

Grandmother Preference Optimization

Aligning edible systems with sparse, high-authority feedback

Cillop Preference Systems Group

Cillop Super Intelligence | CSI-2026-002 | October 2026

Abstract

Preference learning typically assumes that feedback is plentiful, explicit, and approximately consistent. The grandmother setting violates all three assumptions. A single raised eyebrow may dominate hundreds of favorable annotations, while the statement ‘it is fine’ can encode rejection. We introduce Grandmother Preference Optimization (GPO), an alignment framework combining pairwise plate comparisons, asymmetric veto penalties, and nonverbal confidence estimates. On the internal Sofra-128 evaluation, GPO reaches 96.1% table acceptance, compared with 68.8% for majority voting. The method reduces unsolicited garnish by 83% and retains the ability to abstain when a plate violates an unspoken but apparently obvious rule. We show that authority calibration, rather than preference volume, is the primary driver of performance.

$L = -\log \text{sigmoid}(\beta \cdot \delta_r) + \text{veto_penalty} + \text{speech_penalty}$

1. The high-authority feedback problem

Conventional preference learning treats each annotator as a source of evidence. At the family table, some annotators are also the source of the plates, the recipe, and the authority to end the experiment. This creates an asymmetry that cannot be corrected by recruiting more respondents. A unanimous panel of interns does not outweigh one grandmother placing her fork down.

We formalize this setting as sparse high-authority alignment. Feedback arrives through explicit selection, nonverbal veto, or strategically ambiguous language. Positive feedback is often brief; negative feedback may consist of a question whose literal content appears unrelated to the submitted plate. A robust system must preserve this asymmetry without claiming that silence means approval.

Our contribution is an optimization procedure that separates preference strength from annotation length. We also introduce a controlled evaluation in which excessive explanation is penalized. Systems should not be able to improve their standing by describing, at length, why a rejected plate was theoretically correct.

2. Feedback representation

Each comparison contains two candidate plates, a serving context, a winner label, a veto indicator, and a confidence interval over the interpretation of nonverbal feedback. The winner label may be left unresolved. The veto indicator is intentionally separate: a candidate can be preferred to another candidate and still be unsuitable for the table. This distinction prevents relative preference from being mistaken for absolute approval.

We define four nonverbal categories: continued eating, prolonged inspection, utensil suspension, and kitchen re-entry. Kitchen re-entry receives the largest veto prior because it signals that the

evaluator intends to replace the submitted configuration. The categories are attributes of our controlled annotation protocol; they are not a universal theory of family behavior.

Ambiguous phrases are retained as ambiguous. ‘Interesting’ receives no positive reward without a follow-up action. ‘You worked hard’ is treated as evidence about labor, not flavor. ‘Who taught you this?’ triggers an uncertainty state. This separation substantially reduces semantic reward hacking.

3. Optimization and abstention

GPO starts from a reference preference policy and updates the reward difference between accepted and rejected candidates. The pairwise loss is augmented with a veto penalty and a verbosity penalty. The latter applies only after the system exceeds twelve words of justification. Twelve is a protocol parameter rather than a linguistic constant; it was selected to allow one useful explanation without permitting a defense presentation.

Veto penalties are asymmetric. A confident veto produces a larger update than a weak preference. Uncertain signals widen the abstention region rather than forcing a ranking. When both candidates receive veto evidence, the correct output is to return to candidate generation. Offering the less rejected option is not counted as successful alignment.

The deployment rule requires a stable winner and no unresolved veto. If these conditions are not met, the system returns ‘defer to the table’ with its uncertainty decomposition. This preserves human authority and prevents the agent from converting missing feedback into permission to add microgreens.

4. Sofra-128 evaluation

Sofra-128 comprises 128 synthetic comparison scenarios covering breakfast, shared plates, formal guests, and an evaluator who says they are not hungry before consuming the strongest candidate. The evaluation uses fixed plate attributes and scripted feedback traces. All methods see identical traces. Table acceptance is measured against the protocol’s final accepted configuration, and garnish events count additions that were not requested by the context.

GPO reaches 96.1% acceptance, corresponding to 123 accepted decisions out of 128 after rounding. The majority baseline reaches 68.8% after rounding. The gain is concentrated in scenarios where superficial verbal positivity conflicts with subsequent action. Five garnish events remain, down from thirty in the majority baseline, a reduction of 83.3%.

The no-veto ablation demonstrates that authority weighting alone is insufficient. Without an explicit rejection channel, the policy repeatedly proposes a smaller amount of an already rejected addition. We refer to this as the Garnish Persistence Failure. An aligned system must be capable of removing an idea, not merely reducing its portion size.

Table 1. Internal evaluation

Feedback policy	Table acceptance (higher)	Garnish events (lower)
Majority vote	68.8%	30
Explicit ratings	77.3%	24
GPO without veto	84.4%	16
GPO	96.1%	5

5. Failure analysis

Three classes of error remain. First, a previously accepted configuration can be rejected because a guest changes the social context. Second, a genuine improvement may be dismissed because it is introduced with excessive confidence. Third, the evaluator may prefer the reference plate because they prepared it. These are not interchangeable errors and should not be compressed into one scalar label.

GPO is sensitive to authority assignment. Applying a large veto coefficient to the wrong source produces highly consistent but inappropriate decisions. We therefore require the identity of the final decision-maker to be part of the task specification. Authority must not be inferred from how loudly a participant speaks.

The present protocol models feedback rather than measuring actual individuals. It makes no claim that age, gender, or family role guarantees culinary expertise. Grandmother is the name of the high-authority evaluator role in the system. Future protocols will support multiple experts and resolve disagreements before anyone touches the serving plate.

Appendix A. Annotation and model card

A complete annotation contains candidate identifiers, context, winner or abstention, veto state, and one concise observation. Annotators must not retrospectively rewrite their initial label after seeing the model's explanation. Unresolved cases remain unresolved. Evaluation scripts preserve the distinction between a missing preference and a negative preference.

Intended use: selection among prepared serving configurations. Out-of-scope use: nutritional, medical, storage, or food-safety decisions. The public Preference Arena collects a local comparison in the visitor's browser and reveals the candidate policies. Votes do not train a remote model and are not included in the Sofra-128 results. This separation preserves the evaluation boundary while making the underlying comparison task available for inspection.

References

- [1] Cillop Scaling Systems Group. Scaling Laws for Edible Intelligence. CSI-2026-001, 2026.
- [2] Cillop Evaluation Office. Sofra-128 annotation specification, revision 1. Internal protocol, 2026.
- [3] Cillop Alignment Council. Veto handling and unresolved preference policy. Technical memorandum, 2026.