLMo3 Inst.	OLMo3 Math	L3.2 3B	L3.1 8B	L3 8B	Medi Phi	Phi 3.5	Granite 4.2	Ornith 1.5
Median	3.2	1.2	1.2	0	0	0.83	1.0	0.12	0.65	1.3	1.4	0	0
Max	29	11	11	4.3	14	14	7.1	3.7	3.7	19	19	3.3	4.3

Note. Auxiliary names are abbreviated as in Table 1. Statistics are taken over all $\binom{16}{2} = 120$ model pairs.

Assumption (I) states that the task population has no preferred direction on the auxiliary's hypersphere, so only the relative position of two tasks matters in terms of distribution. It implies that each normalized task PAVE is uniform on the hypersphere and we denote this law by ς. Assumption (CI) states that, given both task PAVEs, the two order labels are independent, and each depends only on its own PAVE through a common function $m(x) = \mathbb{E}[\sigma(t) \mid \boldsymbol{q}_t = x]$, the expected order label given the task PAVE, which we call the order propensity. Both assumptions are population-level idealizations (Appendix F.8).

Let $\delta = \mathbb{E}[\sigma(t)]$ be the ordering bias and $R = \mathbb{E}[m(\boldsymbol{q}_t)^2]$. Then $R - \delta^2 = \mathrm{Var}(m(\boldsymbol{q}_t))$ is the part of the order variation explained by the auxiliary's characterization space. Our goal is to study the property of the agreement probability. Intuitively, we analyze the agreement probability by expanding it, like a Fourier series, with power series of ρ up to a quantified and empirically small error. Formally, let $\Pi_k m$ be the degree-k spherical-harmonic component of m, where degree 0 is constant and degree 1 is linear, and let $W_k = \|\Pi_k m\|_2^2$, with the norm in $L^2(\varsigma)$, be the energy of m at degree k. For intuition, W_k can be read simply as nonnegative weights that sum to R. The following Lemma shows how we formally expand the agreement probability and the exact upper bound of the error term. Proof is deferred to Appendix F.

Lemma 3.2 (Power-series form of the agreement probability)**.** *Under* (I) *and* (CI)*,* $W_k \geq 0$*,* $W_0 = \delta^2$*, and* $\sum_k W_k = R \leq 1$*. Let* $p^\sharp(\rho) = \frac{1}{2}\big(1 + \sum_{k\geq 0} W_k \rho^k\big)$ *for* $\rho \in [-1, 1]$*, with* $0^0 = 1$*. Then* $p(1) = p^\sharp(1) = (1 + R)/2$*. If, in addition,* $\mathcal{E}_2 = \sum_k k(k-1)W_k < \infty$*, then*

$$\sup_{\rho\in[-1,1]} |p(\rho) - p^\sharp(\rho)| \leq \varepsilon := \frac{\mathcal{E}_2}{2(d-2)}.$$

The key property is that all coefficients are nonnegative. $W_0 = \delta^2$ shows the agreement due to ordering bias alone, and $R - \delta^2$ is the agreement due to the auxiliary's characterization space. The error ε is small when the energy of m lies on degrees $k \ll \sqrt{d}$, and d_a, the number of LoRA-B coordinates, is very large. For each auxiliary and model pair, we fit the expansion to the observed agreement between task pairs, choosing the degree by cross-validation. The resulting estimates satisfy $\varepsilon < 3 \times 10^{-8}$ throughout (Table 2). If the degree is not too large, the error term 2ε in Theorem 3.3 is equally negligible, and the monotonicity of $p^\sharp$ carries over to p essentially unchanged.

Theorem 3.3 (Agreement is approximately nondecreasing in nonnegative alignment)**.** *Assume* (I) *and* (CI). (i) $p^\sharp$ *is nondecreasing and convex on* $[0, 1]$*. It is strictly increasing if and only if* m *is not constant* ς*-almost everywhere.* (ii) *If* $\mathcal{E}_2 < \infty$*, then for all* $0 \leq \rho \leq \rho' \leq 1$*,*

$$p(\rho') - p(\rho) \geq p^\sharp(\rho') - p^\sharp(\rho) - 2\varepsilon \geq -2\varepsilon.$$

The proof is given in Appendix F.6. At $\rho = 0$, only the constant term $W_0 = \delta^2$ remains, so two tasks with orthogonal PAVEs agree with probability about $(1 + \delta^2)/2$. This reflects ordering bias alone: if one model wins on most tasks, any two tasks tend to agree. At $\rho = 1$ agreement reaches $(1 + R)/2$. The total gain, $(R - \delta^2)/2$, is thus the part of the order variation explained by the auxiliary's geometry. Since the terms with $k \geq 2$ grow mainly near $\rho = 1$, agreement rises fastest at high alignment, so highly aligned tasks are especially informative about each other. The rise is strict unless every W_k with $k \geq 1$ is zero, that is, unless the order propensity is constant.

We admit that Theorem 3.3 relies on two strong population assumptions. Nevertheless, in the evaluation results and ablation studies, RG-MAV in the following section, which is mainly guided by the monotonicity conclusion we proved, achieves great performance, supporting the practical value of our theory. Furthermore, this is the first theory that establishes the connection, though implicit, between cosine similarity of PAVEs and the model-task performance.

4 RG-MAV: LEVERAGING COMPLEMENTARITY FOR ROBUSTNESS

In this section, we turn the complementary PAVE views induced by different auxiliary models into a practical method for model selection. For a new target, *Reference-Guided Multi-Auxiliary Voting* (RG-MAV) uses informative, previously evaluated reference tasks to decide which views should be trusted more for each model comparison, and combines them into a final ranking without evaluating candidates on the target. We then evaluate its performance on model ranking and Top-1 selection, and conduct ablations of components and aggregation.

4.1 METHODOLOGY

Let $\mathcal{A} = \{A_1, \ldots, A_l\}$ be the set of auxiliary models, $\mathcal{M} = \{M_1, \ldots, M_n\}$ the set of candidate models, and $\mathcal{R} = \{R_1, \ldots, R_m\}$ an internal reference task pool maintained by the system, with evaluation score $s_r(i)$ available for each model M_i on each reference R_r. Following Section 3, let $\boldsymbol{q}_{a,i}$, $\boldsymbol{q}_{a,r}$, and $\boldsymbol{q}_{a,t}$ denote the normalized PAVEs of model M_i, reference task R_r, and target task t under A_a. Below, we fix a target task t and candidate pair (M_i, M_j), and suppress these indices.

Define difference between normalized PAVEs $\boldsymbol{d}_a$ under A_a, the evaluation order label $\sigma(r)$ on R_r, and the pairwise decision $v_a(x)$ of A_a on x as

$$\boldsymbol{d}_a = \boldsymbol{q}_{a,i} - \boldsymbol{q}_{a,j} \qquad \sigma(r) = \operatorname{sign}(s_r(i) - s_r(j)) \qquad v_a(x) = \operatorname{sign}(\langle \boldsymbol{q}_{a,x}, \boldsymbol{d}_a \rangle), x \in \{r, t\}.$$

A_a prefers M_i to M_j on x when $v_a(x) > 0$.

Historical pairwise correctness. We define auxiliary A_a's historical correctness $c_{a,r}$ on R_r by

$$c_{a,r} = \frac{1}{2}\left[1 + \sigma(r) v_a(r)\right].$$

This measures agreement with the evaluation-induced ordering, where 1 for a correct strict preference, 0 for an incorrect one, and $\frac{1}{2}$ when either ordering is tied.

Pair-specific reference neighborhood. Within each auxiliary model's PAVE geometry, we measure proximity of t and R_r by $\rho_{a,r} = \langle \boldsymbol{q}_{a,t}, \boldsymbol{q}_{a,r} \rangle$, use an orientation factor $\alpha_{a,r}$ to reverse this proximity's sign for opposite target and reference decisions, then select the k references with the largest adjusted proximities to t:

$$\alpha_{a,r} = 1 - 2\mathbb{I}[v_a(t) v_a(r) < 0], \qquad \mathcal{N}_{a,k} = \operatorname{TopK}_{R_r \in \mathcal{R}} \left[\alpha_{a,r} \rho_{a,r}\right],$$

with $k = 1$ as the default most-local setting.

Weighted pairwise voting. First align historical correctness with the target decision using $\tilde{c}_{a,r} = \frac{1}{2} + \alpha_{a,r}(c_{a,r} - \frac{1}{2})$, which retains $c_{a,r}$ when $\alpha_{a,r} = 1$ and replaces it with $1 - c_{a,r}$ otherwise. The neighborhood average determines each auxiliary model's weight in the target task pairwise voting:

$$w_a = \frac{1}{|\mathcal{N}_{a,k}|} \sum_{R_r \in \mathcal{N}_{a,k}} \tilde{c}_{a,r}, \qquad V = \sum_{a=1}^{l} w_a \, v_a(t).$$

We predict $M_i \succ_t M_j$ if $V > 0$, $M_j \succ_t M_i$ if $V < 0$, and a tie otherwise. After repeating this procedure for all $\binom{n}{2}$ candidate pairs, we use *Copeland scoring* (Copeland, 1951; Hosseini et al., 2024) to obtain the final ranking. The full RG-MAV pseudocode is provided in Appendix E.

4.2 EVALUATION RESULTS

Evaluation protocol. We evaluate RG-MAV on 11 target tasks spanning six domains: MathQA, Math-MCQA, MedQA-USMLE, PubMedQA, Bhasha-Legal, FLARE-Headlines, FLARE-Causal20, CodeMMLU-Syntax, CodeMMLU-Exec, ContractNLI-A, and SciKnowEval. The remaining 11 tasks in the 22-task pool form a disjoint, system-maintained reference pool; full task information and citations are provided in Appendix B.2. Evaluation scores on the reference tasks are available for estimating auxiliary reliability, whereas target-task scores are withheld during ranking inference and used only afterward to construct the ground-truth model rankings for evaluation. We compare RG-MAV with Single Best Auxiliary (SBA), Uniform Multi-Auxiliary Voting (U-MAV), Prompt-Embedding Auxiliary Selection (PE-AS), MODEL SPIDER, and EmbedLLM, with

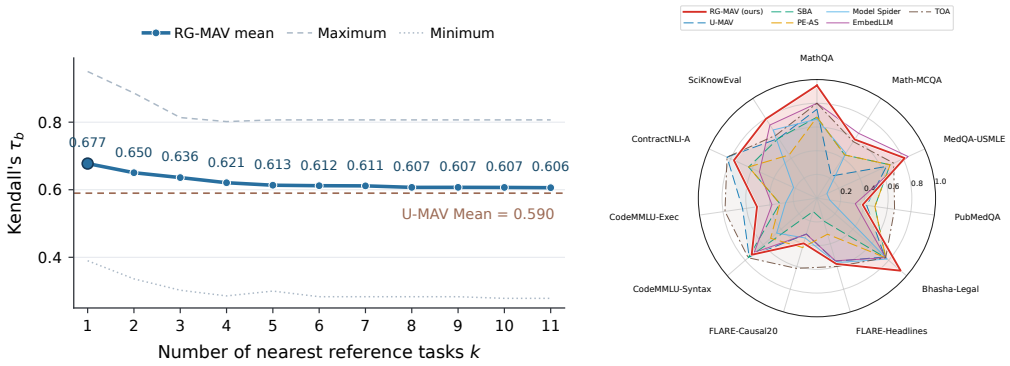


Figure 2: (a) Effect of the reference-neighborhood size k on RG-MAV. (b) Radar visualization of per-task Kendall's τ_b for the model-ranking methods.

Table 3: Comparison of model-selection methods across targets under two evaluation criteria. The left block reports Kendall's τ_b between the evaluation-induced and predicted model rankings, and the right reports Top-1 regret. The best result in each row is bolded.

	Kendall's τ_b ↑							**Top-1 Regret ↓**						
Target task	RG-MAV	U-MAV	SBA	PE-AS	Model Spider	Embed LLM	TOA	RG-MAV	U-MAV	SBA	PE-AS	Model Spider	Embed LLM	TOA
MathQA	**0.950**	0.750	0.683	0.683	0.667	0.800	0.800	**0.000**	0.014	0.014	0.014	0.110	**0.000**	0.014
Math-MCQA	0.588	0.222	0.433	0.433	0.450	**0.650**	0.567	**0.000**	0.074	0.074	0.074	0.032	**0.000**	**0.000**
MedQA-USMLE	0.814	0.644	0.678	0.678	0.092	**0.845**	0.711	0.002	0.135	0.135	0.135	0.143	**0.000**	**0.000**
PubMedQA	0.390	0.424	0.494	0.494	0.109	0.326	**0.661**	0.003	0.003	0.024	0.024	0.070	0.003	0.003
Bhasha-Legal	**0.933**	0.756	0.767	0.767	0.783	0.767	0.767	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.001	**0.000**
FLARE-Headlines	0.577	0.555	0.200	0.317	0.567	0.550	**0.600**	0.026	0.026	0.090	0.092	0.010	0.090	**0.008**
FLARE-Causal20	0.397	0.315	0.117	0.433	0.350	0.317	**0.617**	0.099	0.076	0.233	0.184	0.099	0.110	**0.016**
CodeMMLU-Syntax	0.728	0.756	**0.767**	0.517	0.450	0.700	**0.767**	0.007	0.007	0.030	0.007	**0.000**	0.007	0.007
CodeMMLU-Exec	0.510	0.633	0.317	0.317	0.267	0.383	**0.783**	0.024	0.006	0.024	0.024	**0.000**	0.024	0.006
ContractNLI-A	0.769	**0.833**	0.633	0.633	0.217	0.533	**0.833**	0.012	**0.000**	0.036	0.036	0.059	0.025	**0.000**
SciKnowEval	**0.795**	0.599	0.600	0.433	0.683	0.733	0.633	**0.000**	0.027	**0.000**	0.027	**0.000**	**0.000**	0.027
Avg.	0.677	0.590	0.517	0.519	0.421	0.600	**0.704**	0.016	0.033	0.060	0.056	0.047	0.024	**0.007**

Task-wise Oracle Auxiliary (TOA) serving as a post-hoc reference for target-wise single auxiliary selection. Detailed baseline definitions and implementation protocols are provided in Appendix B.5.

Effect of reference locality. To examine sensitivity to reference locality, we vary the neighborhood size k from 1 to 11. As shown in Figure 2a, the default nearest-reference setting ($k = 1$) achieves an average Kendall's τ_b of 0.677. Performance drops as less-similar references are added and then stabilizes around 0.60, close to the uniform multi-auxiliary voting (U-MAV) baseline; the maximum across target tasks also decreases with k, while the minimum shows no corresponding improvement. This trend is consistent with Section 3.2, where pairwise-order agreement increases with task PAVE similarity for nonnegative alignment. All following comparisons use the default $k = 1$ setting.

Comparison results. As shown in Table 3, RG-MAV achieves the strongest average ranking performance among all deployable methods, with an average Kendall's τ_b of 0.677. It improves over U-MAV by 0.088 and outperforms it on 7 of 11 target tasks, showing the benefit of reference-guided reliability weighting over uniform multi-auxiliary aggregation. RG-MAV also improves over the single-auxiliary baselines SBA and PE-AS by 0.160 and 0.159 on average, respectively, and exceeds the learned model-ranking baselines Model Spider and EmbedLLM by 0.256 and 0.077. Although the post-hoc TOA attains a higher average τ_b of 0.704, RG-MAV surpasses it on 5 target tasks, showing that aggregating complementary pairwise decisions can even outperform target-wise oracle auxiliary. The same advantage extends to Top-1 model selection: RG-MAV achieves an average regret of 0.016, lower than all deployable baselines, with only the post-hoc TOA attaining a lower value. This indicate that RG-MAV's gains in ranking quality also translate to identifying the top candidate, rather than arising only from improved ordering among lower-ranked candidates.

Table 4: Summary of component ablations and robustness analyses.

Setting	Avg. Kendall's τ_b
Complete RG-MAV	**0.677**
Without pair-specific weighting	0.598
Without reference locality	0.606
Without orientation factor	0.658
Seeds 3407 / 42 / 1337†	0.666 / 0.658 / 0.687
Leave-one-auxiliary-out	0.667–0.685

Note. †The seed experiment uses the same controlled subset of 8 auxiliaries under all three seeds; its absolute values therefore differ from the main 13 auxiliary result. The leave-one-auxiliary-out range reports the averages obtained after removing each of the 13 auxiliaries in turn.

Interpretation. When $k = 1$, RG-MAV reduces to pairwise auxiliary selection. For each candidate pair, each auxiliary's nearest reference under orientation-adjusted proximity determines its aligned correctness: 0 excludes the auxiliary, 1 selects it, and $1/2$ retains a neutral weight. The retained auxiliaries then vote on the corresponding model pair on the target task.

4.3 Ablations Study

Beyond the main comparison, we examine whether RG-MAV's gains depend on its individual design components, stochastic PAVE construction, or any particular auxiliary. We summarize the main findings in Table 4; complete results are provided in Appendix D.3, D.1, and D.4.

Component contributions. Removing pair-specific weighting causes the largest degradation, reducing the average Kendall's τ_b from 0.677 to 0.598, removing reference locality decreases to 0.606, while removing the orientation factor yields a smaller reduction to 0.658. Thus, pair-specific weighting and reference locality provide the largest contributions, with orientation supplying an additional complementary signal. The full definitions and detailed results are reported in Appendix D.3.

Robustness of aggregation. Individual PAVE-induced rankings can change noticeably with the fine-tuning seed, as shown in Fig. 5; relative to seed 3407, the mean absolute change in auxiliary–task Kendall's τ_b is 0.083 for seed 42 and 0.092 for seed 1337. Nevertheless, the complementary structure remains stable: across the three seeds, the gap between the task-wise best auxiliary and the best fixed auxiliary remains between 0.100 and 0.110, and every pair among the eight controlled auxiliaries reverses its relative performance on at least one pair of tasks. Moreover, the leave-one-auxiliary-out results in Fig. 7 range only from 0.667 to 0.685, with a maximum decrease of 0.011 from the complete method. These results indicate that RG-MAV does not rely on a particular stochastic PAVE realization or any single auxiliary. Detailed seed statistics are given in Tables 8 and 9, with the complete leave-one-auxiliary-out analysis in Appendix D.4.

Additional analyses are deferred to Appendix D.2, which shows that auxiliary complementarity also persists when evaluation focuses only on Top-1 regret, and Appendix D.5, which evaluates RG-MAV under alternative maintained reference pools. Detailed per-task results for varying the reference-neighborhood size are provided in Appendix D.6.

5 Conclusion

We extend PAVE-based capability characterization to heterogeneous LLMs, where each auxiliary model induces its own characterization view. Empirically, each auxiliary fails on some tasks where another succeeds, and neither scale nor domain specialization predicts which. Theoretically, we prove that, for a fixed auxiliary and under explicit assumptions, the probability that two tasks agree on a model pair's true performance ordering is approximately nondecreasing in their nonnegative PAVE similarity. Building on this, RG-MAV weights each auxiliary's pairwise decisions by its correctness on the target's nearest reference tasks in that auxiliary's own space, outperforming deployable baselines on average without running any candidate on the target task. Our work suggests treating the auxiliary as a variable rather than a fixed design choice, a step toward reusable capability characterization for open, heterogeneous model ecosystems.

AI USE STATEMENT

We used generative AI tools to assist with LaTeX typesetting of author-developed proofs, provide feedback on research methodology and experiments, refine research hypotheses, implement a small portion of the baseline code, and extract option labels from model outputs during data processing. We did not use these tools to develop theoretical models or conceptual frameworks beyond the feedback above, formulate mathematical claims, supply the key arguments underlying the proofs, implement our proposed method, interpret results, or translate text. Synthetic dataset generation and qualitative or thematic data analysis were not applicable to this work. We also used AI tools for manuscript polishing, utility scripts for data downloading and figure plotting, literature searches, and locating library documentation. We reviewed all AI-assisted material, manually checked generated code and LaTeX, audited extracted labels against the original model outputs, and obtained bibliographic entries directly from original sources. The authors take responsibility for the final content of this work, including AI-assisted text, claims, and artifacts.

ETHICS STATEMENT

This work studies model selection in the context of learnware identification: ranking candidate LLMs via parameter-space specifications instead of exhaustive evaluation. It involves no human subjects, user data, or new data collection. All candidate models and auxiliary models are publicly released open-weight checkpoints, used for research in accordance with their licenses (including Apache-2.0, MIT, the Llama Community Licenses, CC-BY-NC-4.0, and model-specific terms from Google and NVIDIA). All benchmarks and probe datasets are publicly available and cited to their original sources. Some datasets contain healthcare, legal, and financial content, including publicly released court judgments; we use them only for evaluation and specification construction, and make no attempt to identify individuals. A third-party LLM API is used solely to extract answer labels from model outputs. Because PAVE-induced rankings are validated against benchmark scores, they inherit the coverage and cultural biases of those benchmarks, several of which center on specific jurisdictions, such as Indian legal and financial examinations, US medical licensing, and US case law. A highly ranked model may therefore not suit users in other contexts.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide anonymized source code for PAVE construction, output filtering, RG-MAV, and all baselines in the supplementary material, together with the canonicalized probe labels and evaluation scores.[3] All candidate models, auxiliary models, benchmarks, and probe datasets are publicly available and are described in Appendices B.1, B.2, and B.3, with subsets, splits, and roles listed in Table 5. We describe the full PAVE pipeline in Appendix B.4, and report all hyperparameters and random seeds in Table 6 (Appendix B.6). The output-filtering procedure, including the exact extractor prompt, is given in Appendix C.2. We specify RG-MAV in Section 4.1 and Algorithm 1 (Appendix E), and baseline implementations are detailed in Appendix B.5. Seed sensitivity, ablations, and robustness analyses are reported in Appendices D.1, D.3, D.4, D.5, and D.6. All theoretical assumptions are stated in Section 3.2, and complete proofs are provided in Appendix F. Computational resources are reported in Appendix B.7.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv:2404.14219, 2024a.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. Technical Report MSR-TR-2024-57, Microsoft Research, 2024b.

[3] https://anonymous.4open.science/r/paper-anon-CEC4

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6430–6439, 2019.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, 2019.

Kendall Atkinson and Weimin Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044 of *Lecture Notes in Mathematics*. 2012.

Stella Biderman. Math-mcqa: A multiple choice adaptation of the math dataset. Hugging Face dataset, 2025.

BIG-bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138, 2023.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. SaulLM-7B: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

Arthur H. Copeland. A reasonable social welfare function. Seminar on Applications of Mathematics to the Social Sciences, University of Michigan, 1951.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19352–19374, 2025.

Alexis Correa-Guillén, Carlos Gómez-Rodríguez, and David Vilares. Head-qa v2: Expanding a healthcare benchmark for reasoning. arXiv:2511.15355, 2025.

Thomas M. Cover. The number of linearly inducible orderings of points in d-space. *SIAM Journal on Applied Mathematics*, 15(2):434–439, 1967.

Rafael M. O. Cruz, Robert Sabourin, and George D. C. Cavalcanti. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. arXiv:2606.19348, 2026.

Vijay Devane, Mohd Nauman, Bhargav Patel, Aniket Mahendra Wakchoure, Yogeshkumar Sant, Shyam Pawar, Viraj Thakur, Ananya Godse, Sunil Patra, Neha Maurya, et al. Bhashabench v1: A comprehensive benchmark for the quadrant of indic domains. arXiv:2510.25409, 2025.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *Regional Conference Series in Probability and Statistics*. 1989.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboards. arXiv:2009.13888, 2021.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. arXiv:2406.09098, 2024.

Gerald B. Folland. *A Course in Abstract Harmonic Analysis*. 2nd edition, 2016.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. The aloe family recipe for open and specialized healthcare llms. arXiv:2505.04388, 2025.

Tilmann Gneiting. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4): 1327–1349, 2013.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

Granite Team, IBM. Granite 4.2 llms: How they're built, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv:2407.21783, 2024.

Hadi Hosseini, Debmalya Mandal, and Amrit Puhan. The surprising effectiveness of sp voting with partial preferences. arXiv:2406.00870, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Kush Juvekar, Arghya Bhattacharya, Sai Khadloya, and Utkarsh Saxena. Are LLMs court-ready? evaluating frontier models on Indian legal reasoning. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pp. 359–369, 2025.

Olav Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. 2nd edition, 2002.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*, 2020.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. arXiv:2506.07962, 2025.

Marko Kojic, Ivan Bondyrev, Aral de Moor, Joseph Shtok, Petr Borovlev, Kseniia Lysaniuk, Madeeswaran Kannan, Ivan Dolgov, and Nikita Pavlichenko. Mellum2 technical report. *arXiv preprint arXiv:2605.31268*, 2026.

Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, 2021.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025.

Pin-Jie Lin, Miaoran Zhang, Marius Mosbach, and Dietrich Klakow. Exploring the effectiveness and consistency of task selection in intermediate-task transfer learning. arXiv:2407.16245, 2024.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, 2017.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pp. 23–32, 2020.

Meta. Llama 3.2 model card, 2024.

Mistral AI. Mistral-Small-3.2-24B-Instruct-2506 model card. Hugging Face model card, 2025.

Claus Müller. *Spherical Harmonics*, volume 17 of *Lecture Notes in Mathematics*. 1966.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7294–7305, 2020.

Dung Nguyen, Thang Phan, Nam Le Hai, Thong Doan, Nam Nguyen, Quang Pham, and Nghi Bui. Codemmlu: A multi-task benchmark for assessing code understanding & reasoning capabilities of codellms. In *Proceedings of the 13th International Conference on Learning Representations*, pp. 2614–2672, 2025.

NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2512.20848*, 2025.

Ryan O'Donnell. *Analysis of Boolean Functions*. 2014.

Ornith Team. Ornith-1.5: From self-scaffolding to self-improvement, 2026.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260, 2022.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10585–10605, 2021.

Qwen Team. Qwen3-Coder-30B-A3B-Instruct model card. Hugging Face model card, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026a.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, 2026b.

Isaac J. Schoenberg. Positive definite functions on spheres. *Duke Mathematical Journal*, 9(1): 96–108, 1942.

John Schulman and Thinking Machines Lab. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025.

Andrew Sellergren et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Hao-Yu Shi, Zhi-Hao Tan, Zi-Chen Zhao, Yang Yu, and Zhi-Hua Zhou. A study on PAVE specification for learnware. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*, volume 32 of *Princeton Mathematical Series*. 1971.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. 4th edition, 1975.

Zhi-Hao Tan, Zi-Chen Zhao, Hao-Yu Shi, Xin-Yu Zhang, Peng Tan, Yang Yu, and Zhi-Hua Zhou. Learnware of language models: Specialized small language models can do big. arXiv:2505.13425, 2025.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. arXiv:2512.13961, 2025.

Norbert Tihanyi, Mohamed Amine Ferrag, Ridhi Jain, Tamas Bisztray, and Merouane Debbah. Cybermetric: A benchmark dataset based on retrieval-augmented generation for evaluating llms in cybersecurity knowledge. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 296–302, 2024.

tvall43. Qwen3.6-14B-A3B-FableVibes model card. Hugging Face community model card, 2026.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. 2018.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5039–5059, 2022.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025.

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, et al. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models. arXiv:2512.13607, 2026.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Kevin Woods, W. Philip Kegelmeyer, and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (4):405–410, 1997.

Xi-Zhu Wu, Wenkai Xu, Song Liu, and Zhi-Hua Zhou. Model reuse with reduced kernel mean embedding specification. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):699–710, 2023.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54104–54132, 2024.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. arXiv:2306.05443, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. The finben: An holistic financial benchmark for large language models. arXiv:2402.12659, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv:2505.09388, 2025.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12133–12143, 2021.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xfinder: Large language models as automated evaluators for reliable evaluation. arXiv:2405.11874, 2025.

Z.ai. GLM-4.7-Flash model card. Hugging Face model card, 2026.

ZeroShot. Twitter Financial News Sentiment. Hugging Face dataset, 2022.

Yi-Kai Zhang, Ting-Ji Huang, Yao-Xiang Ding, De-Chuan Zhan, and Han-Jia Ye. Model spider: Learning to rank pre-trained models efficiently. In *Advances in Neural Information Processing Systems 36*, pp. 13692–13719, 2023.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 159–168, 2021.

Zhi-Hua Zhou. Learnware: on the future of machine learning. *Frontiers of Computer Science*, 10 (4):589–590, 2016.

Zhi-Hua Zhou. Learnware and ai model management system, 2026.

Zhi-Hua Zhou and Zhi-Hao Tan. Learnware: small models do big. *Science China Information Sciences*, 67(1):112102, 2024.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. EmbedLLM: Learning compact representations of large language models. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

A Related Work

Learnware specifications and task representations. Learnware identifies reusable models through specifications of their capabilities (Zhou, 2016; Zhou & Tan, 2024). Distribution-based specifications such as RKME (Wu et al., 2023) summarize training data; PAVE instead represents a model and a task by the parameter displacements obtained when a shared pretrained network is fine-tuned on each (Tan et al., 2025; Shi et al., 2026). The prior LLM study uses specialized models derived from a small number of base models and constructs specifications with one fixed network (Tan et al., 2025). We study independently released LLMs and vary this network, which we call the *auxiliary model*. Candidate models need not share weights: each is represented through its generations on a common probe dataset. Related representations also derive task information from a fixed network, including Fisher-information embeddings (Achille et al., 2019), soft-prompt embeddings (Vu et al., 2022), and projected LoRA gradients (Xia et al., 2024). Their applications differ, but together they motivate an unresolved question for PAVE: how much does the chosen network determine the usefulness of the resulting geometry?

Model ranking from historical tasks. Transferability estimators such as LEEP and LogME score candidate models using their predictions or features on target data (Nguyen et al., 2020; You et al., 2021). Historical-evaluation methods reduce that requirement. MODEL SPIDER learns task–model compatibility from known rankings (Zhang et al., 2023), while EmbedLLM learns model representations from question-level correctness records to forecast performance (Zhuang et al., 2025). RG-MAV also uses scored reference tasks, but uses them to estimate the reliability of PAVE rankings rather than to learn candidate-model scores directly. Once model specifications have been constructed on the shared probe dataset, ranking a new task requires its task PAVE and no candidate-model inference on that task. This distinction matters when the candidate pool is heterogeneous and the target-task evaluation budget is limited.

Local competence and auxiliary geometry. Dynamic classifier selection estimates a predictor's competence from nearby validation examples (Woods et al., 1997; Cruz et al., 2018). Our unit of locality is a *task*: for each candidate-model pair, RG-MAV consults scored reference tasks near the target in each auxiliary's own PAVE space, then weights that auxiliary's pairwise vote by its historical correctness. The geometry explains why one shared notion of task similarity need not suffice. Pairwise differences between model PAVEs define hyperplane boundaries; their chambers encode the rankings an auxiliary can produce, connecting our construction to linear orderings and hyperplane arrangements (Cover, 1967). Under the paper's population assumptions, spherical-harmonic analysis expresses cross-task ordering agreement through nonnegative Gegenbauer coefficients (Müller, 1966; Schoenberg, 1942), analogous to a noise-stability profile (O'Donnell, 2014). The controlled power-series approximation in Section 3.2 supplies the near-monotonicity result that motivates nearby references; nonnegative Gegenbauer coefficients alone do not establish it. Finally, Copeland scoring (Copeland, 1951) converts the weighted pairwise decisions into a ranking even when their majority relation is intransitive.

B Experimental Settings

This appendix provides the experimental details underlying our study. We first describe the auxiliary pool, datasets, and candidate models (Appendices B.1, B.2, B.3), and then detail the end-to-end PAVE ranking pipeline and comparison methods (Appendices B.4, B.5). Finally, we report the unified experimental configuration and computational resources (Appendices B.6, B.7).

B.1 Auxiliary Description

The auxiliary pool comprises 13 models spanning different families, scales from 1.7B to 9B, architectures, and post-training strategies, providing diverse views of candidate–task relationships. The pool also contains several informative within-family contrasts in model scale, domain specialization, and post-training objective, particularly within the Qwen, Phi, Olmo, and Llama families. These complementary characteristics yield heterogeneous PAVE geometries and allow us to examine how auxiliary properties affect learnware identification. Detailed descriptions of auxiliaries are provided below.

- **Phi-3.5-mini-instruct** is Microsoft's 3.8B-parameter general-purpose instruction model (Abdin et al., 2024a). Its relatively small scale and distinct training recipe provide a lightweight general-purpose auxiliary and a direct counterpart to MediPhi-Instruct.
- **MediPhi-Instruct** is a 3.8B-parameter clinical language model derived from Phi-3.5-mini-instruct and further adapted to medical data (Corbeil et al., 2025). Together with Phi-3.5-mini-instruct, it provides a useful comparison between general-purpose and domain-specialized auxiliary models.
- **Olmo-3-7B-Instruct** is Ai2's 7B general-purpose instruction model from the fully open Olmo 3 family (Team Olmo et al., 2025). It provides a general-purpose Olmo auxiliary and forms a same-family comparison with Olmo-3-7B-RL-Zero-Math.
- **Olmo-3-7B-RL-Zero-Math** is a 7B Olmo 3 checkpoint specialized for mathematical reasoning through reinforcement learning with verifiable rewards (Team Olmo et al., 2025). Its specialization provides a useful contrast with the general one at the same scale.
- **Qwen3-1.7B** is a compact 1.7B-parameter dense model from the Qwen3 family (Yang et al., 2025). As the smallest auxiliary in our pool, it provides a lightweight endpoint for examining the effect of auxiliary scale.
- **Qwen3-8B** is an 8.2B-parameter dense model from the same Qwen3 family (Yang et al., 2025). Together with Qwen3-1.7B, it provides a useful within-family comparison across substantially different model scales.
- **Llama-3-8B-Instruct** is Meta's 8B instruction-tuned model from the Llama 3 family (Grattafiori et al., 2024). It serves as a widely used general-purpose auxiliary and provides a same-scale family comparison with Llama-3.1-8B-Instruct.
- **Llama-3.1-8B-Instruct** is Meta's updated 8B instruction model from the Llama 3.1 family (Grattafiori et al., 2024). Its pairing with Llama-3-8B-Instruct allows us to compare successive model generations at approximately fixed scale.
- **Llama-3.2-3B-Instruct** is Meta's 3B-parameter instruction-tuned model from the Llama 3.2 family (Meta, 2024). It is optimized for multilingual dialogue and provides a compact Llama-family auxiliary, complementing the larger 8B Llama 3 and Llama 3.1 models in our pool.
- **Qwen3.5-2B** is a compact 2B-parameter model from the Qwen3.5 family, which adopts a hybrid linear- and full-attention architecture (Qwen Team, 2026a). It introduces both a small-scale auxiliary and an architecture distinct from the Qwen3 models.
- **Qwen3.5-9B** is a 9B-parameter model from the same Qwen3.5 family (Qwen Team, 2026a). Together with Qwen3.5-2B, it provides a second within-family comparison for studying the effect of auxiliary scale.
- **Granite-4.2-8B** is IBM's 8B dense reasoning model from the Granite 4.2 family (Granite Team, IBM, 2026). It adds another model family at the common 7–9B scale with a reasoning-oriented post-training strategy.
- **Ornith-1.5-9B** is a 9B reasoning model from the Ornith-1.5 family (Ornith Team, 2026). Its distinctive self-improvement and reinforcement-learning pipeline introduces a substantially different training strategy into the auxiliary pool.

B.2 Dataset Descriptions

We use 22 task benchmarks across all experiments of this work. Together, these benchmarks and datasets cover healthcare, finance, math, code, legal, and science. The 22 task benchmarks are divided into reference and target tasks: reference tasks support the reference-guided multi-auxiliary voting method described in Section 4.1, while target tasks are used to evaluate the effectiveness of our method for learnware identification. The diversity of the reference and target tasks in domain, knowledge coverage, and required capabilities provides a broad reference task pool while allowing us to evaluate the generalizability of our method across different tasks. The roles, subsets, and splits of all datasets are summarized in Table 5.

Split selection. Following LoRA Without Regret (Schulman & Thinking Machines Lab, 2025), we keep the adaptation hyperparameters fixed across benchmarks, requiring 5,000 examples per task.

Table 5: Dataset subsets, splits and roles.

Dataset	Domain	Role	Subset	Split
HEAD-QA V2	Healthcare	Reference	`en`	`train`
MedMCQA	Healthcare	Reference	–	`train`
Bhasha-Fin	Finance	Reference	`English`	`test`
Twitter-FinSent	Finance	Reference	–	`train`
ElemMathQA	Math	Reference	`elementary_math_qa`	`validation`
AQuA-RAT	Math	Reference	–	`train`
CyberMetric	Code	Reference	–	`train`
Indian-Legal	Legal	Reference	`clat_ug`, `clat_pg`, `djs_dhjs`	`test`
CaseHOLD	Legal	Reference	–	`train`
SciQ	Science	Reference	–	`train`
QASC	Science	Reference	–	`train`
MedQA-USMLE	Healthcare	Target	–	`train`
PubMedQA	Healthcare	Target	`pqa_artificial`	`train`
FLARE-Causal20	Finance	Target	–	`test`
FLARE-Headlines	Finance	Target	–	`test`
MathQA	Math	Target	`default`	`train`
Math-MCQA	Math	Target	–	`train`
CodeMMLU-Exec	Code	Target	`execution_prediction`	`test`
CodeMMLU-Syntax	Code	Target	`programming_syntax`	`test`
Bhasha-Legal	Legal	Target	`English`	`test`
ContractNLI-A	Legal	Target	–	`train`
SciKnowEval	Science	Target	`l1`, `l2`	`test`

Note: "–" denotes no specific subset. Benchmark names are abbreviated for compactness.

We therefore use the training split when the official test or validation split is unavailable or too small. Although prior exposure to such splits may affect the induced candidate rankings, all model-selection methods are evaluated against the same ranking obtained from the same candidate pool and evaluation set. The comparison therefore remains consistent with our objective of learnware identification through candidate ranking recovery.

Probe dataset construction. Because high-quality datasets with suitable formats are limited across the six domains, we construct the probe dataset from 15 source datasets that are disjoint from the reference and target tasks data used in our experiments. We randomly sample 5,000 examples from their pooled eligible examples to form a single shared probe dataset, which is then used for all candidate models. We treat this construction as a default setting rather than a design choice to be optimized, since selecting or tuning the probe sources is outside the scope of this work.

Specific information about these datasets is listed below.

Healthcare

- **HEAD-QA**: A multilingual multiple-choice benchmark derived from professional healthcare examinations in Spain (Correa-Guillén et al., 2025). It covers several healthcare disciplines and evaluates specialized medical knowledge and examination-style reasoning over clinically plausible alternatives.
- **PubMedQA**: A biomedical question-answering dataset constructed from PubMed abstracts (Jin et al., 2019). Given a research question and its associated abstract, models determine whether the available evidence supports the answer, evaluating scientific reading comprehension and evidence-based biomedical reasoning.
- **MedMCQA**: A large-scale multiple-choice benchmark collected from Indian medical entrance examinations, spanning 21 medical subjects and thousands of topics (Pal et al., 2022). It evaluates both factual medical knowledge and the reasoning required to solve professional examination questions.
- **MedQA-USMLE-4-options**: An English medical benchmark based on third-party preparation materials for United States medical licensing examinations (Jin et al., 2020). Its four-

choice questions frequently use clinical vignettes and evaluate the application of biomedical knowledge to diagnosis, mechanism, and treatment decisions.

Finance

- **FLARE-Causal20-SC**: A single-choice financial causality task derived from FinCausal 2020 (Mariko et al., 2020). It asks models to distinguish causal statements from non-causal financial text, evaluating understanding of cause–effect relations in financial news and regulatory contexts.
- **BhashaBench-Finance**: A benchmark built from Indian government and institutional examinations covering banking, accounting, taxation, insurance, financial markets, and related topics (Devane et al., 2025). It provides broad evaluation of financial knowledge with substantial coverage of India-specific institutions, regulations, and economic practices.
- **Twitter Financial News Sentiment**: An English corpus of finance-related tweets labeled as *bearish*, *bullish*, or *neutral* (ZeroShot, 2022). It evaluates market-oriented sentiment understanding in short and often noisy financial text containing headlines, ticker symbols, and informal commentary.
- **FLARE Headlines**: A financial headline understanding task formulated as binary-choice questions (Xie et al., 2023; 2024). Models identify attributes of asset-price statements, including whether a price is mentioned, its movement direction, temporal information, and related market signals.

Math

- **MathQA**: A large collection of five-choice mathematical word problems annotated with natural-language rationales, formulas, and operation programs (Amini et al., 2019). It covers areas such as algebra, geometry, probability, and arithmetic, evaluating multi-step quantitative reasoning from natural-language problems.
- **AQuA-RAT**: A dataset of approximately 98,000 five-choice algebraic word problems accompanied by human-readable rationales (Ling et al., 2017). It evaluates multi-step mathematical reasoning in which intermediate calculations are expressed jointly through natural language and mathematical expressions.
- **Elementary Math QA**: A multiple-choice mathematical word-problem task from BIG-bench (BIG-bench authors, 2023). It focuses on elementary arithmetic and logical reasoning expressed through natural-language problem statements, testing whether models can translate simple quantitative descriptions into correct calculations.
- **MATH-MCQA**: A four-choice adaptation of the MATH benchmark containing 12,498 competition-level mathematics problems across algebra, geometry, number theory, counting and probability, precalculus, and related areas (Biderman, 2025). The distractors are designed to represent plausible solution errors, making the task useful for evaluating both mathematical problem solving and discrimination between closely related answers.

Code

- **CodeMMLU Execution Prediction**: A program-understanding task from CodeMMLU that asks models to determine the output or runtime behavior of code examples (Nguyen et al., 2025). It evaluates mental execution, control-flow tracking, state updates, and understanding of programming-language semantics.
- **CodeMMLU Programming Syntax**: A CodeMMLU task evaluating programming syntax and language-specific conventions across languages such as C, C++, Java, JavaScript, Python, R, and TypeScript (Nguyen et al., 2025). Models must distinguish valid constructs from syntactically incorrect alternatives, providing a measure of broad programming-language knowledge.
- **CyberMetric**: A cybersecurity benchmark containing 10,180 four-choice questions designed to assess professional-level security knowledge (Tihanyi et al., 2024). It covers areas including cryptography, security controls, threat intelligence, vulnerability assessment, security architecture, and PCI DSS, evaluating domain knowledge across a broad range of cybersecurity concepts.

⚖ Legal

- **BhashaBench-Legal**: A benchmark of English and Hindi questions collected from Indian judicial-service, bar, and legal-education examinations (Devane et al., 2025). It covers constitutional law, civil and criminal procedure, family law, property, taxation, and other areas, providing jurisdiction-specific evaluation beyond predominantly U.S.-centered legal benchmarks.
- **CaseHOLD**: A multiple-choice benchmark derived from United States judicial opinions (Zheng et al., 2021). Given a citing passage and five candidate holdings, models identify the correct holding of the referenced precedent, testing legal reading and fine-grained case-law reasoning.
- **ContractNLI-A**: A three-way natural-language inference task for contract review (Koreeda & Manning, 2021). Given contractual text and a hypothesis, models predict *entailment*, *contradiction*, or *neutral*, evaluating precise inference over specialized legal clauses.
- **Indian Legal Exam**: A benchmark containing 6,218 questions from 38 Indian legal examination papers published between 2008 and 2025 (Juvekar et al., 2025). We combine its CLAT Undergraduate, CLAT Postgraduate, and Delhi Judicial Service/Higher Judicial Service components to evaluate legal knowledge and reasoning across different educational and professional levels.

🧪 Science

- **SciKnowEval**: A multi-level scientific benchmark covering biology, chemistry, physics, and materials science (Feng et al., 2024). We use its multiple-choice components to evaluate both fundamental scientific knowledge and cross-disciplinary scientific reasoning.
- **QASC**: A benchmark containing 9,980 eight-choice questions primarily about grade-school science (Khot et al., 2020). Many questions require combining multiple pieces of scientific knowledge through an intermediate concept, testing knowledge retrieval and compositional reasoning.
- **SciQ**: A dataset containing 13,679 four-choice science examination questions covering physics, chemistry, biology, and related subjects (Welbl et al., 2017). Supporting evidence is provided for most questions, while the multiple-choice task evaluates scientific knowledge and the ability to distinguish the correct answer from plausible distractors.

B.3 Candidate Model Description

We selected 16 instruction-following models that span dense and mixture-of-experts architectures, general and domain-specific training, from different model family, and parameter counts from 7B to approximately 35B. The diversity of the models well reflects the heterogeneity encountered by a learnware dock system in an open-world setting. The descriptions below summarize the publicly documented checkpoints used as candidates.

MedGemma 27B Text IT is Google's 27B-parameter medical language model built on Gemma 3 and trained specifically on medical text (Sellergren et al., 2025). It shows strong medical knowledge and reasoning performance on corresponding healthcare benchmarks, making it a major healthcare specialist in our candidate pool.

Mistral Small 3.2 24B Instruct is a 24B-parameter dense instruction model from Mistral AI (Mistral AI, 2025). It exhibits broad performance across general knowledge, mathematics, scientific reasoning, and coding benchmarks, serving as a strong dense general-purpose candidate without specialization to a particular domain.

NVIDIA Nemotron 3 Nano 30B-A3B is a 30B-total, 3.5B-active hybrid Mamba–Transformer mixture-of-experts model developed by NVIDIA (NVIDIA, 2025). Its evaluations show broad competence in general knowledge, mathematical and scientific reasoning, providing a sparse general-purpose candidate with relatively balanced capabilities.

Nemotron-Cascade-8B is NVIDIA's 8B general-purpose language model developed from the Qwen3-8B family (Wang et al., 2026). Despite its relatively small scale, it reports strong per-

formance across general knowledge, scientific reasoning, and mathematics benchmarks such as MMLU, MMLU-Pro, GPQA, and AIME, providing a compact but capable generalist in the candidate pool.

GLM-4.7-Flash is a 30B-total, approximately 3B-active mixture-of-experts model released by Z.ai (Z.ai, 2026). It performs strongly on mathematical and scientific reasoning benchmarks such as AIME and GPQA, contributing a reasoning-oriented generalist from the GLM family.

Mellum2-12B-A2.5B-Instruct is JetBrains' 12B-total, approximately 2.5B-active mixture-of-experts instruction model (Kojic et al., 2026). Its evaluations cover general knowledge, mathematics, scientific reasoning, and coding, providing a relatively compact sparse generalist with broad knowledge-oriented capabilities.

Qwen3-Coder-30B-A3B-Instruct is a code-specialized mixture-of-experts model with 30.5B total and 3.3B active parameters (Qwen Team, 2025). Its training is centered on programming and software-related knowledge and reasoning, making it the principal coding specialist in our candidate pool.

Qwen3-32B is a dense 32.8B-parameter model from the Qwen3 family (Yang et al., 2025). Qwen3 emphasizes broad improvements in mathematical, coding, and commonsense reasoning while maintaining strong general-purpose knowledge, making it a representative high-capacity generalist.

Qwen3.8-27B is a 27B-parameter dense model from the Qwen3.8 family (Qwen Team, 2026b) with a hybrid architecture combining Gated DeltaNet and attention layers. Its reported results are particularly strong in scientific and multidisciplinary reasoning, while also retaining broad general knowledge, making it another high-capacity reasoning-oriented generalist.

Qwen3.6-14B-A3B-FableVibes is a community-developed model with approximately 14B total and 3B active parameters, obtained by pruning a larger Qwen-derived model and adapting it on reasoning traces (tvall43, 2026). It is intended as a compact general-purpose reasoning model and introduces both a community-developed checkpoint and a distinct sparse training recipe into the candidate pool.

Ornith-1.5-35B-A3B is a 35B-total mixture-of-experts model with approximately 3B active parameters per token (Ornith Team, 2026). Its reported evaluations show strong scientific reasoning and coding performance, including high results on GPQA Diamond, providing a sparse candidate with pronounced reasoning-oriented capabilities.

Llama3.1-Aloe-Beta-8B is an 8B healthcare-specialized model built on Llama 3.1 and trained across a broad collection of medical tasks by Barcelona Supercomputing Center (Garcia-Gasulla et al., 2025). It performs strongly on medical question-answering benchmarks including MedQA, MedMCQA, PubMedQA, and medical MMLU, providing a smaller healthcare specialist complementary to MedGemma.

Saul-7B-Instruct-v1 is a 7B model developed from Mistral-7B through continued pretraining on a large English legal corpus followed by instruction tuning (Colombo et al., 2024). Its specialization in legal text comprehension and legal-domain tasks makes it the primary legal specialist in our candidate pool.

Phi-4 is Microsoft's 14B-parameter dense language model trained with extensive synthetic and curated data (Abdin et al., 2024b). It is particularly strong in mathematical and scientific reasoning while retaining competitive general knowledge and coding performance, providing a compact reasoning-oriented dense model.

OLMo-2-0325-32B-Instruct is Ai2's 32B-parameter dense instruction model from the fully open OLMo 2 family (Walsh et al., 2025; Lambert et al., 2025). It shows broad capabilities across general knowledge, reading comprehension, commonsense reasoning, and mathematics, providing a high-capacity generalist from a model family distinct from Qwen, Mistral, and Llama.

Granite-4.2-30B is IBM's 30B-parameter dense model from the Granite family (Granite Team, IBM, 2026). Its evaluations show strong mathematical and scientific reasoning together with

competitive multidisciplinary knowledge performance on MMLU-Pro, adding another high-capacity dense generalist from an independent model family.

B.4 Detailed Model Ranking by PAVE Pipeline

This section details the pipeline behind Figure 1 and Table 1. Auxiliaries, datasets, and candidates are described in Appendices B.1–B.3, output filtering in Appendix C.2, and all numerical settings in Table 6; here we only describe how these components are connected. The pipeline is run once per auxiliary, and every other component is shared across auxiliaries.

1. **Task data (Fig. 1B).** Each example of a task, drawn from the subset and split in Table 5, is formatted as a prompt x that presents the question with its answer options and asks for the answer only. Its reference y is the ground-truth label. For every task t, 5,000 examples are sampled with the benchmark sampling seed to form the task dataset $D_t = \{(x, y)\}$.
2. **Candidate inference (Fig. 1A).** Each of the 16 candidates generates once on the probe dataset $\mathcal{P}$ which is sampled from datasets spanning six domains and is disjoint from the data of previous 22 tasks, with the inference settings in Table 6. The probe dataset is shared by all candidates and sampled from source datasets disjoint from the 22 tasks (Table 5). Each probe generation is canonicalized to a valid label, or to an empty string if it cannot be resolved (Appendix C.2).
3. **Specification data.** For candidate M_i, the candidate-specific pseudo dataset $D_i = \{(x, \hat{y}_i(x)) : x \in \mathcal{P}\}$ pairs each probe prompt with M_i's canonicalized answer $\hat{y}_i(x)$, which serves as a pseudo-label. The task dataset D_t uses ground-truth labels instead. Both datasets have the same size and the same format: x is the prompt field, and the label is the completion field on which the loss is computed. Model and task PAVEs thus differ only in their training data, not in the training procedure. Because D_i depends only on the candidate, all auxiliaries are trained on identical data.
4. **PAVE extraction (Fig. 1C).** For auxiliary A_a and each of the $16 + 22 = 38$ datasets D, we fine-tune the auxiliary with LoRA on its MLP projection layers, freeze the A matrices, and train only the B matrices. Every run uses the unified configuration and training seed of Table 6. All 38 runs of an auxiliary therefore start from the same LoRA initialization, with identical frozen A matrices and B matrices at zero, so their B updates share one coordinate system. As in Section 3.1, the PAVE $\boldsymbol{p}_{a,D} \in \mathbb{R}^{d_a}$ flattens the trained B matrices across target modules, i.e., their displacement from initialization. The dimension d_a depends on the auxiliary's depth and MLP widths. We write $\boldsymbol{p}_{a,i}$ and $\boldsymbol{p}_{a,t}$ for the PAVEs of D_i and D_t. Each PAVE is computed once and cached, so a model PAVE is reused across all tasks. In total, the experiment performs $13 \times 38 = 494$ fine-tuning runs.
5. **PAVE-induced ranking (Fig. 1D).** For each auxiliary A_a and task t, we compute $\cos(\boldsymbol{p}_{a,t}, \boldsymbol{p}_{a,i})$ for all 16 candidates, which yields a 22×16 similarity table per auxiliary. Candidates are then ranked by decreasing similarity. This step uses only PAVEs and requires no candidate inference on t.
6. **Evaluation (Fig. 1E).** The evaluation score $s_t(i)$ is M_i's accuracy on the sampled examples of t, and ranking the candidates by $s_t(i)$ gives the evaluation-induced ranking. Kendall's τ_b (Section 2.1) between the evaluation-induced and PAVE-induced rankings gives one entry of Table 1, with tied scores handled by its tie correction. Evaluation scores are used only in this step and never enter PAVE construction.

Only the auxiliary varies across the columns of Table 1. Section 4 reuses these PAVEs and evaluation scores.

B.5 Comparison Methods Details

Our baselines are designed to distinguish the effects of single-auxiliary selection, multi-auxiliary aggregation, prompt-embedding-guided auxiliary selection, and direct learning of model rankings from historical evaluations, with an oracle auxiliary included as a post-hoc reference.

- **Single Best Auxiliary (SBA).** SBA selects the single auxiliary with the highest average ranking quality on the reference tasks and uses this fixed auxiliary for all target tasks. It represents the strongest fixed single-auxiliary choice available when only historical reference-task evaluations are accessible, and serves to test whether dynamically exploiting reference-task information is necessary.
- **Uniform Multi-Auxiliary Voting (U-MAV).** U-MAV aggregates the pairwise ranking decisions of all auxiliaries with uniform weights. Unlike RG-MAV, it does not use reference tasks to estimate auxiliary reliability. It therefore isolates the benefit of reliability-aware weighting from that of multi-auxiliary aggregation itself.
- **Prompt-Embedding Auxiliary Selection (PE-AS).** Inspired by embedding-based task selection methods (Poth et al., 2021; Lin et al., 2024), we represent each target and reference task using prompt embeddings and retrieve the most similar reference task by cosine similarity. We then select the auxiliary that performs best on this retrieved reference task and use it to rank candidate models on the target task. This baseline adapts prompt-embedding-based nearest-task retrieval to auxiliary selection, testing whether the auxiliary favored by the semantically closest reference task is sufficient for reliable ranking on the target task.
- **MODEL SPIDER.** We adapt MODEL SPIDER (Zhang et al., 2023) to our model-ranking setting. MODEL SPIDER learns representations for candidate models and target tasks and uses a lightweight Transformer-based compatibility function, trained from historical model rankings, to directly predict model preferences on unseen tasks. In our setting, observed rankings on the reference tasks provide the historical supervision. This baseline contrasts RG-MAV's explicit estimation and aggregation of auxiliary reliability with an alternative that learns a latent task–model compatibility function for ranking.
- **EmbedLLM.** We adapt EmbedLLM (Zhuang et al., 2025), a recent approach that learns compact model representations from model–question correctness observations. Given a target instance, EmbedLLM predicts the probability that each candidate model answers it correctly; we average these probabilities over the target task and rank models by the resulting scores. It provides an instance-level performance-prediction baseline, contrasting RG-MAV's task-level, auxiliary-based ranking mechanism with direct prediction from historical per-instance correctness.
- **Task-wise Oracle Auxiliary (TOA).** For each target task, TOA retrospectively selects the auxiliary that achieves the highest ranking agreement with the ground-truth candidate-model ranking. Since this selection requires target-task evaluations, TOA is not deployable in practice. Instead, it represents the target-wise performance ceiling of any non-ensemble single-auxiliary strategy within the given auxiliary pool, providing a reference for how close practical auxiliary-selection methods can approach the best available auxiliary.

Implementation details.

For **PE-AS**, we use the frozen `sentence-transformers/all-mpnet-base-v2` encoder. Each task representation is obtained by averaging prompt embeddings over its sampled task prompts, followed by ℓ_2 normalization. For prompts exceeding the encoder context length, we split the prompt into non-overlapping chunks and combine their embeddings using token-count-weighted averaging. We compute cosine similarity between each target task and all reference tasks and retrieve the nearest reference task ($k = 1$). The auxiliary with the highest Kendall's τ_b on this reference task is then selected for the target task.

For **MODEL SPIDER**, we adapt the original framework to our textual LLM setting while preserving its task–model representation learning, Transformer-based compatibility scoring, and ranking objective. We use a frozen `sentence-transformers/sentence-T5` encoder for task representations, learnable 256-dimensional model representations, and a one-layer eight-head Transformer to predict task–model compatibility scores. For each reference benchmark, we sample 16 disjoint 10-example pseudo-tasks and derive model rankings from their observed accuracies as ranking supervision. The model is trained with the ranking loss used by MODEL SPIDER. We adopt its general-token setting ($k = 0$) to maintain the same target-inference-free protocol as the other methods, without the optional candidate-specific re-ranking stage that requires running selected models on the target task. For a target task, we partition its sampled task prompts into disjoint 10-example pseudo-tasks and average the predicted model scores to obtain the final ranking. Since our reference-

task evaluations directly provide historical model rankings, we use these observed rankings as supervision in place of RankAgg, whose role in the original setting is to construct training rankings from transferability signals.

For **EmbedLLM**, we construct the training supervision from the per-instance correctness matrix of all candidate models on the reference tasks. Questions are encoded with frozen `sentence-transformers/all-mpnet-base-v2` representations, while each candidate model is assigned a learnable 232-dimensional embedding. Following EmbedLLM, the projected question representation and model representation are combined element-wise and passed to a binary classifier for correctness prediction. We train the model for 50 epochs with Adam, a learning rate of 10^{-4}, weight decay of 10^{-5}, and batch size 2048. At inference time, the predicted correctness probabilities are averaged across all target samples for each candidate model, and the resulting scores are used to produce the final ranking.

B.6 Experiment Configuration

We use a unified training configuration across all auxiliaries, tasks, and model specifications to ensure that the resulting PAVE geometries are comparable. The complete data-processing, optimization, LoRA, and evaluation settings are summarized in Table 6.

B.7 Computation Resources

Experiments were conducted using two NVIDIA A100 80GB PCIe GPUs, two Intel Xeon Platinum 8358 CPUs with 32 cores each (base clock 2.6 GHz, turbo boost 3.4 GHz), and 1 TiB of RAM. The experiments reported in this paper required approximately 1,550 A100 GPU-hours, including approximately 494 GPU-hours for PAVE vector generation and 1056 GPU-hours for model inference. Importantly, more than half of the inference compute was used to obtain ground-truth evaluation scores on the target benchmarks for experimental evaluation, and is not required by our identification method. Model inference required by the method is performed on the probe dataset to construct model PAVE specifications and on reference tasks for reference-guided multi-auxiliary voting.

C Output Format Effects and Filtering

In this section, we examine how variations in candidate-model output format can affect PAVE similarity. We then introduce a lightweight filtering procedure to canonicalize probe outputs and evaluate its effect at both the pairwise-ordering and overall-ranking levels.

C.1 Impact of Output Format on PAVE Similarity

The model dataset used to construct each model specification consists of the candidate model's own generations on the probe dataset. As a result, without output normalization, the learned specification may reflect not only *which answer* the candidate model selects, but also *how that answer is expressed*. For example, one model may return only the option label, while another may append the option text, an explanation, or intermediate reasoning, even when both are prompted to produce only the answer.

Such differences change the completion sequences used to fine-tune the auxiliary. Under the language-modeling objective, different output tokens induce different training signals and therefore different parameter updates. Additional explanatory content can consequently introduce supervision unrelated to the candidate model's final decision, causing the resulting specification to partially capture response style or verbosity rather than only the model's decision behavior on the probe dataset.

Figure 3 illustrates this difference on a probe instance. `Qwen3-Coder-30B` returns the selected option together with a detailed explanation, whereas `Mellum2-12B` produces only an option label. Such variation can introduce an unintended factor into specification construction and thereby affect PAVE similarity.

To examine this effect empirically, we consider pairs containing one model from the top three and one from the bottom three of the evaluation-induced ranking, thereby focusing on models with clearly different target-task performance. We identify 19 cases across 11 benchmarks

Table 6: Hyperparameters and implementation settings used in our experiments.

Component	Hyperparameter	Value
Data	Samples per task	5,000
Data	Training examples per specification	5,000
Data	Benchmark sampling seed	12,123
Data	Probe-dataset sampling seed	6,791
Data	Training format	Separate `prompt` and `completion` fields; loss is computed on completion tokens only
Inference	Precision	bfloat16
Inference	Batch size	8
Inference	Maximum new tokens	256
Inference	Decoding	Greedy (`do_sample=false`)
Auxiliary	Attention implementation	Eager
Auxiliary	Model/training precision	FP32
Tokenization	Maximum sequence length	2,048 tokens
Tokenization	Padding	Right padding; the EOS token is used as the PAD token when no PAD token is defined
LoRA	Rank	256
LoRA	lora alpha	32
LoRA	Dropout	0.1
LoRA	Bias	None
LoRA	Task type	Causal language modeling
LoRA	Target modules	MLP projection layers
LoRA	Trainable parameters	LoRA-B matrices only; LoRA-A matrices are frozen after initialization
Optimization	L1 regularization (`l1_weight_decay`)	0.01
Optimization	Optimizer	PyTorch AdamW (`adamw_torch`)
Optimization	Learning rate	1.1×10^{-4}
Optimization	LR schedule	Cosine decay
Optimization	Warmup(`warmup_steps`)	16 steps
Optimization	L2 weight decay	0.01
Optimization	Maximum gradient norm	1.0
Optimization	Maximum training steps	1,250; this step limit overrides epoch-based stopping
Batching	Accelerate processes	2
Batching	Per-device batch size	4
Batching	Gradient accumulation	1
Batching	Effective global batch size	8 examples ($2 \times 4 \times 1$)
Training	Training seed	3,407 for PyTorch, CUDA, and `SFTConfig`; deterministic cuDNN enabled

in which the higher-evaluation model nevertheless receives lower PAVE similarity under the `Phi-3.5-mini-instruct` auxiliary, as an example for illustration. After canonicalizing the probe outputs using the procedure in Section C.2, all 19 recover the evaluation-consistent ordering. Table 7 shows one representative pair for each affected benchmark.

C.2 Output Filtering

To reduce the influence of response formatting, we canonicalize each candidate model's probe generation to its predicted label before constructing the model specification. We first apply deterministic matching against the valid labels of the corresponding probe task. Non-empty outputs that cannot be resolved deterministically are passed to `DeepSeek-v4-Flash-0731` as a secondary label

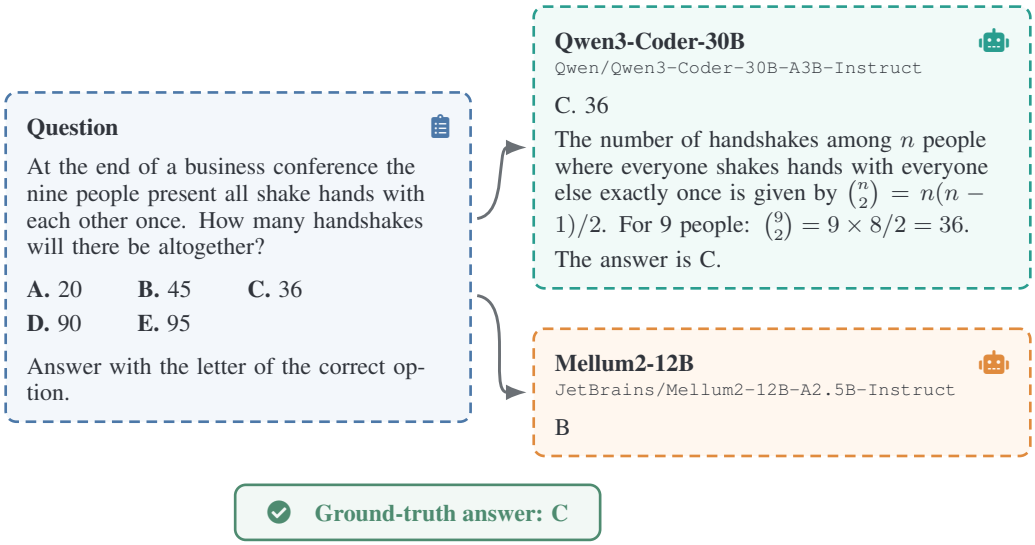


Figure 3: Example of two candidate-model responses to the same question. The responses may differ in both the predicted answer and output format, while the ground-truth answer is shown separately.

Table 7: Representative unexpected PAVE similarity orderings before and after output filtering under the Phi-3.5-mini-instruct auxiliary. Evaluation scores are shown in parentheses.

Benchmark	Higher-evaluation model	Lower-evaluation model	Unfiltered	Filtered
PubMedQA	Aloe-8B (0.974)	Nemo-Cascade-8B (0.920)	$0.215 < 0.216$	$0.233 > 0.216$
ContractNLI-A	Phi-4 (0.811)	GLM-4.7 (0.607)	$0.191 < 0.204$	$0.256 > 0.204$
Bhasha-Legal	Phi-4 (0.717)	Q3.6-Fable (0.518)	$0.300 < 0.338$	$0.371 > 0.338$
Twitter-FinSent	Granite-30B (0.787)	GLM-4.7 (0.471)	$0.227 < 0.229$	$0.241 > 0.229$
SciKnowEval	Phi-4 (0.858)	Saul-7B (0.738)	$0.257 < 0.262$	$0.310 > 0.261$
QASC	Q3-Coder-30B (0.658)	Mellum2-12B (0.539)	$0.227 < 0.241$	$0.246 > 0.242$
Indian-Legal	Q3-32B (0.718)	Q3.6-Fable (0.530)	$0.338 < 0.339$	$0.362 > 0.339$
MedQA-USMLE	Q3-32B (0.753)	Nemo-Cascade-8B (0.562)	$0.200 < 0.201$	$0.213 > 0.201$
Math-MCQA	Q3-32B (0.649)	Saul-7B (0.365)	$0.175 < 0.176$	$0.190 > 0.181$
ElemMathQA	Q3-32B (0.477)	Nemo-3-Nano-30B (0.266)	$0.162 < 0.171$	$0.175 > 0.172$
FLARE-Headlines	Granite-30B (0.763)	Nemo-3-Nano-30B (0.492)	$0.172 < 0.183$	$0.183 > 0.182$

Note. Benchmark and model names are abbreviated for compactness.

extractor. The extractor receives only the raw generation and the valid candidate set, without access to the reference answer, and is instructed to recover the model's final decision rather than solve the original question. We use temperature 0, a maximum of 32 output tokens, and disabled reasoning. Extracted labels are validated against the candidate set; successfully resolved outputs are replaced by their canonical labels, while unresolved outputs are represented by an empty string.

LLM-based fallback prompt. For reproducibility, we provide the exact system prompt used for the LLM-based fallback extractor:

```
You extract the final option label chosen in a raw model
output.  Use only the raw output; do not solve the original
question yourself.

Return exactly one valid candidate label.  If no candidate
can be determined, return NONE.

Do not explain or output any other text.
```

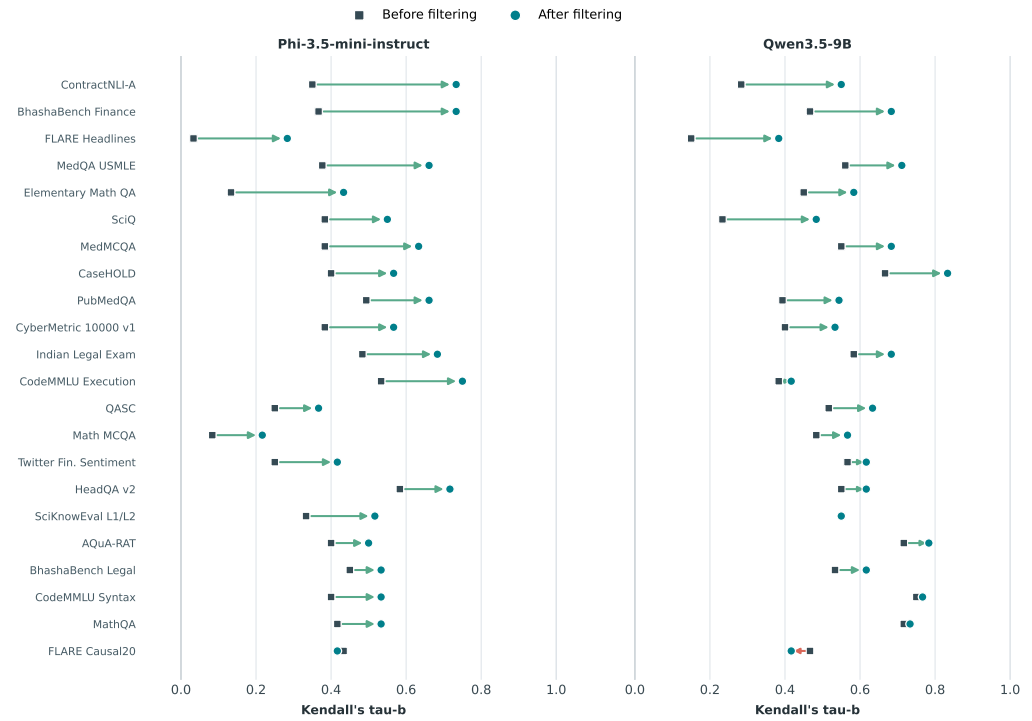


Figure 4: Kendall's τ_b between the PAVE-induced candidate-model ranking and the ranking determined by target-task evaluation scores, before and after output filtering, under two auxiliaries.

C.3 Effect of Output Filtering

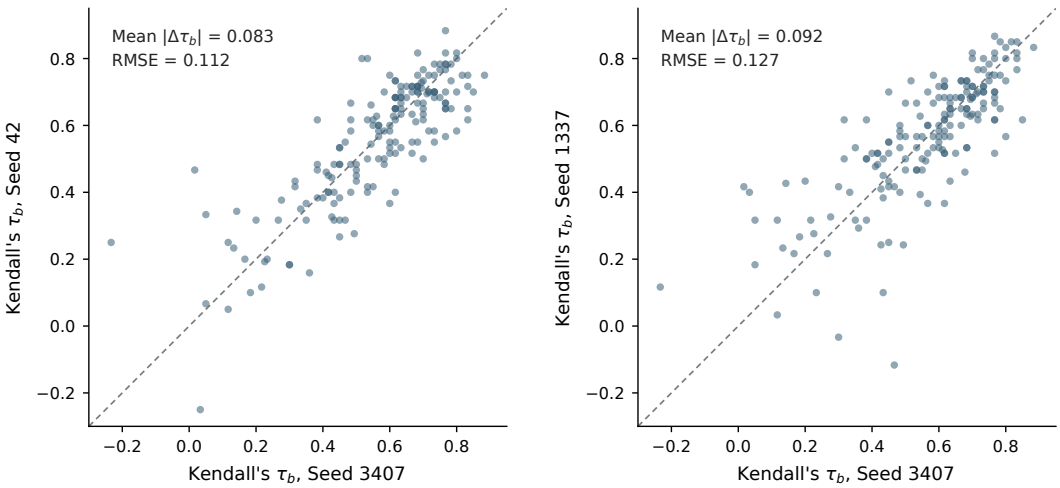
The pairwise results in Table 7 show that output canonicalization can recover individual similarity orderings. We further examine whether this effect extends to the overall candidate-model ranking. For each task, we compute Kendall's τ_b between the PAVE-induced and evaluation-induced rankings, once using model specifications constructed from the original probe outputs and once using specifications constructed from the filtered outputs. We perform this comparison under `Phi-3.5-mini-instruct` and `Qwen3.5-9B`, as shown in Figure 4.

Filtering improves or preserves ranking agreement on almost all benchmarks under both auxiliaries. The largest gains under `Phi-3.5-mini-instruct` include ContractNLI-A, BhashaBench Finance, FLARE Headlines, and Elementary Math QA, while `Qwen3.5-9B` shows pronounced improvements on ContractNLI-A, SciQ, FLARE Headlines, and CaseHOLD. In contrast, benchmarks with already strong unfiltered agreement, such as AQuA-RAT and CodeMMLU Syntax, change only slightly. FLARE-Causal20 is the main exception, with a small decrease after filtering.

The consistent trend across two different auxiliaries suggests that the benefit is not specific to a particular PAVE representation. At the same time, the varying magnitude of improvement across benchmarks indicates that output-format effects are not uniform: they can substantially distort the induced ranking in some tasks while having little influence in others. Together with the pairwise recoveries in Table 7, these results support canonicalizing probe outputs as a practical and necessary way to reduce unintended format-related variation in PAVE similarity.

D Experimental Results

In this appendix, we provide additional experiments supporting the robustness and design choices of the main results. We examine seed sensitivity and the stability of auxiliary complementarity (Appendix D.1), evaluate complementarity and RG-MAV under Top-1 regret (Appendix D.2), ablate the main RG-MAV components and individual auxiliaries (Appendices D.3, D.4), test robustness



(a) Seed 42 **(b) Seed 1337**

Figure 5: Seed sensitivity of individual auxiliary–task ranking quality. Each point corresponds to one auxiliary–task combination and compares its Kendall's τ_b under the default seed 3407 with an alternative training seed. The dashed diagonal denotes identical ranking quality under the two seeds.

Table 8: Seed sensitivity and cross-auxiliary complementarity across 8 auxiliaries and 22 benchmarks.

Seed	Overall Mean	Best Fixed Auxiliary	Task-wise Max	Gap	Winning Auxiliaries	Crossing Pairs
3407	0.562	0.615	0.725	0.110	8/8	28/28
42	0.555	0.633	0.740	0.107	7/8	28/28
1337	0.566	0.630	0.730	0.100	8/8	28/28

Note. Overall Mean averages Kendall's τ_b over all auxiliary–task combinations. Best Fixed Auxiliary is the highest average τ_b achieved by a single auxiliary across all tasks. Task-wise Max averages the best auxiliary performance separately for each task, and Gap is its difference from Best Fixed Auxiliary. Winning Auxiliaries counts auxiliaries that attain the task-wise best on at least one benchmark. A Crossing Pair indicates that each auxiliary in the pair outperforms the other on at least one benchmark.

to reference-pool composition (Appendix D.5), and provide detailed per-task analysis of reference locality (Appendix D.6).

D.1 Seed Sensitivity and Robustness of Auxiliary Complementarity

The PAVE construction involves stochastic fine-tuning and may therefore vary with the training seed. We examine both whether individual auxiliary performance is sensitive to this randomness and whether the cross-auxiliary complementarity observed in the main experiments persists under different training outcomes. We repeat PAVE construction using two alternative seeds, 42 and 1337, for a representative subset of eight auxiliaries: Qwen3-1.7B, Qwen3.5-9B, OLMo-3-7B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, MediPhi-Instruct, Granite-4.2-8B, and Ornith-1.5-9B. This subset spans multiple model families and parameter scales. We use eight rather than all 13 auxiliaries because this analysis targets the structural behavior of the auxiliary pool, while regenerating PAVEs under each additional seed requires another full round of fine-tuning for every auxiliary–model and auxiliary–task combination. All other experimental settings are kept unchanged.

Results. Figure 5 shows that individual auxiliary–task ranking quality can vary noticeably with the training seed. Relative to seed 3407, the mean absolute change in Kendall's τ_b is 0.083 for seed 42 and 0.092 for seed 1337. Among the 176 auxiliary–task combinations, seed 42 yields 83 improvements, 82 declines, and 11 unchanged results, while seed 1337 yields 84 improvements, 84 declines, and 8 unchanged results. Despite these local variations, the overall means remain similar

Table 9: RG-MAV performance under different PAVE training seeds using the same 8 auxiliary subset.

Target task	Seed 3407	Seed 42	Seed 1337
MathQA	0.917	0.929	0.929
Math-MCQA	0.604	0.544	0.633
MedQA-USMLE	0.807	0.828	0.768
PubMedQA	0.414	0.287	0.464
Bhasha-Legal	0.917	0.917	0.915
FLARE-Headlines	0.616	0.487	0.678
FLARE-Causal20	0.322	0.479	0.350
CodeMMLU-Syntax	0.751	0.773	0.783
CodeMMLU-Exec	0.487	0.538	0.538
ContractNLI-A	0.678	0.672	0.678
SciKnowEval	0.819	0.785	0.817
Avg.	0.666	0.658	0.687

Note. For a controlled comparison, RG-MAV uses the same 8 auxiliary subset under all three seeds. The seed-3407 results therefore differ from the main-text results.

at 0.562, 0.555, and 0.566, indicating that aggregate similarity can mask substantial changes for individual auxiliaries and tasks.

Importantly, the complementary structure persists across all three seeds. The gap between the task-wise maximum and the best fixed auxiliary remains 0.110, 0.107, and 0.100 for seeds 3407, 42, and 1337, respectively. The task-wise winners are distributed across 8/8, 7/8, and 8/8 auxiliaries, and all 28 auxiliary pairs exhibit cross-task performance reversals under every seed. Thus, although the effectiveness of a particular auxiliary can depend on stochastic fine-tuning, task-dependent specialization across auxiliaries remains stable.

We further evaluate whether this persistent complementarity can be exploited by RG-MAV. For a controlled comparison, Table 9 uses the same eight auxiliaries under all three seeds, including seed 3407; hence these values differ from the 13-auxiliary results reported in the main text. RG-MAV achieves average Kendall's τ_b values of 0.666, 0.658, and 0.687 under seeds 3407, 42, and 1337, respectively. Despite noticeable variation in individual auxiliary–task rankings, the aggregated ranking quality remains similar across seeds. Together, these results support the use of complementary auxiliaries to reduce dependence on any single seed-sensitive PAVE geometry and obtain more robust learnware identification.

D.2 Top-1 Regret and Auxiliary Complementarity

Kendall's τ_b evaluates recovery of the full candidate-model ranking, whereas some deployment settings may care primarily about whether the top-ranked model is competitive with the true best candidate. We therefore additionally evaluate Top-1 regret. For target task t, let $m_t^\star$ denote the candidate with the highest evaluation score and let $\hat{m}_t$ denote the model ranked first by a selection method. We define

$$R_t = s_t(m_t^\star) - s_t(\hat{m}_t),$$

where lower values are better. For a tie among predicted top-ranked models, we report the average regret over all tied models, avoiding optimistic tie-breaking with target-task labels.

Auxiliary complementarity under Top-1 regret. Table 10 shows that the complementarity observed under Kendall's τ_b is not specific to full-ranking recovery. All 13 auxiliaries attain the minimum Top-1 regret on at least one benchmark when ties are counted. The best fixed auxiliary, Qwen3.5-9B, obtains an average regret of 0.019, whereas selecting the best auxiliary separately for each benchmark reduces the average to 0.002. Thus, no single auxiliary consistently provides the best Top-1 model selection across the task suite, and task-dependent auxiliary specialization remains evident even when only the selected model matters.

Exploiting complementarity with RG-MAV. Figure 6 further evaluates whether this complementarity translates into improved Top-1 model selection. RG-MAV achieves an average regret of

Table 10: Top-1 regret of individual auxiliaries across benchmarks. Lower is better.

Benchmark	Q3 1.7B	Q3 8B	Q3.5 2B	Q3.5 9B	OLMo3 Inst.	OLMo3 Math	L3.2 3B	L3.1 8B	L3 8B	Medi Phi	Phi 3.5	Granite 4.2	Ornith 1.5	Min.
AQuA-RAT	**0.000**	0.006	0.006	**0.000**	0.006	0.006	0.006	0.142	0.006	0.006	0.116	0.006	**0.000**	**0.000**
Bhasha-Fin	0.050	0.022	0.037	0.037	0.097	0.050	0.037	0.097	0.037	0.022	0.022	**0.000**	0.037	**0.000**
Bhasha-Legal	**0.000**	0.020	**0.000**	0.020	0.001	0.070	0.001	**0.000**	0.120	**0.000**	**0.000**	**0.000**	0.001	**0.000**
CaseHOLD	**0.000**	0.012	**0.000**	0.012	**0.000**	**0.000**	0.057	0.012	**0.000**	**0.000**	**0.000**	0.012	**0.000**	**0.000**
CodeMMLU-Exec	0.033	0.049	0.006	**0.000**	0.006	0.013	0.006	0.033	0.006	0.006	0.006	0.024	**0.000**	**0.000**
CodeMMLU-Syntax	0.062	**0.007**	0.062	**0.007**	0.030	0.030	**0.007**	0.062	0.075	0.030	0.030	0.030	**0.007**	**0.007**
ContractNLI-A	**0.000**	0.012	**0.000**	0.074	0.012	0.005	0.012	0.012	0.035	**0.000**	0.012	0.036	0.025	**0.000**
FLARE-Causal20	0.082	0.099	0.105	0.076	0.099	0.072	0.184	0.233	0.233	**0.016**	0.076	0.233	0.233	**0.016**
HEAD-QA v2	0.068	0.027	0.049	**0.000**	0.050	0.050	0.060	0.049	**0.000**	0.050	0.050	**0.000**	**0.000**	**0.000**
MathQA	**0.000**	0.014	0.014	0.014	0.014	0.014	**0.000**	0.014	**0.000**	0.014	0.110	0.014	0.014	**0.000**
MedMCQA	0.013	0.013	0.034	**0.000**	0.008	0.014	**0.000**	**0.000**	0.013	0.014	0.014	**0.000**	**0.000**	**0.000**
PubMedQA	0.003	0.003	0.026	**0.000**	0.003	0.003	0.035	0.010	0.035	0.003	0.003	0.024	0.003	**0.000**
QASC	0.059	0.028	0.019	0.019	0.019	0.018	0.018	0.028	0.018	**0.017**	**0.017**	**0.017**	0.019	**0.017**
SciKnowEval	0.027	0.027	0.061	**0.000**	0.027	0.027	0.027	0.027	0.061	0.027	0.015	**0.000**	0.022	**0.000**
Twitter-FinSent	0.048	0.046	0.048	0.046	0.064	0.048	**0.000**	0.039	0.221	0.000	0.000	0.000	**0.000**	**0.000**
CyberMetric	0.050	0.019	0.019	0.009	0.044	0.014	**0.000**	0.019	0.019	0.019	0.019	0.019	**0.000**	**0.000**
ElemMathQA	0.008	0.008	**0.000**	**0.000**	0.008	0.008	0.021	0.008	0.008	0.008	0.021	**0.000**	0.021	**0.000**
FLARE-Headlines	0.026	0.026	0.010	0.026	0.026	0.060	0.145	0.090	0.145	**0.008**	0.082	0.090	0.092	**0.008**
Indian-Legal	**0.000**	0.028	**0.000**	0.048	0.027	0.027	0.012	0.003	0.012	**0.000**	**0.000**	0.012	0.012	**0.000**
Math-MCQA	**0.000**	0.074	0.074	**0.000**	0.074	0.074	0.074	0.074	0.119	0.074	0.032	0.074	0.032	**0.000**
MedQA-USMLE	0.096	0.135	0.096	**0.000**	0.143	0.135	0.135	0.135	0.137	0.096	0.135	0.135	0.096	**0.000**
SciQ	0.007	0.017	0.023	0.030	0.028	0.016	**0.000**	0.017	**0.000**	0.017	0.017	0.007	**0.000**	**0.000**
Average	0.029	0.032	0.031	**0.019**	0.036	0.034	0.038	0.050	0.059	0.019	0.035	0.033	0.028	0.002

Note: Auxiliary and benchmark names are abbreviated for compactness. Q and L denote Qwen and Llama. The best value in each benchmark row is bolded, and the worst value in each auxiliary column is underlined. The final column reports the minimum regret across auxiliaries.

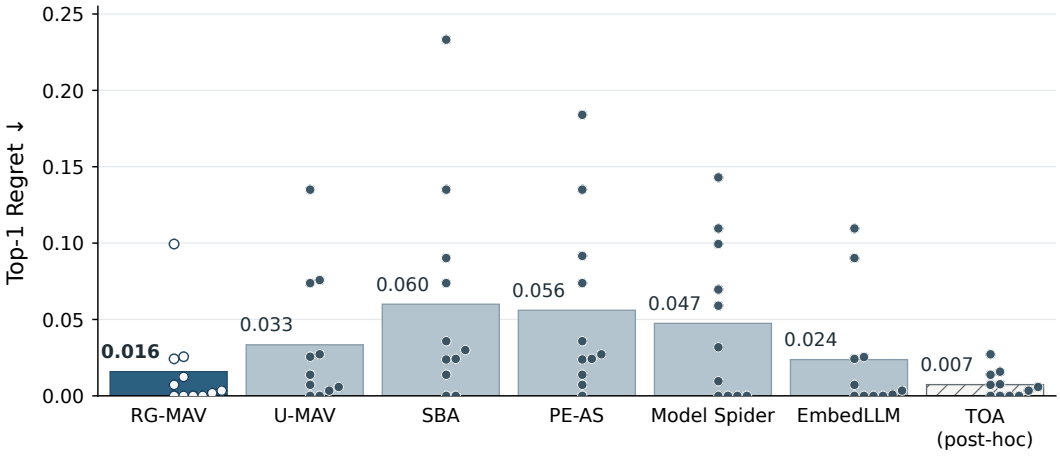


Figure 6: Top-1 regret of model-selection methods across the 11 target tasks. Bars report the mean regret and points show individual target-task results. Lower is better.

0.0158, compared with 0.0334 for U-MAV, 0.0600 for SBA, 0.0560 for PE-AS, 0.0475 for Model Spider, and 0.0237 for EmbedLLM. The post-hoc TOA attains 0.0073. RG-MAV therefore reduces average regret by approximately 53% relative to uniform multi-auxiliary voting and also improves over the non-oracle comparison methods. These results show that the benefit of complementary auxiliaries extends beyond recovery of the full ranking: RG-MAV also identifies high-performing Top-1 candidates when the downstream objective is concerned primarily with the selected model.

D.3 Component Ablation of RG-MAV

We further ablate the main components of RG-MAV to examine where its improvement comes from. Starting from the full method with $k = 1$, we consider the following three variants.

Table 11: Component ablation of RG-MAV across the 11 target tasks.

Target task	Without Pair-Specific Weighting	Without Orientation Factor	Without Reference Locality	Complete RG-MAV
MathQA	0.750	0.937	0.756	0.950
Math-MCQA	0.271	0.286	0.565	0.588
MedQA-USMLE	0.644	0.723	0.728	0.814
PubMedQA	0.424	0.430	0.397	0.390
Bhasha-Legal	0.756	0.912	0.783	0.933
FLARE-Headlines	0.555	0.553	0.447	0.577
FLARE-Causal20	0.352	0.403	0.279	0.397
CodeMMLU-Syntax	0.756	0.785	0.807	0.728
CodeMMLU-Exec	0.633	0.661	0.538	0.510
ContractNLI-A	0.833	0.840	0.667	0.769
SciKnowEval	0.599	0.711	0.700	0.795
Avg.	0.598	0.658	0.606	**0.677**

Without reference locality. We set $k = 11$ and use all reference tasks to estimate auxiliary reliability, rather than restricting the reference neighborhood to the nearest task. This ablation removes the locality used by RG-MAV and tests whether nearby reference tasks provide more informative evidence of auxiliary reliability than a broader reference neighborhood.

Without the orientation factor. We remove the orientation factor$\alpha_{a,r}$ when constructing the reference neighborhood and retrieve reference tasks only according to their PAVE similarity $\rho_{a,r} = \langle \boldsymbol{q}_{a,t}, \boldsymbol{q}_{a,r} \rangle$. The nearest reference task is therefore shared across candidate pairs under the same auxiliary, without considering whether the target and reference have opposite orientations along $\boldsymbol{d}_a$. This ablation tests the contribution of the orientation factor in our method RG-MAV.

Without pair-specific weighting. We keep the pair-specific reference neighborhood and orientation factor unchanged, and first compute w_a exactly as in RG-MAV for every candidate pair. Writing $w_a^{i,j}$ for the weight corresponding to pair (M_i, M_j), we then replace these pair-specific weights with a single auxiliary-level weight

$$\bar{w}_a = \frac{1}{\binom{n}{2}} \sum_{i<j} w_a^{i,j},$$

which is shared by all pairwise votes from auxiliary A_a on target task t. This ablation isolates the effect of pair-specific auxiliary reliability estimation while retaining the same reference selection and orientation correction as the full method.

Results. Table 11 shows that all three components contribute to the overall performance of RG-MAV. Removing pair-specific weighting reduces the average Kendall's τ_b from 0.677 to 0.598, the largest degradation among the three ablations, indicating that estimating auxiliary reliability separately for different candidate pairs is important for exploiting auxiliary complementarity. Removing reference locality decreases the average to 0.606, supporting the use of nearby reference tasks rather than a broad reference neighborhood. Removing the orientation factor yields a smaller but still consistent reduction in the average, from 0.677 to 0.658, indicating that pair-specific orientation provides additional information when selecting reference tasks.

The complete method does not dominate every ablation on every target task. For example, removing the orientation factor improves the results on CodeMMLU-Syntax, CodeMMLU-Exec, and ContractNLI-A, while removing reference locality improves PubMedQA, CodeMMLU-Syntax, and CodeMMLU-Exec. Similarly, removing pair-specific weighting yields higher scores on several individual targets despite substantially reducing the overall average. Such task-level exceptions are expected because reference-based reliability estimation is target dependent, and a simplified variant may occasionally produce more favorable auxiliary decisions for a particular task. Nevertheless, none of these gains persists consistently across the target suite, and every ablation decreases the cross-task average. Overall, the results indicate that pair-specific weighting and reference locality provide the largest contributions, while the orientation factor offers a smaller complementary improvement, jointly improving the robustness of RG-MAV across heterogeneous target tasks.



Figure 7: Leave-one-auxiliary-out ablation of RG-MAV. Each row corresponds to the removal of one auxiliary, and each cell reports Kendall's τ_b on the corresponding target task. The final column reports the average over the 11 target tasks.

D.4 Leave-One-Auxiliary-Out Ablation

We further examine whether RG-MAV depends critically on any individual auxiliary. We remove each of the 13 auxiliaries in turn and rerun RG-MAV with the remaining 12 auxiliaries, while keeping all other components unchanged. This yields 13 leave-one-auxiliary-out variants and directly measures the sensitivity of the aggregation to each individual auxiliary.

Results. Figure 7 shows that RG-MAV remains stable when any single auxiliary is removed. Across the 13 leave-one-auxiliary-out variants, the average Kendall's τ_b ranges from 0.667 to 0.685, compared with 0.677 for complete RG-MAV. The largest decrease in average performance is only 0.011, while removing several auxiliaries slightly improves the average result. Although individual target tasks can exhibit larger fluctuations, no auxiliary consistently determines the performance across tasks. These results indicate that RG-MAV does not rely critically on any single auxiliary and that its overall performance is robust to single-auxiliary removal.

D.5 Robustness to Reference-Pool Composition

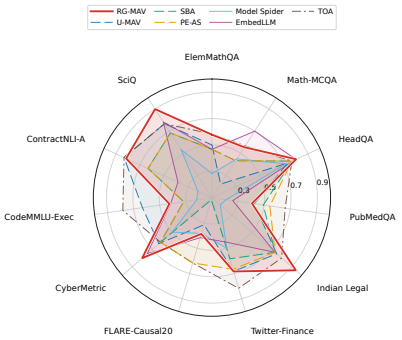
RG-MAV assumes a maintained pool of reference tasks, while target tasks arrive at deployment time and need not be known in advance. We therefore examine whether its performance depends on a particular choice of reference tasks. We construct four alternative suites with different reference pools and evaluate RG-MAV on the corresponding held-out target tasks, while keeping the candidate models, auxiliary pool, hyperparameters, and evaluation protocol unchanged. The reference pools are designed to retain task diversity across multiple domains, reflecting the fact that their composition is a controllable system-level choice in practical deployment.

Results. RG-MAV remains effective under different compositions of the maintained reference pool. Across the four suites, its average Kendall's τ_b is 0.663, 0.728, 0.640, and 0.770, respectively. Compared with U-MAV, the corresponding differences are $+0.073$, $+0.049$, -0.005, and $+0.025$. Thus, reference-guided weighting improves over uniform voting in three of the four suites and is nearly tied in the remaining one.

The variation across suites also highlights the role of reference-task coverage. Because RG-MAV estimates auxiliary reliability from nearby reference tasks, its effectiveness naturally depends on how informative the maintained reference pool is for an incoming target. A reference pool with broader task and domain coverage provides more opportunities to retrieve relevant evidence, whereas a less representative pool can weaken the advantage over uniform voting. Importantly, the reference pool is maintained by the system and is therefore a controllable design choice, unlike the incoming target task.

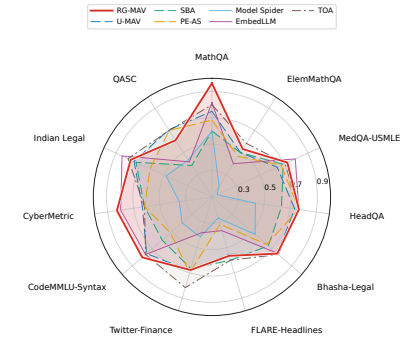
Suite 1

Target task	RG-MAV	U-MAV	SBA	PE-AS	Model Spider	Embed LLM	TOA
ElemMathQA	0.577	0.500	0.467	0.467	0.283	0.467	**0.583**
Math-MCQA	0.555	0.222	0.433	0.433	0.450	**0.700**	0.567
HeadQA	**0.802**	0.734	0.767	0.767	0.717	0.800	0.767
PubMedQA	0.407	0.424	0.494	0.544	0.209	0.259	**0.661**
Indian-Legal	**0.941**	0.750	0.733	0.733	0.183	0.733	0.800
Twitter-FinSent	0.684	0.683	0.583	0.667	0.500	0.450	**0.817**
FLARE-Causal20	0.387	0.315	0.117	**0.617**	0.383	0.417	**0.617**
CyberMetric	**0.800**	0.633	0.617	0.617	0.500	0.750	0.617
CodeMMLU-Exec	0.427	0.633	0.317	0.317	0.217	0.417	**0.783**
ContractNLI-A	0.816	**0.833**	0.633	0.633	0.217	0.383	**0.833**
SciQ	**0.898**	0.767	0.683	0.683	0.533	0.783	0.767
Avg.	0.663	0.590	0.531	0.589	0.381	0.560	**0.710**



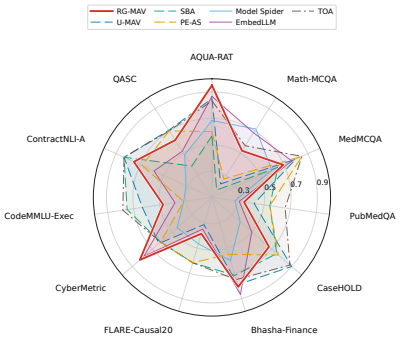
Suite 2

Target task	RG-MAV	U-MAV	SBA	PE-AS	Model Spider	Embed LLM	TOA
MathQA	**0.962**	0.750	0.600	0.683	0.600	0.817	0.800
ElemMathQA	0.532	0.500	0.500	0.467	0.200	0.400	**0.583**
MedQA-USMLE	0.729	0.644	0.695	0.678	0.142	**0.795**	0.711
HeadQA	**0.767**	0.734	0.633	**0.767**	0.433	**0.767**	**0.767**
Bhasha-Legal	0.756	0.756	0.667	0.650	0.533	0.750	**0.767**
FLARE-Headlines	0.565	0.555	**0.600**	0.317	0.267	0.367	**0.600**
Twitter-FinSent	0.678	0.683	0.667	0.700	0.417	0.383	**0.817**
CodeMMLU-Syntax	**0.800**	0.756	0.600	0.517	0.400	0.767	0.767
CyberMetric	**0.828**	0.633	0.617	0.617	0.350	0.800	0.617
Indian-Legal	0.778	0.750	0.733	0.617	0.483	**0.850**	0.800
QASC	0.611	**0.706**	0.383	0.700	0.433	0.417	0.700
Avg.	**0.728**	0.679	0.609	0.610	0.387	0.647	0.721



Suite 3

Target task	RG-MAV	U-MAV	SBA	PE-AS	Model Spider	Embed LLM	TOA
AQuA-RAT	**0.950**	0.850	0.567	0.600	0.683	0.867	0.833
Math-MCQA	0.521	0.222	0.167	0.267	**0.717**	0.683	0.567
MedMCQA	0.695	0.797	0.650	**0.850**	0.600	0.767	**0.850**
PubMedQA	0.346	0.424	0.544	0.544	0.276	0.310	**0.661**
CaseHOLD	0.672	**0.900**	0.750	0.767	0.383	0.500	0.883
Bhasha-Finance	0.807	0.790	0.717	0.550	0.600	**0.867**	0.750
FLARE-Causal20	0.387	0.315	**0.617**	**0.617**	0.467	0.350	**0.617**
CyberMetric	**0.824**	0.633	0.617	0.617	0.450	0.783	0.617
CodeMMLU-Exec	0.475	0.633	0.750	0.317	0.317	0.383	**0.783**
ContractNLI-A	0.751	**0.833**	**0.833**	0.683	0.317	0.583	**0.833**
QASC	0.617	**0.706**	0.383	0.700	0.483	0.517	0.700
Avg.	0.640	0.646	0.599	0.592	0.481	0.601	**0.736**



Suite 4

Target task	RG-MAV	U-MAV	SBA	PE-AS	Model Spider	Embed LLM	TOA
MathQA	0.617	0.750	0.600	0.733	0.067	**0.800**	**0.800**
AQuA-RAT	0.740	**0.850**	0.567	0.783	0.250	0.733	0.833
MedQA-USMLE	0.809	0.644	0.695	0.678	0.510	**0.845**	0.711
MedMCQA	0.711	0.797	0.650	**0.850**	0.567	0.733	**0.850**
Bhasha-Legal	**0.933**	0.756	0.667	0.767	0.700	0.867	0.767
FLARE-Headlines	0.582	0.555	**0.600**	0.350	0.467	**0.600**	**0.600**
Bhasha-Finance	0.852	0.790	0.717	0.683	0.500	**0.883**	0.750
CodeMMLU-Syntax	0.700	0.756	0.600	**0.767**	0.000	0.733	**0.767**
CyberMetric	**0.831**	0.633	0.617	0.483	0.500	0.817	0.617
CaseHOLD	0.785	**0.900**	0.750	0.767	0.333	0.500	0.883
SciQ	**0.912**	0.767	0.617	0.683	0.600	0.867	0.767
Avg.	**0.770**	0.745	0.643	0.686	0.409	0.762	0.759

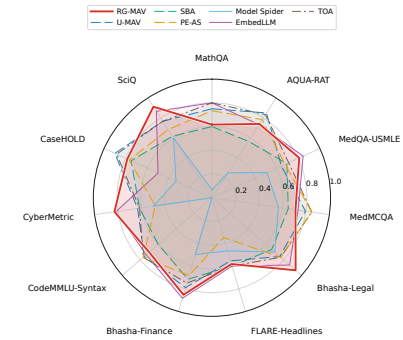


Figure 8: Robustness to alternative reference–target task partitions. Each row corresponds to one alternative suite. The left panel reports Kendall's τ_b on the target tasks, with the best result in each row bolded, and the right panel visualizes the corresponding per-task performance.

Table 12: Detailed effect of reference locality on RG-MAV. Kendall's τ_b is reported for each target task as the reference-neighborhood size k varies from 1 to 11.

Target task	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$
MathQA	**0.950**	0.886	0.814	0.802	0.783	0.790	0.773	0.773	0.773	0.768	0.756
Math-MCQA	**0.588**	0.542	0.555	0.565	0.547	0.549	0.565	0.565	0.565	0.561	0.565
MedQA-USMLE	**0.814**	0.785	0.746	0.723	0.728	0.728	0.728	0.712	0.711	0.728	0.728
PubMedQA	0.390	**0.410**	0.390	0.400	0.407	0.390	0.390	0.397	0.397	0.397	0.397
Bhasha-Legal	**0.933**	0.852	0.812	0.802	0.778	0.795	0.783	0.767	0.773	0.768	0.783
FLARE-Headlines	**0.577**	0.553	0.559	0.492	0.498	0.492	0.498	0.481	0.464	0.464	0.447
FLARE-Causal20	**0.397**	0.336	0.303	0.286	0.300	0.283	0.283	0.283	0.283	0.279	0.279
CodeMMLU-Syntax	0.728	0.773	0.773	0.773	**0.807**	**0.807**	**0.807**	**0.807**	**0.807**	**0.807**	**0.807**
CodeMMLU-Exec	0.510	0.532	**0.570**	0.549	0.538	0.532	0.532	0.525	0.538	0.538	0.538
ContractNLI-A	**0.769**	0.740	0.712	0.706	0.678	0.667	0.667	0.667	0.667	0.667	0.667
SciKnowEval	**0.795**	0.745	0.763	0.731	0.683	0.700	0.700	0.700	0.700	0.700	0.700
Avg.	**0.677**	0.650	0.636	0.621	0.613	0.612	0.611	0.607	0.607	0.607	0.606

Note. Bold values indicate the best Kendall's τ_b for each target task across neighborhood sizes.

Overall, these results show that the gains of RG-MAV are not tied to one specific reference-task configuration. They also suggest a practical deployment principle: maintaining a diverse reference pool spanning multiple task domains can improve the reliability of reference-guided auxiliary weighting for previously unseen target tasks.

D.6 Detailed Analysis of Reference Locality

The main results show that RG-MAV performs best with a highly local reference neighborhood. We provide the complete per-task results here to examine how the neighborhood size k affects model-ranking quality beyond the averaged trend. We vary k from 1 to 11, where $k = 1$ uses only the nearest reference task and larger values progressively incorporate a broader reference neighborhood. All other components of RG-MAV are kept unchanged.

Results. The detailed results confirm the aggregate locality trend reported in the main experiments. The mean Kendall's τ_b is highest at $k = 1$ (0.677), decreases to 0.650 at $k = 2$ and 0.636 at $k = 3$, and gradually approaches approximately 0.607 as the reference neighborhood becomes broader. Using all 11 reference tasks yields an average τ_b of 0.606, a decrease of 0.071 from the nearest-reference setting.

The effect is also pronounced on several individual targets. For example, Kendall's τ_b decreases from 0.950 to 0.756 on MathQA, from 0.933 to 0.783 on Bhasha-Legal, from 0.577 to 0.447 on FLARE-Headlines, and from 0.769 to 0.667 on ContractNLI-A as k increases from 1 to 11. This behavior is consistent with the locality assumption underlying RG-MAV: nearby reference tasks provide more informative evidence of auxiliary reliability, whereas progressively incorporating less related references can dilute this signal.

The trend need not hold for every individual target. For example, CodeMMLU-Syntax performs better with broader neighborhoods, while PubMedQA and CodeMMLU-Exec attain their best results at intermediate values of k. Nevertheless, the clear decrease in the cross-task average supports the use of a highly local reference neighborhood, with $k = 1$ providing the strongest overall performance.

E Algorithm Details

This section provides the pseudocode for the reference-guided multi-auxiliary voting method described in Section 4.1. All quantities follow the notation defined in the main text.

F The pairwise agreement profile

This appendix develops the argument in Section 3.2. We first specify the task model and the required harmonic identities. Invariant disintegration and conditional independence then reduce agreement to a rotation average, whose harmonic expansion gives the spectral representation. A uniform approx-

Algorithm 1 Reference-Guided Multi-Auxiliary Voting (RG-MAV)

Require: auxiliary models $\mathcal{A}$, candidate models $\mathcal{M}$, reference tasks $\mathcal{R}$ with evaluation scores $s_r(i)$, target task t, normalized PAVEs $\boldsymbol{q}_{a,i}, \boldsymbol{q}_{a,r}, \boldsymbol{q}_{a,t}$, neighborhood size k
Ensure: ranking of candidate models

for each auxiliary model $A_a \in \mathcal{A}$ **do**
  $\rho_{a,r} \leftarrow \langle \boldsymbol{q}_{a,t}, \boldsymbol{q}_{a,r} \rangle$ for all $R_r \in \mathcal{R}$
end for
for each candidate pair (M_i, M_j) with $i < j$ **do**
  $\sigma(r) \leftarrow \operatorname{sign}\big(s_r(i) - s_r(j)\big)$ for all $R_r \in \mathcal{R}$
  for each auxiliary model $A_a \in \mathcal{A}$ **do**
    $\boldsymbol{d}_a \leftarrow \boldsymbol{q}_{a,i} - \boldsymbol{q}_{a,j}$
    $v_a(t) \leftarrow \operatorname{sign}(\langle \boldsymbol{q}_{a,t}, \boldsymbol{d}_a \rangle)$
    $v_a(r) \leftarrow \operatorname{sign}(\langle \boldsymbol{q}_{a,r}, \boldsymbol{d}_a \rangle)$ for all $R_r \in \mathcal{R}$
    $c_{a,r} \leftarrow \frac{1}{2}\left[1 + \sigma(r) v_a(r)\right]$ for all $R_r \in \mathcal{R}$
    $\alpha_{a,r} \leftarrow 1 - 2\mathbb{I}[v_a(t) v_a(r) < 0]$ for all $R_r \in \mathcal{R}$
    $\mathcal{N}_{a,k} \leftarrow \operatorname{TopK}_{R_r \in \mathcal{R}} \left[\alpha_{a,r} \rho_{a,r}\right]$
    $\tilde{c}_{a,r} \leftarrow \frac{1}{2} + \alpha_{a,r}\left(c_{a,r} - \frac{1}{2}\right)$ for all $R_r \in \mathcal{N}_{a,k}$
    $w_a \leftarrow \frac{1}{|\mathcal{N}_{a,k}|} \sum_{R_r \in \mathcal{N}_{a,k}} \tilde{c}_{a,r}$
  end for
  $V \leftarrow \sum_{a=1}^{l} w_a v_a(t)$
  if $V > 0$ **then**
    record $M_i \succ_t M_j$
  else if $V < 0$ **then**
    record $M_j \succ_t M_i$
  else
    record a tie between M_i and M_j
  end if
end for
compute Copeland scores from all recorded pairwise preferences
return candidate models ranked by decreasing Copeland score

imation of the Gegenbauer polynomials yields the power-series surrogate. We use its nonnegative coefficients to establish monotonicity and the agreement bounds, and then discuss equality cases and statistical interpretation.

F.1 Notation and assumptions

Fix an auxiliary a and a model pair $i \neq j$. Let $d = d_a \geq 3$, and write $\boldsymbol{q}_t = \boldsymbol{q}_{a,t} \in \mathbb{S}^{d-1}$ for the normalized task PAVE. The evaluation-induced order is

$$\sigma_{ij}(t) = \operatorname{sign}(s_t(i) - s_t(j)) \in \{-1, +1\},$$

where unequal evaluation scores are assumed almost surely. Throughout the pairwise analysis, we suppress (i, j) and write $\sigma(t)$, m, p, Φ, δ, and R for the corresponding quantities. Table 13 summarizes the analytic notation.

For a task pair (t, t'), define its alignment $\rho = \langle \boldsymbol{q}_t, \boldsymbol{q}_{t'} \rangle$ and agreement profile

$$p(\rho) = \Pr[\sigma(t) = \sigma(t') \mid \rho], \qquad \Phi(\rho) = 2p(\rho) - 1. \tag{1}$$

Since $\sigma(t)\sigma(t')$ equals 1 on agreement and -1 on disagreement,

$$\Phi(\rho) = \mathbb{E}[\sigma(t)\sigma(t') \mid \rho]. \tag{2}$$

This conditional product moment is analogous to the noise-stability quantities studied in Boolean function analysis (O'Donnell, 2014). Conditional expectations initially determine the profile only λ-almost everywhere. Lemma F.5 provides the version used throughout the appendix.

We restate the assumptions of Section 3.2.

Table 13: Notation for the pairwise agreement analysis.

Symbol	Meaning
ς	Uniform probability measure on $\mathbb{S}^{d-1}$
dg	Haar probability measure on $O(d)$
$\|\|f\|\|_2$	Norm in $L^2(\mathbb{S}^{d-1}, \varsigma)$
$\langle x, y\rangle$, $\|\|x\|\|$	Euclidean inner product and norm
$\mathcal{H}_k^d$, $N(d,k)$	Degree-k real spherical harmonics and their dimension
Π_k	Orthogonal projection from $L^2(\varsigma)$ onto $\mathcal{H}_k^d$
$P_k^{(d)}$	Gegenbauer polynomial normalized by $P_k^{(d)}(1) = 1$
λ	Distribution of $\rho = \langle \boldsymbol{q}_t, \boldsymbol{q}_{t'} \rangle$

Assumption (I) (Isotropic exchangeable task geometry). For every $g \in O(d)$, $(g\boldsymbol{q}_t, g\boldsymbol{q}_{t'}) \overset{\mathrm{d}}{=} (\boldsymbol{q}_t, \boldsymbol{q}_{t'})$.

Assumption (CI) (Mediation and conditional independence). There exists a measurable $m\colon \mathbb{S}^{d-1} \to [-1, 1]$ such that, for every $u, v \in \{-1, +1\}$,

$$\Pr[\sigma(t) = u, \sigma(t') = v \mid \boldsymbol{q}_t, \boldsymbol{q}_{t'}] = \frac{1}{4}(1 + u\, m(\boldsymbol{q}_t))(1 + v\, m(\boldsymbol{q}_{t'}))$$

almost surely.

Assumption (CI) states both conditional independence of the order labels and dependence of each conditional label distribution only on its own task PAVE. In particular, $m(x) = \mathbb{E}[\sigma(t) \mid \boldsymbol{q}_t = x]$ almost everywhere. We call m the *order propensity*, and set

$$\delta = \mathbb{E}[\sigma(t)], \qquad R = ||m||_2^2.$$

The quantity δ records population ordering bias. A value of zero means that each ordering has marginal probability $1/2$; it does not require the ordering to be independent of the task PAVE.

Remark F.1 (Consequences of the assumptions). Under (I), each task PAVE has distribution ς, by uniqueness of the invariant probability measure on the sphere (Folland, 2016; Eaton, 1989). Thus isotropy is a population assumption in addition to normalization. Independently sampled task-label pairs with the same distribution satisfy (CI), but (CI) also permits dependent PAVEs. Shared latent structure can violate mediation or conditional independence. The sign and alignment dependence of the resulting discrepancy require further assumptions.

F.2 Classical harmonic identities

We collect the standard identities needed below (Müller, 1966; Atkinson & Han, 2012; Stein & Weiss, 1971). All function inner products in this subsection use the probability measure ς.

F1. Harmonic decomposition. The spaces $\mathcal{H}_k^d$ are finite dimensional and invariant under $O(d)$. They give the orthogonal decomposition

$$L^2(\varsigma) = \bigoplus_{k \geq 0} \mathcal{H}_k^d, \qquad f = \sum_{k \geq 0} \Pi_k f \text{ in } L^2(\varsigma), \qquad ||f||_2^2 = \sum_{k \geq 0} ||\Pi_k f||_2^2.$$

In particular, $\mathcal{H}_0^d$ consists of constants, $\Pi_0 f = \int f \, d\varsigma$, and $\mathcal{H}_1^d = \{x \mapsto \langle x, v\rangle : v \in \mathbb{R}^d\}$.

F2. Addition theorem. For any orthonormal basis $\{Y_{k,r}\}_{r=1}^{N(d,k)}$ of $\mathcal{H}_k^d$,

$$\sum_{r=1}^{N(d,k)} Y_{k,r}(x) Y_{k,r}(y) = N(d,k) P_k^{(d)}(\langle x, y\rangle). \tag{3}$$

Here $N(d, 0) = 1$ and $N(d,k) = (2k + d - 2)(k + d - 3)!/(k!(d-2)!)$ for $k \geq 1$.

F3. Uniqueness of the zonal harmonic. For fixed $x \in \mathbb{S}^{d-1}$, the elements of $\mathcal{H}_k^d$ invariant under all orthogonal transformations fixing x form a one-dimensional space. It is spanned by $y \mapsto P_k^{(d)}(\langle x, y\rangle)$.

F4. Uniform bound. For $s \in [-1,1]$, $|P_k^{(d)}(s)| \le 1$. Indeed, applying Cauchy–Schwarz to equation 3 and using its diagonal identity gives

$$N(d,k)|P_k^{(d)}(\langle x,y\rangle)| \le \Big(\sum_r Y_{k,r}(x)^2\Big)^{1/2}\Big(\sum_r Y_{k,r}(y)^2\Big)^{1/2} = N(d,k).$$

F5. Gegenbauer recurrence. The normalized polynomials satisfy

$$(k+d-2)P_{k+1}^{(d)}(s) = (2k+d-2)sP_k^{(d)}(s) - kP_{k-1}^{(d)}(s), \qquad k \ge 1, \tag{4}$$

with $P_0^{(d)}(s) = 1$ and $P_1^{(d)}(s) = s$ (Atkinson & Han, 2012; Szegő, 1975).

F6. Parity. Homogeneity gives $Y(-x) = (-1)^k Y(x)$ for $Y \in \mathcal{H}_k^d$, and consequently $P_k^{(d)}(-s) = (-1)^k P_k^{(d)}(s)$.

F7. Spherical gradient energy. The Laplace–Beltrami operator satisfies $\Delta_{\mathbb{S}^{d-1}} Y = -k(k+d-2)Y$ for $Y \in \mathcal{H}_k^d$. Thus, for $f \in H^1(\mathbb{S}^{d-1})$,

$$||\nabla_{\mathbb{S}^{d-1}} f||_2^2 = \sum_{k\ge 0} k(k+d-2)||\Pi_k f||_2^2.$$

F.3 Invariant conditioning and the order propensity

We first show that conditioning on alignment is equivalent to averaging over an orthogonal-group orbit. Assumption (CI) then identifies the function to be averaged.

Lemma F.2 (Transitivity on pairs)**.** *If $x, y, x', y' \in \mathbb{S}^{d-1}$ satisfy $\langle x, y\rangle = \langle x', y'\rangle$, there exists $g \in O(d)$ with $gx = x'$ and $gy = y'$. Hence $A_\rho = \{(x,y) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} : \langle x, y\rangle = \rho\}$ is a single $O(d)$-orbit for every $\rho \in [-1,1]$.*

Proof. For $\rho = \pm 1$, we have $y = \pm x$ and $y' = \pm x'$, so any orthogonal transformation taking x to x' suffices. For $|\rho| < 1$, the pairs (x, y) and (x', y') are linearly independent and have the same Gram matrix. The linear map sending x to x' and y to y' is therefore an isometry between their spans. Extending it by an isometry between the orthogonal complements gives the required element of $O(d)$. □

Lemma F.3 (Invariant disintegration)**.** *Let ν be an $O(d)$-invariant probability measure on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, and let λ be its image under $(x, y) \mapsto \langle x, y\rangle$. For each ρ, let ω_ρ be the law of (gx_ρ, gy_ρ) for Haar-distributed g and any $(x_\rho, y_\rho) \in A_\rho$. Then $\nu = \int \omega_\rho \, d\lambda(\rho)$. In particular, under (I), every bounded measurable F satisfies*

$$\mathbb{E}[F(\boldsymbol{q}_t, \boldsymbol{q}_{t'}) \mid \rho] = \int_{O(d)} F(gx_\rho, gy_\rho)\, dg \tag{5}$$

for λ-almost every ρ.

Proof. The sphere product is a compact metric space. The disintegration theorem (Kallenberg, 2002, Thm. 6.3) gives a conditional probability kernel ν_ρ supported on A_ρ for λ-almost every ρ. For fixed $g \in O(d)$, invariance of ν and preservation of the inner product imply that $g_*\nu_\rho$ is another such kernel. Uniqueness of disintegration yields $g_*\nu_\rho = \nu_\rho$ almost everywhere.

Choose a countable dense subset of $O(d)$ and remove the union of the corresponding exceptional sets. For each remaining ρ, weak continuity of $g \mapsto g_*\nu_\rho$ extends invariance to every $g \in O(d)$. Lemma F.2 identifies A_ρ as a compact homogeneous space, whose invariant probability measure is unique (Folland, 2016). Thus $\nu_\rho = \omega_\rho$ almost everywhere. The family ω_ρ is a measurable kernel, as can be seen by choosing $x_\rho = e_1$ and $y_\rho = \rho e_1 + \sqrt{1-\rho^2}e_2$. This proves the disintegration and equation 5. □

Under (I), the propensity is an element of $L^\infty(\varsigma)$ with $||m||_\infty \le 1$. Choose a Borel representative with values in $[-1,1]$.

Definition F.4 (Canonical profile)**.** For $\rho \in [-1, 1]$ and $(x_\rho, y_\rho) \in A_\rho$, define

$$\Phi(\rho) = \int_{O(d)} m(gx_\rho)m(gy_\rho)\, dg, \qquad p(\rho) = \frac{1 + \Phi(\rho)}{2}. \tag{6}$$

Lemma F.5 (Well-posedness and reduction)**.** *The integral in equation 6 is finite and independent of both the representative pair in A_ρ and the Borel version of m. Under (I) and (CI), it agrees λ-almost everywhere with equation 2.*

Proof. Boundedness of m gives finiteness. For each fixed $z \in \mathbb{S}^{d-1}$, the image of Haar measure under $g \mapsto gz$ is ς. Changing m on a ς-null set therefore leaves the integral unchanged. If another pair in A_ρ is obtained by applying $h \in O(d)$, right invariance of Haar measure shows that replacing g by gh leaves the integral unchanged. Such an h exists by Lemma F.2.

Assumption (CI) gives $\mathbb{E}[\sigma(t)\sigma(t') \mid \boldsymbol{q}_t, \boldsymbol{q}_{t'}] = m(\boldsymbol{q}_t)m(\boldsymbol{q}_{t'})$. The tower property and Lemma F.3, applied to $F(x, y) = m(x)m(y)$, identify the resulting conditional expectation with equation 6. □

We henceforth use the canonical profile at every alignment. At alignments outside the support of λ, its value is specified by the model through equation 6, rather than identified by the conditional expectation alone.

Lemma F.6 (Moment identities)**.** *Under (I) and (CI),*

$$\int m\, d\varsigma = \delta, \qquad \delta^2 \le R \le 1,$$

$$R - \delta^2 = \mathrm{Var}(m(\boldsymbol{q}_t)), \qquad 1 - R = \mathbb{E}[\mathrm{Var}(\sigma(t) \mid \boldsymbol{q}_t)].$$

Moreover, $R = 1$ if and only if $\sigma(t)$ is almost surely a function of $\boldsymbol{q}_t$, and $R = \delta^2$ if and only if m is constant almost everywhere.

Proof. The tower property gives $\int m\, d\varsigma = \mathbb{E}[\sigma(t)] = \delta$. Since $\sigma(t)^2 = 1$, the conditional variance equals $1 - m(\boldsymbol{q}_t)^2$. Taking expectations gives the identity for $1 - R$, while the definition of variance gives the identity for $R - \delta^2$. Nonnegativity of these variances yields the bounds and their equality cases. □

The quantity $R - \delta^2$ measures variation in the ordering explained by the task PAVE. Because the label is binary, a constant propensity means that its conditional distribution is independent of the PAVE.

F.4 Spectral representation

We next evaluate the rotation average by harmonic degree. The key step is that distinct degrees do not interact under averaging.

Lemma F.7 (Harmonic rotation averages)**.** *Let $u \in \mathcal{H}_k^d$, $v \in \mathcal{H}_l^d$, and $B_{u,v}(x, y) = \int_{O(d)} u(gx)v(gy)\, dg$. This function depends only on $\langle x, y\rangle$. If $k \neq l$, then $B_{u,v} = 0$. For $u \in \mathcal{H}_k^d$,*

$$\int_{O(d)} u(gx)u(gy)\, dg = ||u||_2^2 P_k^{(d)}(\langle x, y\rangle).$$

Proof. Right invariance of Haar measure gives $B_{u,v}(hx, hy) = B_{u,v}(x, y)$ for every $h \in O(d)$. Lemma F.2 therefore gives a function b such that $B_{u,v}(x, y) = b(\langle x, y\rangle)$.

For fixed x, the function $B_{u,v}(x, \cdot)$ lies in the finite-dimensional invariant space $\mathcal{H}_l^d$ and is invariant under transformations fixing x. By F3, it is a multiple of $P_l^{(d)}(\langle x, \cdot\rangle)$. Applying the same argument in the first variable shows that b is also a multiple of $P_k^{(d)}$. If $k \neq l$, these polynomials have different degrees, so both multiples must vanish.

When $v = u$, the same argument yields $B_{u,u}(x, y) = \beta P_k^{(d)}(\langle x, y\rangle)$. On the diagonal, $B_{u,u}(x, x) = \int u(gx)^2\, dg = ||u||_2^2$. Since $P_k^{(d)}(1) = 1$, we obtain $\beta = ||u||_2^2$. □

Theorem. *Under (I) and (CI), let $W_k = ||\Pi_k m||_2^2$. For every $\rho \in [-1, 1]$,*

$$\Phi(\rho) = \sum_{k\geq 0} W_k P_k^{(d)}(\rho), \qquad W_k \geq 0, \qquad \sum_{k\geq 0} W_k = R \leq 1, \qquad W_0 = \delta^2.$$

The series converges absolutely and uniformly on $[-1, 1]$. In addition, the kernel $(x, y) \mapsto \Phi(\langle x, y\rangle)$ is positive semidefinite.

Proof. Let $m_k = \Pi_k m$ and $S_K m = \sum_{k=0}^K m_k$. By F1 and Lemma F.6, $\sum_k W_k = ||m||_2^2 = R$ and $W_0 = (\int m\, d\varsigma)^2 = \delta^2$. The bound in F4 and the Weierstrass test imply absolute and uniform convergence of $\Psi(\rho) = \sum_k W_k P_k^{(d)}(\rho)$.

For $(x, y) \in A_\rho$, Lemma F.7 gives

$$\int_{O(d)} S_K m(gx) S_K m(gy)\, dg = \sum_{k=0}^K W_k P_k^{(d)}(\rho).$$

For every fixed $z \in \mathbb{S}^{d-1}$, $\int |f(gz)|^2\, dg = ||f||_2^2$. Applying Cauchy–Schwarz after adding and subtracting $S_K m(gx)m(gy)$ therefore yields

$$\left|\int_{O(d)} m(gx)m(gy)\, dg - \int_{O(d)} S_K m(gx) S_K m(gy)\, dg\right|$$
$$\leq (||m||_2 + ||S_K m||_2)||m - S_K m||_2 \leq 2||m||_2||m - S_K m||_2 \longrightarrow 0.$$

Thus $\Phi(\rho) = \Psi(\rho)$ for every ρ.

Finally, for any $x_1, \ldots, x_N \in \mathbb{S}^{d-1}$ and $c_1, \ldots, c_N \in \mathbb{R}$, Lemma F.5 gives

$$\sum_{r,s=1}^N c_r c_s \Phi(\langle x_r, x_s\rangle) = \int_{O(d)} \left(\sum_{r=1}^N c_r m(gx_r)\right)^2 dg \geq 0.$$

This proves positive semidefiniteness. □

The weights describe the harmonic energy of the propensity. Degree zero records ordering bias, degree one gives the linear component, and higher degrees describe nonlinear dependence on task location. Even a deterministic halfspace rule generally has energy at degrees greater than one.

Proposition F.8 (Realizable spectra)**.** *Every measurable $m\colon \mathbb{S}^{d-1} \to [-1, 1]$ is the propensity of a pairwise task model satisfying (I) and (CI). Consequently, the realizable spectra are exactly $\{(||\Pi_k m||_2^2)_{k\geq 0} : ||m||_\infty \leq 1\}$. For every finitely supported nonnegative sequence (b_k), there exists $c > 0$ such that $(c^2 b_k)$ is realizable.*

Proof. Draw $\boldsymbol{q}_t, \boldsymbol{q}_{t'}$ independently from ς. Given the PAVEs, draw the labels independently with conditional success probabilities $(1 + m(\boldsymbol{q}_t))/2$ and $(1 + m(\boldsymbol{q}_{t'}))/2$. This construction satisfies both assumptions and has the required propensity. Conversely, every model under consideration has a propensity bounded by one, which proves the characterization.

For the last assertion, choose $u_k \in \mathcal{H}_k^d$ with $||u_k||_2^2 = b_k$ and let $\widetilde{m} = \sum_k u_k$. The sum is finite and hence bounded on the sphere. Set $c = 1/\max\{1, ||\widetilde{m}||_\infty\}$ and $m = c\widetilde{m}$. Orthogonality gives $||\Pi_k m||_2^2 = c^2 b_k$. □

Remark F.9 (Relation to Schoenberg's theorem). Schoenberg's characterization identifies continuous isotropic positive semidefinite kernels on the sphere with summable nonnegative Gegenbauer coefficients (Schoenberg, 1942; Gneiting, 2013). The agreement profiles form a subclass because their coefficients must arise from a propensity bounded by one. For example, $\Phi(\rho) = \rho$ defines a positive semidefinite kernel and satisfies $\Phi(1) = 1$. Realizing it as an agreement profile would require $m(x) = \langle x, v\rangle$ with $||m||_2^2 = ||v||^2/d = 1$, whereas $||m||_\infty = ||v|| \leq 1$. These conditions are incompatible for $d > 1$.

Example F.10 (Halfspace propensity)**.** Let $m(x) = \operatorname{sign}\langle x, w\rangle$ for a fixed unit vector w. The choice of the sign on the equator is immaterial. For Haar-distributed g, the vector $g^{-1}w$ is uniform on the sphere. The random-hyperplane identity (Goemans & Williamson, 1995) therefore gives

$$p_{\mathrm{hp}}(\rho) = 1 - \frac{\arccos \rho}{\pi}, \qquad \Phi_{\mathrm{hp}}(\rho) = \frac{2}{\pi} \arcsin \rho.$$

Here $\delta = 0$ and $R = 1$. The odd propensity has no even-degree harmonic components, consistently with F6. Although the exact profile also has a nonnegative power-series expansion, its monomial coefficients need not equal its finite-dimensional Gegenbauer weights.

F.5 Approximation by a power series

The next lemma controls the effect of replacing each Gegenbauer polynomial by a monomial. Both the weights and the basis may depend on dimension; the approximation holds for each fixed auxiliary.

Lemma F.11 (Uniform monomial approximation)**.** *For $d \geq 3$ and $k \geq 0$,*

$$\sup_{s \in [-1,1]} |P_k^{(d)}(s) - s^k| \leq \frac{k(k-1)}{d-2}.$$

Proof. Set $\beta_k = k/(k+d-2)$. The recurrence equation 4 becomes

$$P_{k+1}^{(d)}(s) = sP_k^{(d)}(s) + \beta_k\big(sP_k^{(d)}(s) - P_{k-1}^{(d)}(s)\big).$$

Let $E_k(s) = P_k^{(d)}(s) - s^k$ and $e_k = \sup_{|s|\leq 1} |E_k(s)|$. Subtracting s^{k+1} and using F4 gives $e_{k+1} \leq e_k + 2\beta_k$. Since $e_0 = e_1 = 0$ and $\beta_k \leq k/(d-2)$,

$$e_k \leq 2\sum_{j=1}^{k-1} \beta_j \leq \frac{2}{d-2} \sum_{j=1}^{k-1} j = \frac{k(k-1)}{d-2}.$$

□

Corollary F.12 (Power-series approximation)**.** *Under (I) and (CI), define*

$$\Phi^\sharp(\rho) = \sum_{k\geq 0} W_k \rho^k, \qquad p^\sharp(\rho) = \frac{1 + \Phi^\sharp(\rho)}{2}, \qquad \mathcal{E}_2 = \sum_{k\geq 0} k(k-1)W_k.$$

The series converges absolutely and uniformly on $[-1, 1]$. If $\mathcal{E}_2 < \infty$, then

$$||\Phi - \Phi^\sharp||_{\infty,[-1,1]} \leq \frac{\mathcal{E}_2}{d-2}, \qquad ||p - p^\sharp||_{\infty,[-1,1]} \leq \varepsilon, \qquad \varepsilon = \frac{\mathcal{E}_2}{2(d-2)}.$$

If $m \in H^1(\mathbb{S}^{d-1})$, then $\mathcal{E}_2 \leq ||\nabla_{\mathbb{S}^{d-1}} m||_2^2$. If $W_k = 0$ for $k > K$, then $\mathcal{E}_2 \leq RK(K-1)$.

Proof. Uniform convergence follows from $|W_k\rho^k| \leq W_k$ and $\sum_k W_k = R$. Applying Lemma F.11 termwise gives

$$|\Phi(\rho) - \Phi^\sharp(\rho)| \leq \sum_k W_k |P_k^{(d)}(\rho) - \rho^k| \leq \frac{\mathcal{E}_2}{d-2}.$$

The bound for p follows by division by two. The gradient estimate follows from F7 and $k(k-1) \leq k(k+d-2)$. The final estimate follows from $k(k-1) \leq K(K-1)$ on the spectral support. □

Remark F.13 (Conditions for a small approximation error). Large dimension alone does not ensure a small error. The preceding bound is useful when $\mathcal{E}_2/(d-2)$ is small. For example, if $d = 1.6 \times 10^6$ and $W_k = 0$ for $k > 10$, the error in Φ is below 6×10^{-5}. This illustrates the bound under a specified spectral cutoff; it does not estimate the spectrum of the experimental propensities. If $\mathcal{E}_2 = \infty$, the surrogate remains well defined, but this estimate provides no finite error bound.

F.6 MONOTONICITY AND PARITY

A function is absolutely monotone on an interval if it has derivatives of every order there and each derivative, including the function itself, is nonnegative. Nonnegativity of the spectral weights gives this property for $\Phi^\sharp$ on $[0, 1)$.

Corollary F.14 (Monotonicity for nonnegative alignment)**.** *Under (I) and (CI), $\Phi^\sharp$ is absolutely monotone on $[0, 1)$, and $p^\sharp$ is nondecreasing and convex on $[0, 1]$. It is strictly increasing on $[0, 1]$ if and only if m is not almost everywhere constant. If $\mathcal{E}_2 < \infty$, then for $0 \le \rho \le \rho' \le 1$,*

$$p(\rho') - p(\rho) \ge -2\varepsilon.$$

Proof. Since $\sum_k W_k < \infty$, the series can be differentiated termwise on $(-1, 1)$. For every integer $r \ge 0$ and $0 \le \rho < 1$,

$$\frac{d^r}{d\rho^r}\Phi^\sharp(\rho) = \sum_{k \ge r} W_k \frac{k!}{(k-r)!}\rho^{k-r} \ge 0.$$

Uniform convergence gives continuity at 1 and extends monotonicity and convexity to $[0, 1]$. If some $W_k > 0$ for $k \ge 1$, then for $0 \le \rho < \rho' \le 1$ the corresponding term satisfies $W_k\rho^k < W_k(\rho')^k$. Conversely, if all such weights vanish, F1 implies that m is constant almost everywhere and $p^\sharp$ is constant. Finally, Corollary F.12 gives

$$p(\rho') - p(\rho) \ge p^\sharp(\rho') - p^\sharp(\rho) - 2\varepsilon \ge -2\varepsilon.$$

□

This result establishes exact monotonicity for the surrogate and an approximate monotonicity inequality for the finite-dimensional profile. Neither statement asserts monotonicity at negative alignments.

Lemma F.15 (Parity)**.** *The following are equivalent: $m(-x) = m(x)$ for ς-almost every x; $W_k = 0$ for every odd k; and Φ is even. The corresponding equivalence holds between odd m, vanishing even-degree weights, and odd Φ. The same parity characterizations hold for $\Phi^\sharp$.*

Proof. By F6, the antipodal map acts on $\mathcal{H}_k^d$ as multiplication by $(-1)^k$. Completeness in F1 therefore identifies even functions with vanishing odd-degree components, and odd functions with vanishing even-degree components. Since $W_k = ||\Pi_k m||_2^2$, this proves the equivalences involving m and the weights.

The Gegenbauer polynomials are orthogonal under the distribution of the inner product of two independent uniform points. Uniqueness of the expansion, together with $P_k^{(d)}(-\rho) = (-1)^k P_k^{(d)}(\rho)$, proves the assertions for Φ. Uniqueness of a power series gives the corresponding assertions for $\Phi^\sharp$. □

Example F.16 (A quadratic propensity)**.** Fix a unit vector w and let

$$m(x) = P_2^{(d)}(\langle x, w\rangle) = \frac{d\langle x, w\rangle^2 - 1}{d-1}.$$

By F4, $||m||_\infty \le 1$, so Proposition F.8 makes this an admissible propensity. The addition theorem gives $||m||_2^2 = 1/N(d, 2)$, and $m \in \mathcal{H}_2^d$. Thus $W_2 = 1/N(d, 2)$ and all other weights vanish. The exact profile is

$$\Phi(\rho) = \frac{d\rho^2 - 1}{N(d, 2)(d-1)}, \qquad p(\rho) = \frac{1}{2} + \frac{d\rho^2 - 1}{2N(d, 2)(d-1)}.$$

It is even and strictly decreasing on $[-1, 0]$. In particular, $p(-1) = p(1) > 1/2$, whereas $p(0) < 1/2$. Since $\delta = 0$, this example also shows why the surrogate lower bound at zero alignment requires an approximation correction for the exact profile.

Example F.17 (A parity order rule)**.** For orthonormal w_1, w_2, consider the deterministic propensity

$$m(x) = \operatorname{sign}\langle x, w_1\rangle \operatorname{sign}\langle x, w_2\rangle.$$

It is even, has mean zero, and satisfies $m^2 = 1$ almost everywhere. Hence $\delta = 0$, $R = 1$, and both profiles are even. The exact profile satisfies $p(\pm 1) = 1$, while the surrogate satisfies $p^\sharp(\pm 1) = 1$ and $p^\sharp(0) = 1/2$.

For comparison, let (G, H) be jointly Gaussian with zero mean, marginal covariance I_d, and cross-covariance ρI_d. In coordinates aligned with w_1, w_2, the pairs (G_1, H_1) and (G_2, H_2) are independent. The Gaussian sign identity (Vershynin, 2018; Goemans & Williamson, 1995) yields

$$\Phi_{\mathrm{G}}(\rho) = \left(\frac{2}{\pi}\arcsin\rho\right)^2, \qquad p_{\mathrm{G}}(\rho) = \frac{1 + \Phi_{\mathrm{G}}(\rho)}{2}.$$

This is a different coupling. Its monomial coefficients are Gaussian Hermite weights and need not equal the spherical weights W_k. The calculation therefore does not identify Φ_{G} with $\Phi^\sharp$, or provide an error estimate for the discontinuous spherical propensity.

Remark F.18 (Interpretation of negative alignment). Even propensities show that negative alignment can retain substantial ordering agreement. Global monotonicity requires further restrictions on the combined spectral contributions. In particular, nonnegativity of the weights alone is insufficient, while vanishing of all positive even-degree weights is not a necessary condition for global monotonicity.

F.7 Agreement bounds and equality cases

The approximation and monotonicity results lead to the bounds.

Corollary. *Under (I) and (CI), for $\rho \in [0, 1]$,*

$$\frac{1 + \delta^2}{2} \le p^\sharp(\rho) \le \frac{1 + \delta^2 + (R - \delta^2)\rho}{2}.$$

If $\mathcal{E}_2 < \infty$, then

$$\frac{1 + \delta^2}{2} - \varepsilon \le p(\rho) \le \frac{1 + \delta^2 + (R - \delta^2)\rho}{2} + \varepsilon.$$

Proof. For $k \ge 1$ and $\rho \in [0, 1]$, $0 \le \rho^k \le \rho$. Hence

$$\delta^2 = W_0 \le \Phi^\sharp(\rho) \le W_0 + \rho \sum_{k \ge 1} W_k = \delta^2 + (R - \delta^2)\rho.$$

The upper bound also follows from convexity in Corollary F.14, since

$$\Phi^\sharp(\rho) \le (1 - \rho)\Phi^\sharp(0) + \rho\Phi^\sharp(1) = (1 - \rho)\delta^2 + \rho R.$$

Applying $z \mapsto (1 + z)/2$ gives the surrogate bounds, and Corollary F.12 gives the bounds for p. □

The surrogate lower bound is attained at $\rho = 0$, and the upper bound at both endpoints. Their equality conditions follow directly from nonnegativity of the coefficients. For any $0 < \rho \le 1$, equality in the lower bound holds if and only if $W_k = 0$ for all $k \ge 1$, or equivalently m is constant almost everywhere. For any $0 < \rho < 1$, equality in the upper bound holds if and only if $W_k = 0$ for all $k \ge 2$. Indeed, its gap is

$$\frac{1}{2} \sum_{k \ge 2} W_k(\rho - \rho^k),$$

and every parenthesized term is strictly positive in the interior.

Thus equality in the upper bound at an interior alignment characterizes an affine propensity

$$m(x) = \delta + \langle x, v\rangle, \qquad R - \delta^2 = \frac{||v||^2}{d}, \qquad |\delta| + ||v|| \le 1.$$

The last condition is exactly the requirement $||m||_\infty \le 1$. Endpoint attainment alone does not imply that the bounds are optimal at every interior alignment for fixed δ, R, and d.

For the halfspace profile in Example F.10, the exact inequality $1/2 \le p_{\mathrm{hp}}(\rho) \le (1 + \rho)/2$ follows directly from $0 \le (2/\pi)\arcsin\rho \le \rho$ on $[0, 1]$. The latter upper bound follows from convexity of $\arcsin$ and its endpoint values. This verifies the bound for that example without identifying its exact and surrogate profiles.

F.8 STATISTICAL INTERPRETATION AND SCOPE

The surrogate lower bound $(1+\delta^2)/2$ accounts for population ordering bias. The slope of its upper bound is

$$\frac{R-\delta^2}{2} = \frac{1}{2}\operatorname{Var}(m(\boldsymbol{q}_t)),$$

which measures variation in the ordering explained by the auxiliary. At the canonical endpoints,

$$p^\sharp(0) = \frac{1+\delta^2}{2}, \qquad p(1) = p^\sharp(1) = \frac{1+R}{2}.$$

When $\mathcal{E}_2 < \infty$, $p(0)$ differs from the first value by at most ε. Also, $1 - R = \mathbb{E}[\operatorname{Var}(\sigma(t) \mid \boldsymbol{q}_t)]$ measures the residual ordering uncertainty given the PAVE.

To relate the pairwise analysis to rankings, suppose the assumptions hold for every pair and that evaluation rankings are tie free. With n candidate models, define

$$\tau_{\mathrm{gold}}(t,t') = \binom{n}{2}^{-1} \sum_{i<j} \sigma_{ij}(t)\sigma_{ij}(t').$$

Restoring pair indices, linearity of expectation gives a canonical conditional mean equal to $\binom{n}{2}^{-1}\sum_{i<j}\Phi_{ij}(\rho)$. At zero alignment, provided every $\mathcal{E}_2^{(ij)}$ is finite,

$$\left|\mathbb{E}[\tau_{\mathrm{gold}}(t,t') \mid \rho = 0] - \binom{n}{2}^{-1} \sum_{i<j} \delta_{ij}^2\right| \le \binom{n}{2}^{-1} \sum_{i<j} \frac{\mathcal{E}_2^{(ij)}}{d-2}.$$

The conditional mean at a null alignment is understood through this canonical version. The baseline biases can be estimated from evaluation scores, whereas estimating R_{ij} requires conditional-propensity estimation or suitable conditionally independent replicates. Splitting a benchmark does not itself guarantee either nearly identical PAVEs or the required conditional independence.

The results describe a population model under (I) and (CI). They need not hold for an arbitrary finite benchmark collection. Moreover, the bounds constrain conditional mean agreement, rather than individual task-pair observations. Empirical evaluation should account for the approximation error and for dependence among pairs sharing a task. Under independent task sampling, averages over all pairs are U-statistics, so treating those pairs as independent observations generally gives incorrect uncertainty estimates. Finally, agreement between evaluation-induced orderings does not alone establish transfer of an auxiliary's ranking accuracy.