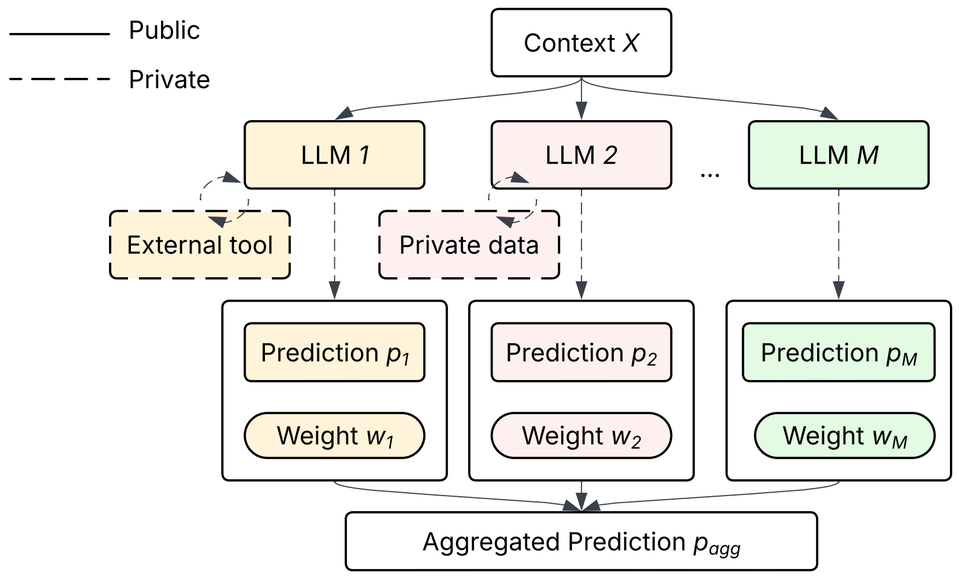
Decentralized Aggregation of LLM Predictions via Wagering Mechanisms

Yuhong Luo David M. Pennock Xintong Wang

Scan for Paper

Aggregation of LLM Predictions



Research question: Given a question of interest, can we design a decentralized mechanism to incentivize truthful predictions and elicit expertise-based aggregation weights from participants, for better collective predictions?

Wagering Mechanisms

These goals inspire us to adopt a wagering mechanism:

- Each participant reports a prediction p_i and places a wager w_i.
- The mechanism rewards predictions based on *a scoring rule*, scaled by the amount wagered.

The weighted-score wagering mechanisms (WSWM) [1] adopts the following net payout function

$$\pi_i^{\text{WSWM}}(\mathbf{p}, \mathbf{w}, y) = w_i \left(s(p_i, y) - \frac{\sum_{j=1}^{M} w_j s(p_j, y)}{\sum_{j=1}^{M} w_j} \right), \quad (1)$$

where $\mathbf{p}$ is the prediction profile, $\mathbf{w}$ the wagering profile, y the revealed outcome, and s a strictly proper scoring rule, such as the *Brier score*.

WSWM satisfies properties including incentive compatibility (IC) under immutable beliefs, individual rationality (IR), budget balance, etc.

Advantage-Aligned Wagering Mechanisms for LLM Aggregation (WALLA)

New properties may be needed for LLM or learning-based agents:

- Dominant-strategy IC of predictions under any belief structure.
- The net payout function can be treated as an objective for optimization.
- Bounded optimal wager, reflecting the expected score advantage that can be used as an agent's weight in prediction aggregation.

We propose WALLA with the following net payout function:

$$\pi_i(\mathbf{p}, \mathbf{w}, y) := w_i(s(p_i, y) - b_{-i}(\mathbf{p}_{-i}, \mathbf{w}_{-i}, y) - cw_i), \quad (2)$$

where c is a mechanism parameter and b_{-i} is the *leave-one-out baseline* that depends only on other agents' predictions and wagers, but not on (p_i, w_i).

Properties of WALLA

As an agent optimizes for net payout, our proposed mechanism achieves four desirable properties:

- DSIC of prediction under both mutable and immutable beliefs.
- Advantage–wager alignment: the best-response wager is proportional to the agent's expected score advantage.
- Prediction-agnostic wager optimization: the best-response wager is well-defined for any fixed prediction.
- Constant worst-case deficit of WALLA with a bounded scoring rule.

Learning a Wager Policy from Payouts

WALLA enables a learning agent to learn the optimal wager using the mechanism's payout as feedback, even when opponents' predictions, wagers, and payouts are not observed.

Learning the wager while freezing prediction is desirable because

- It is more realistic for an LLM with domain expertise to learn when and how much to wager than to become a universally capable predictor.
- The wager network can be a lightweight prediction head on the LLM's hidden states without LLM finetuning.

Empirical Evaluation

Our evaluation spans three scenarios:

1. Homogeneous models (common prior) with distinct private contexts.
2. Heterogeneous models (different priors).
3. Heterogeneous models with distinct private contexts.

Table 1. Categorization of baselines.

Methods	Uncertainty-Aware	Decentralized	Learning	IC / Manipulation-Resistant
Pre-inference Routers	✗	✗	✓	–
Stacked Generalization	✓	✗	✓	–
Uniform Averaging	✗	✓	✗	✗
Weight by Model Confidence	✓	✓	✗	✗
Weight by Context Perplexity	✓	✓	✗	✗
WALLA (Ours)	✓	✓	✓	✓

Results

WALLA matches centralized aggregation baselines (e.g., stacked generalization) on QA and forecasting benchmarks, while simultaneously achieving advantage-weighted aggregation, uncertainty-awareness, fully decentralized learning, and incentive-compatibility guarantees.

Table 2. Four heterogeneous models with private contexts on the PubMedQA.

Method	One Randomly Assigned Relevant Context (Metrics ×100)					
	AUC ↑	ACC ↑	ECE ↓	MRR ↑	K-Tau ↑	DRegret ↓
UniformAvg	73.43	76.46	11.51	-	0.00	24.28
SelfCertainty	69.84	77.25	3.13	86.11	46.30	21.60
PackLLM	77.85	83.60	15.38	66.42	12.32	18.05
RouterDC w/ context	73.33±0.63	76.55±1.26	11.58±1.91	54.66±37.92	10.64±64.71	24.23±0.34
NIRTRouter w/ context	73.20±0.58	76.92±0.84	12.04±7.80	43.90±50.11	-7.29±152.52	24.15±3.73
RouteLLM w/o context	72.15±5.90	76.27±0.21	8.26±1.51	55.46±56.78	11.11±36.00	23.09±3.05
RouteLLM w/ context	74.46±0.57	79.11±0.59	5.66±1.40	60.84±2.83	14.24±2.72	19.18±0.32
StackedGen	79.28±0.36	81.02±0.20	5.79±0.24	68.41±0.10	23.03±0.11	15.72±0.12
WALLA	79.76±0.65	82.11±0.18	11.58±0.52	72.59±0.75	28.06±0.59	16.95±0.05

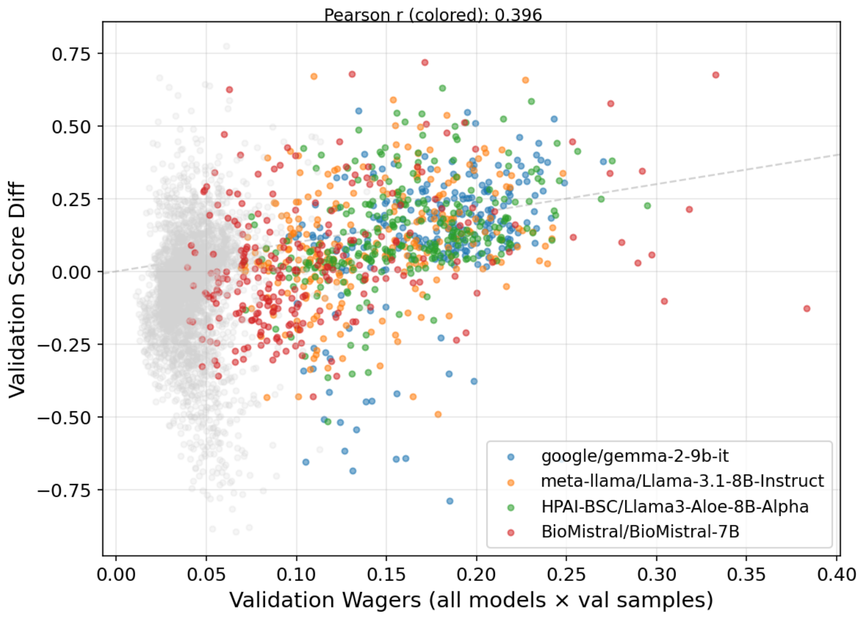


Figure 1. Advantage–wager alignment for WALLA on Scenario 3 with heterogeneous models and one randomly assigned relevant context. Colored dots highlight models with contexts.

[1] Nicolas S. Lambert et al. "An axiomatic characterization of wagering mechanisms". In: *Journal of Economic Theory* 156 (2015), pp. 389–416.

y.luo@rutgers.edu dpennock@dimacs.rutgers.edu xintong.wang@rutgers.edu