Measuring Procedural Reuse Judgments with Response Curves

Anonymous authors
Paper under double-blind review

Abstract

Procedural knowledge such as recipes, protocols, and workflows is often adapted rather than executed directly. Understanding such reuse behavior requires analyzing how alternative procedures are comparatively evaluated for their applicability.

We introduce a measurement framework for studying *reuse judgments*, defined as comparative decisions about which of two candidate procedures is preferred for reuse with respect to a reference procedure under a given context. We construct controlled differences between candidate procedures along predefined structural axes and aggregate pairwise decisions into *response curves*, which characterize how alignment probabilities vary with the magnitude of differences along each axis.

We instantiate this framework in the domain of cooking recipes and use a large language model as a controlled decision-maker to generate comparative judgments at scale. The resulting response curves are stable under sampling variation, vary across structural axes and reference procedures, form structured but overlapping patterns, and shift systematically under contextual constraints.

These results indicate that, under this controlled evaluation setup, reuse judgments cannot be fully captured by a single similarity score, but instead exhibit structured variation in alignment probability over procedural differences. More broadly, this work proposes a measurement-oriented perspective on procedural reuse, enabling direct observation of how reuse judgments vary under controlled conditions.

1 Introduction

Procedural knowledge is commonly shared through artifacts such as cooking recipes, experimental protocols, and operational workflows.

A large body of work has focused on representing and interpreting procedural knowledge in structured, machine-readable forms. This includes the interpretation of instructional texts such as recipes (Kiddon et al., 2015), machine reading of wet-lab protocols (Kulkarni et al., 2018), formal semantic representations of biomedical protocols (Soldatova et al., 2008; 2014), and workflow modeling and process mining (van der Aalst & van Hee, 2002; Van Der Aalst, 2011). More recent work has explored the automatic construction of domain-specific representations of procedural knowledge (Shi et al., 2024).

These approaches primarily aim to support representation, interpretation, or execution. In such settings, the relevant question is not only whether a procedure can be executed, but which alternative procedure is *preferred for reuse* with respect to a given reference.

Execution-based evaluation captures one notion of reuse, where a procedure is applied without modification. However, it does not capture situations in which procedures are evaluated for their potential applicability under adaptation. Evaluations based solely on execution outcomes therefore provide limited insight into how procedural alternatives are comparatively assessed for reuse.

A key challenge is that such reuse judgments are inherently context-dependent and not directly observable from procedural descriptions alone. To address this, we introduce a measurement framework that makes reuse judgments observable through controlled pairwise comparisons between candidate procedures.

We define *reuse judgments* as decisions about which of two candidate procedures is preferred for reuse with respect to a reference procedure under a given context. This formulation enables us to elicit relative preferences when absolute judgments are difficult to specify.

A central aspect of our approach is the introduction of *structural axes*, each of which defines a specific aspect of comparison between procedures (e.g., step similarity or ingredient similarity). These axes are not intended as predictive features, but as measurement coordinates that define the space in which reuse judgments are observed. For each axis, we define *contrast* as the degree to which two candidate procedures differ with respect to that aspect. By operating within an existing procedural corpus, we can vary these contrasts in a controlled manner while holding other factors fixed, enabling systematic comparison across axes within a shared procedural context.

Under this framework, individual decisions are treated as noisy observations of an underlying evaluation process. By aggregating decisions across different levels of contrast, we obtain response patterns that characterize how reuse judgments vary with contrast along each axis, without assuming a fixed or monotonic relationship. In practice, contrasts are grouped into bins, allowing response curves to be estimated as functions of contrast.

This perspective is related to classical models of comparative judgment (Thurstone, 1927) and to measurement paradigms in psychophysics, where responses are analyzed as functions of controlled stimulus differences. Psychometric functions are commonly used to characterize how response probabilities vary with stimulus differences in such settings, and their estimation and statistical properties have been studied extensively in psychophysics (Wichmann & Hill, 2001). Inspired by these traditions, we treat reuse judgments not as scalar scores but as structured response patterns that can be analyzed under controlled variation. Unlike many standard psychometric analyses that assume monotonic relationships between stimulus differences and responses, we do not impose such constraints here.

We implement this framework using a large language model (LLM) as a decision-maker to generate comparative judgments at scale under controlled conditions, following prior work on using LLMs as evaluators (Zheng et al., 2023; Liu et al., 2023). In the domain of cooking recipes, we construct candidate procedures that vary along structural axes such as step similarity, ingredient similarity, ingredient rarity, and style variation. Aggregating decisions within contrast bins yields response curves that characterize how alignment probability varies with contrast along each axis.

We find that reuse judgments exhibit stable response patterns, vary systematically across structural axes and reference procedures, form structured but overlapping distributions, and shift under contextual constraints.

To address this, we make the following contributions:

- A measurement framework that operationalizes reuse judgments through controlled pairwise comparisons over naturally occurring procedural data.
- An analysis showing that aggregated pairwise decisions induce response curves that capture systematic variation in reuse judgments over contrast magnitude, rather than as a single scalar quantity, enabling comparison across multiple axes within a shared procedural space.
- Empirical evidence that contextual factors systematically shift these response patterns.

By separating reuse judgments from execution-based evaluation, this work provides a measurement-oriented perspective on procedural comparison, with implications for systems that generate, adapt, or select procedures under varying constraints.

2 A Measurement Framework for Reuse Evaluation

2.1 Problem Setup

Let $\mathcal{R}$ denote a set of procedural artifacts. Each artifact $r \in \mathcal{R}$ represents a procedure described by structured information consisting of (i) a sequence of procedural steps and (ii) associated procedural attributes.

Depending on the domain, these attributes may include required resources (e.g., ingredients or tools) and descriptions specifying the intended goal of the procedure.

Our objective is to analyze how a decision-maker makes *reuse judgments* over candidate procedures relative to a reference procedure. A reuse judgment refers to the judged applicability of a candidate procedure for reuse with respect to a reference procedure under a given context.

In this work, we instantiate the framework in the domain of cooking recipes (Section 3).

2.2 Measurement Construction

For a reference procedure $r_q \in \mathcal{R}$, a decision-maker is presented with two candidate procedures $\{r_1, r_2\} \subset \mathcal{R} \setminus \{r_q\}$, with $r_1 \neq r_2$, and selects the candidate that is *preferred for reuse* with respect to the reference r_q.

Each evaluation instance is represented as a trial

$$\tau = (r_q, r_1, r_2), \tag{1}$$

with a decision outcome

$$a(\tau) \in \{r_1, r_2\}, \tag{2}$$

where $a(\tau)$ denotes the candidate selected by the decision-maker. This formulation is invariant to the ordering or presentation of the candidates.

To relate these decisions to structural properties of procedures, we introduce a family of structural axes indexed by $k \in \mathcal{K}$. All axes are defined over the same underlying set of procedures $\mathcal{R}$, which serves as a shared procedural space for constructing controlled comparisons.

Each axis is defined by a feature mapping

$$\psi_k : \mathcal{R} \to \mathcal{Z}_k \tag{3}$$

that maps each procedure $r \in \mathcal{R}$ to an axis-specific representation $\psi_k(r) \in \mathcal{Z}_k$, together with a comparison function

$$\Delta_k : \mathcal{Z}_k \times \mathcal{Z}_k \to \mathbb{R}. \tag{4}$$

The framework does not assume that Δ_k satisfies metric properties such as symmetry or the triangle inequality. Depending on the axis, Δ_k may behave like a distance (e.g., semantic dissimilarity) or as a signed comparison value (e.g., differences in ingredient rarity).

For a trial $\tau = (r_q, r_1, r_2)$, we compute the reference-relative values

$$\delta_{k,1} = \Delta_k(\psi_k(r_q), \psi_k(r_1)), \qquad \delta_{k,2} = \Delta_k(\psi_k(r_q), \psi_k(r_2)), \tag{5}$$

and define the signed contrast

$$\tilde{x}_k(\tau) = \delta_{k,1} - \delta_{k,2}, \tag{6}$$

with magnitude

$$x_k(\tau) = |\tilde{x}_k(\tau)|. \tag{7}$$

The sign of $\tilde{x}_k(\tau)$ determines the candidate indicated by the structural comparison under axis k. Let

$$p_k(\tau) = \begin{cases} r_1 & \text{if } \tilde{x}_k(\tau) < 0, \\ r_2 & \text{if } \tilde{x}_k(\tau) > 0, \end{cases} \tag{8}$$

denote this candidate, where ties (i.e., $\tilde{x}_k(\tau) = 0$) are excluded from analysis.

Each axis therefore induces a direction of variation between candidates relative to the reference procedure. The framework does not assume that this direction corresponds to correctness or optimality; rather, it provides a consistent coordinate system for analyzing how decisions vary with contrast along each axis.

Because all axes are defined over the same set of procedures, comparisons across axes can be performed consistently within the same set of trials.

To connect decisions with structural contrasts, we define an alignment indicator

$$y^{(k)}(\tau) = \mathbf{1}\big[a(\tau) = p_k(\tau)\big]. \tag{9}$$

This indicator converts each trial into a binary outcome that records whether the reuse judgment aligns with the directional comparison induced by the axis. Aggregating this indicator over trials enables us to characterize how consistently decisions align with the axis as a function of contrast magnitude $x_k(\tau)$.

The variable $y^{(k)}(\tau)$ serves as the response variable used to analyze reuse judgments along axis k.

2.3 Output Characterization

For each reference procedure r_q and axis k, we define the response function

$$f^{(k)}_{r_q}(x) = \mathbb{P}\big(y^{(k)}(\tau) = 1 \mid x_k(\tau) \in \mathcal{B}(x),\, r_q\big), \tag{10}$$

where $\mathcal{B}(x)$ denotes a bin centered at x, used to discretize the contrast variable. The probability is taken over trials τ with reference procedure r_q and nonzero contrast (i.e., $x_k(\tau) > 0$).

The function $f^{(k)}_{r_q}(x)$ describes the probability that the decision-maker's reuse judgments align with the directional comparison induced by axis k as the structural contrast between candidates increases.

Rather than summarizing behavior using a single scalar, we treat the full response function as a primary object of measurement, capturing how decision behavior varies as a function of structural contrast along each axis.

This formulation enables systematic characterization of how decision behavior varies with contrast magnitude across axes.

3 Experimental Setup

This section describes how the measurement framework introduced in Section 2 is instantiated in our experiments. Following the pipeline illustrated in Figure 1, we describe each stage of the measurement process: starting from the recipe corpus, defining comparison axes, constructing candidate procedures, obtaining pairwise reuse judgments, and aggregating decisions into response curves. Implementation details of dataset construction, candidate sampling, and prompt design are provided in Appendix A.

The experiments examine how pairwise reuse judgments vary under controlled differences between candidate procedures along predefined axes, and how these response patterns change under contextual interventions.

3.1 Procedural Domain

We instantiate the framework in the domain of structured cooking recipes. The recipe corpus is derived from the Food.com dataset, which contains approximately 520,000 user-contributed recipes.

From this dataset we construct a working set of approximately 20,000 recipes through cluster-based sampling to maintain diversity of procedural patterns.

Each recipe corresponds to a procedural artifact $r \in \mathcal{R}$, consisting of a title, an ingredient list, and a sequence of procedural steps.

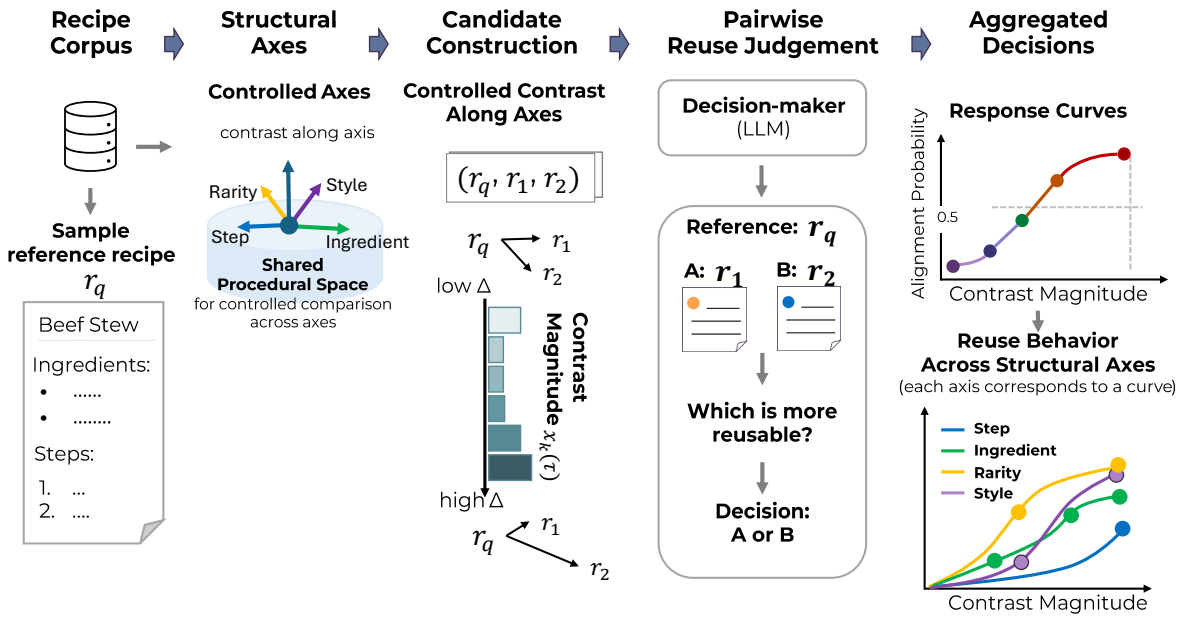

Figure 1: Overview of the measurement pipeline. From a corpus of procedural texts, a reference procedure r_q is sampled and candidate procedures are constructed under controlled contrasts along multiple structural axes defined within a shared procedural space. Pairs of candidate procedures (r_1, r_2) are compared with respect to the reference using an automated decision-maker. Aggregating these pairwise decisions across bins of contrast magnitude $x_k(\tau)$ yields response curves that characterize how the probability that the selected candidate matches the ordering defined by each axis varies as a function of contrast magnitude, enabling comparison of reuse behavior across axes.

In the experimental setup, one recipe is selected as the reference procedure r_q. From the sampled corpus, 300 recipes are randomly selected as reference procedures in the main experiments. The remaining recipes serve as candidate procedures.

3.2 Comparison Axes

We define a set of comparison axes that capture distinct aspects of variation between recipes (e.g., step semantic similarity, ingredient similarity, ingredient rarity, and recipe style). Each axis induces a reference-relative comparison between a reference recipe and candidate recipes, which defines the contrast magnitude used in the analysis, as formalized in Section 2.

These axes specify how contrast magnitude is defined and allow analysis of how reuse judgments vary with respect to contrast magnitude along each axis.

The axes are defined over existing recipes to ensure that the resulting variations correspond to naturally occurring procedural differences.

Implementation details for each axis are provided in Appendix C.

3.3 Candidate Construction

Given a reference recipe r_q, candidate recipes are compared with the reference along each axis to obtain axis-induced comparison values. Pairwise comparison trials are then constructed by selecting two candidate recipes r_1 and r_2, forming a trial $\tau = (r_q, r_1, r_2)$.

To enable controlled comparison across axes by holding the set of trials fixed, we adopt a *shared-pairs design* in which the same candidate pairs are reused across all conditions.

In the main experiments, candidate pairs are generated using the step semantic axis. The choice of sampling axis serves only to define the distribution of candidate pairs and is held fixed across axes; it is not treated as privileged in the subsequent analysis.

Alternative pair construction strategies were also considered. These induce different trial distributions and can lead to different response patterns, reflecting the dependence of response curves on the underlying set of candidate pairs.

For this reason, the shared-pairs design is adopted in the main analysis to enable controlled comparison across axes.

3.4 Reuse Judgment via a Controlled Decision-Maker

For each trial $\tau = (r_q, r_1, r_2)$, a decision-maker selects which candidate is preferred for reuse with respect to the reference recipe.

In the main experiments, this decision-maker is implemented using a large language model (LLM), *Llama-3.1-8B-Instruct* (Grattafiori et al., 2024), under fixed prompting conditions.

The decision-maker is provided with the procedural steps of the reference and candidate recipes. In the baseline condition, only step sequences are shown, without additional metadata such as titles or ingredient lists, and this condition is used for comparison with contextual interventions.

The role of the decision-maker is to provide a consistent and replicable signal for comparative evaluation under controlled conditions. The analysis focuses on how decisions vary with respect to contrast magnitude $x_k(\tau)$, rather than on the absolute correctness of individual decisions.

Decisions are mapped to alignment indicators following the formulation in Section 2, where alignment means that the selected candidate is consistent with the comparison induced by the corresponding axis (i.e., the candidate with the higher axis value is selected).

To reduce positional bias, the order of candidate recipes is randomly swapped across trials with equal probability (0.5), ensuring that an alignment probability of 0.5 corresponds to chance-level agreement with the axis-induced comparison.

3.5 Contextual Interventions

To examine how reuse judgments change under different task contexts, we introduce contextual interventions by modifying the prompts presented to the decision-maker.

These interventions are evaluated relative to the baseline condition described above and alter the information or instructions provided to the decision-maker while keeping the underlying candidate pairs fixed.

We consider the following conditions:

- **Goal specification** (including recipe titles to make task goals explicit)
- **Novice reader** (framing the task for a less experienced user)
- **Time-limited context** (introducing constraints on preparation time)
- **Limited equipment** (restricting available tools or resources)

These interventions introduce different constraints under which reuse decisions are made. Differences in response curves across conditions reflect systematic changes in comparative judgment under these contextual modifications.

Implementation details of prompt modifications are provided in Appendix B.

Table 1: Seed-wise stability of response curves measured by cosine similarity. For each reference procedure and axis, response curves obtained under different pairing seeds are compared by treating each curve as a vector over contrast bins. Cosine similarities are computed between all pairs of curves obtained from different seeds. Values report mean ± standard deviation of cosine similarity aggregated over all reference–axis combinations. The number of reference procedures may vary slightly across axes, as some cases are excluded when insufficient variation in contrast values is available under a given axis, resulting in insufficient data to populate contrast bins.

Axis	Cosine similarity	#References
Step semantic	0.881 ± 0.095	300
Ingredient similarity	0.828 ± 0.133	290
Ingredient rarity	0.871 ± 0.085	300
Recipe style	0.868 ± 0.090	300

3.6 Aggregation and Response Curves

For each trial and axis, the contrast magnitude $x_k(\tau)$ between the two candidate procedures relative to the reference is computed as defined in Section 2. Trials are grouped into bins based on $x_k(\tau)$.

Within each bin, the alignment probability is estimated as the empirical mean of $y^{(k)}(\tau)$ over trials in the bin, corresponding to the fraction of decisions that select the candidate consistent with the axis-induced comparison.

This yields response curves that characterize how alignment probability varies as a function of contrast magnitude.

4 Empirical Results

We analyze reuse judgments using response curves constructed from aggregated pairwise decisions, as defined in Section 2. Each response curve represents the alignment probability as a function of contrast magnitude for a given axis.

Unless otherwise stated, all results are reported under the steps-only baseline condition described in Section 3, in which the decision-maker is provided only with procedural steps. Results under contextual interventions are then compared against this baseline.

We address the following questions:

- (R1) Stability: Are response curves consistent under different random pairings?
- (R2) Axis dependence: How do response curves vary with contrast magnitude across axes?
- (R3) Distribution: How are response curves distributed in curve space?
- (R4) Intervention: How do contextual interventions shift response curves?

4.1 R1: Stability

We first assess whether response curves are stable with respect to random pairing seeds.

For each reference procedure and axis, we construct response curves under three independent pairing seeds using a shared-pairs design. Each response curve is represented as a vector over contrast bins, and stability is quantified by computing cosine similarity between all pairs of curves obtained from different seeds.

Figure 2 (left) shows example curves for a single reference procedure. Individual bins exhibit noticeable variability across seeds, indicating that response curves are locally sensitive to pairing differences.

To quantify this variability, Figure 2 (right) shows the distribution of pairwise cosine similarities between curves constructed from different pairing seeds. For each reference procedure and axis, similarities are

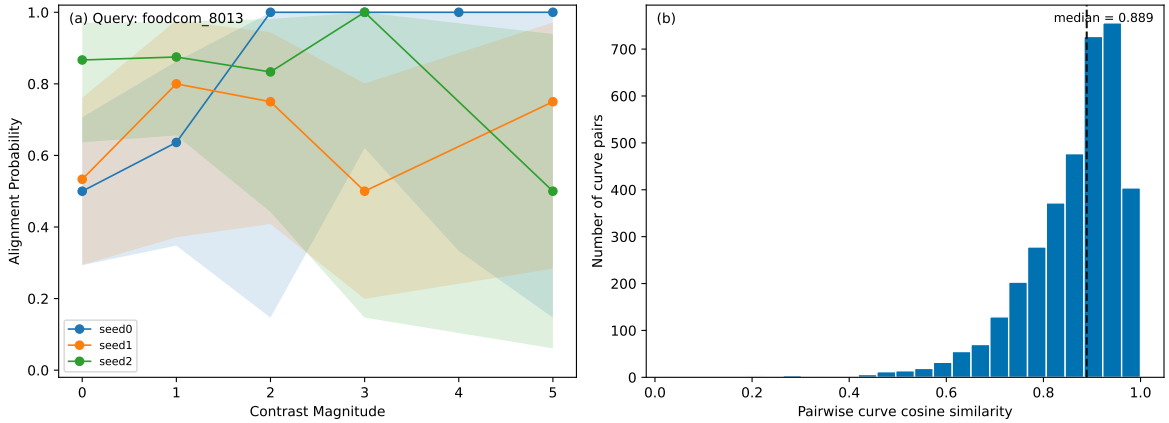

Figure 2: Stability of response curves across pairing seeds. For each reference procedure and axis, response curves are constructed under three independent pairing seeds using a shared-pairs design. Left: example response curves for a single reference procedure under three seeds, where each curve is represented over contrast bins. Individual bins exhibit noticeable variability across seeds, resulting in local fluctuations across curves. Right: distribution of pairwise cosine similarities between response curves constructed with different pairing seeds. For each reference procedure and axis, similarities are computed between all pairs of curves obtained under different seeds and aggregated across all reference–axis combinations. The vertical axis shows the number of such curve pairs.

computed between all pairs of curves obtained under different seeds and aggregated across all reference–axis combinations. The distribution is skewed toward high similarity values, with a median of 0.889.

Table 1 summarizes these similarities as mean ± standard deviation over the same set of comparisons.

Despite local variability, cosine similarities remain relatively high (approximately 0.8–0.9) across axes. This indicates that response curves are globally consistent across seeds, suggesting that the observed response patterns reflect stable properties of the comparison structure rather than artifacts of a particular random pairing configuration.

4.2 R2: Axis Dependence

We next examine how response curves vary with contrast magnitude across different axes.

Figure 3 shows mean response curves for each axis, obtained by grouping trials by contrast magnitude and averaging alignment probabilities across reference procedures within each bin. Because all axes are evaluated over a shared set of candidate pairs under the shared-pairs design, differences between curves reflect axis-dependent variation in reuse judgments under a shared sampling condition.

Because candidate pairs are constructed using the step semantic axis, the resulting trial distribution is not uniform across axes and may favor variation along that axis. As a result, observed differences between axes should be interpreted as reflecting axis-dependent responses under this shared sampling condition, rather than as intrinsic properties of the axes alone.

The mean curves exhibit axis-dependent differences in their overall trends. For example, the ingredient rarity axis shows a weak decreasing pattern with increasing contrast magnitude, whereas the step semantic axis exhibits a more consistent increasing pattern. Ingredient similarity shows a weakly increasing but non-monotonic pattern, while recipe style does not exhibit a clear or consistent trend across contrast bins.

At the same time, substantial variability is observed across reference procedures, as indicated by the spread of individual curves around the mean.

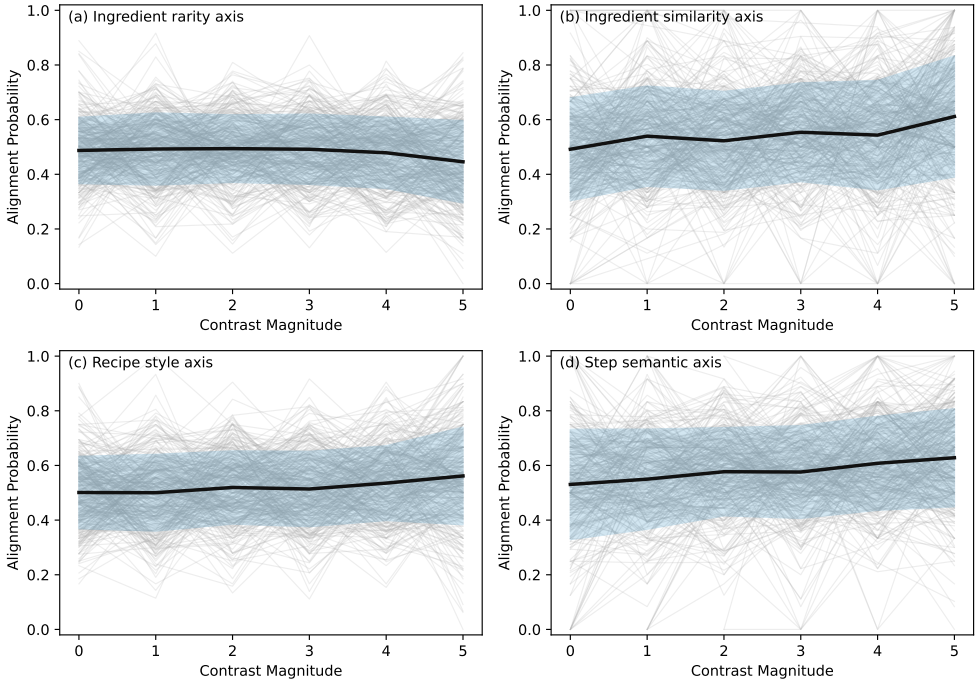

Figure 3: Mean response curves across structural axes under the shared-pairs design. For each axis, trials are grouped by contrast magnitude, and alignment probabilities are averaged across reference procedures within each bin. The black line shows the mean response curve, and the shaded region indicates variability across reference procedures. Because all axes are evaluated over a shared set of candidate pairs, differences between curves reflect axis-dependent variation in reuse judgments under a shared sampling condition.

Figure 4 shows an example for a single reference procedure, illustrating that individual response curves are often non-monotonic and that axis-dependent patterns are not consistently observed at the level of individual instances.

Taken together, these results suggest that reuse judgments vary with contrast magnitude in an axis-dependent manner at the population level, despite substantial variability at the level of individual reference procedures.

Axis-specific diagnostic experiments provide a complementary view under axis-specific sampling (Appendix A.4). The resulting response curves (Appendix D) exhibit related but distribution-dependent patterns, with both the magnitude and shape of the curves varying according to the underlying trial distribution.

4.3 R3: Distribution in Curve Space

To further characterize variation in reuse judgments beyond axis-wise comparisons, we examine how response curves are distributed in a low-dimensional space.

Each response curve is represented as a vector over contrast bins. Pairwise similarities between curves are then used to construct a low-dimensional embedding (see Appendix F for details). In this space, each point corresponds to a reference–axis pair.

Figure 5 shows the resulting embedding. Response curves occupy a largely overlapping region with gradual transitions, without clear separation into a small number of distinct clusters, although weak axis-dependent shifts in the distribution can be observed.

This suggests that variation in reuse judgments is better described as gradual variation in response curves, rather than as a small set of discrete response types. However, we emphasize that this interpretation is qualitative, and we do not claim the absence of clustering.

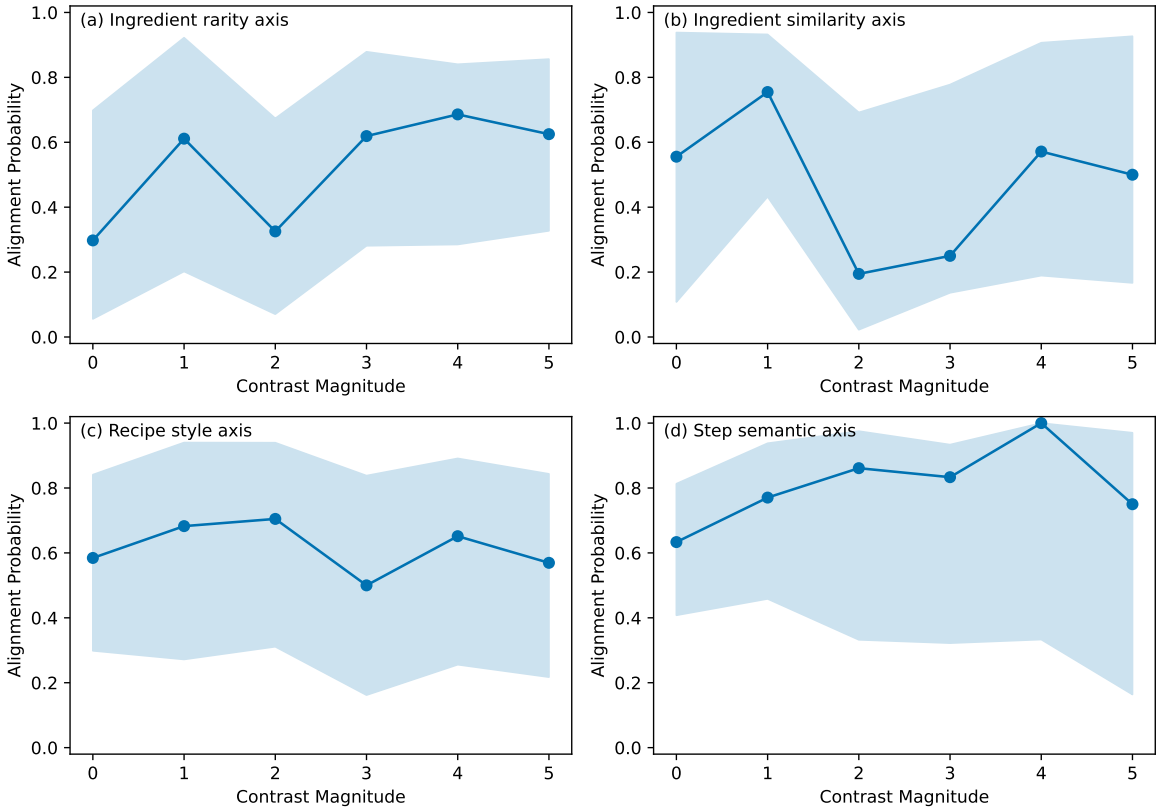

Figure 4: Example response curves for a single reference procedure across axes. Each curve shows alignment probability as a function of contrast magnitude for a given axis. Individual curves are often non-monotonic and exhibit substantial fluctuations, illustrating variability at the level of individual reference procedures.

Table 2: Displacement between baseline and intervention conditions. For each reference procedure, response curves are constructed under both conditions using the same set of candidate pairs under the step semantic axis. Each response curve is represented as a vector over contrast bins, and displacement is defined as $1-$ cosine similarity between the baseline and intervention curves. Values report mean $\pm$ standard deviation of displacement aggregated over reference procedures.

Condition	Displacement	#References
Goal specification	0.504 ± 0.166	300
Novice profile	0.520 ± 0.164	300
Limited equipment	0.529 ± 0.162	300
Time limited	0.535 ± 0.180	300

We also note that this analysis depends on the choice of embedding method and similarity metric, and should be interpreted as an exploratory visualization of relationships between response curves rather than a definitive geometric characterization.

Qualitatively similar structure is observed when using an independent judge model (Appendix E.1), suggesting that similar geometric structure can be observed across different decision-makers.

4.4 R4: Intervention-Induced Shifts

Finally, we examine how contextual interventions affect response curves.

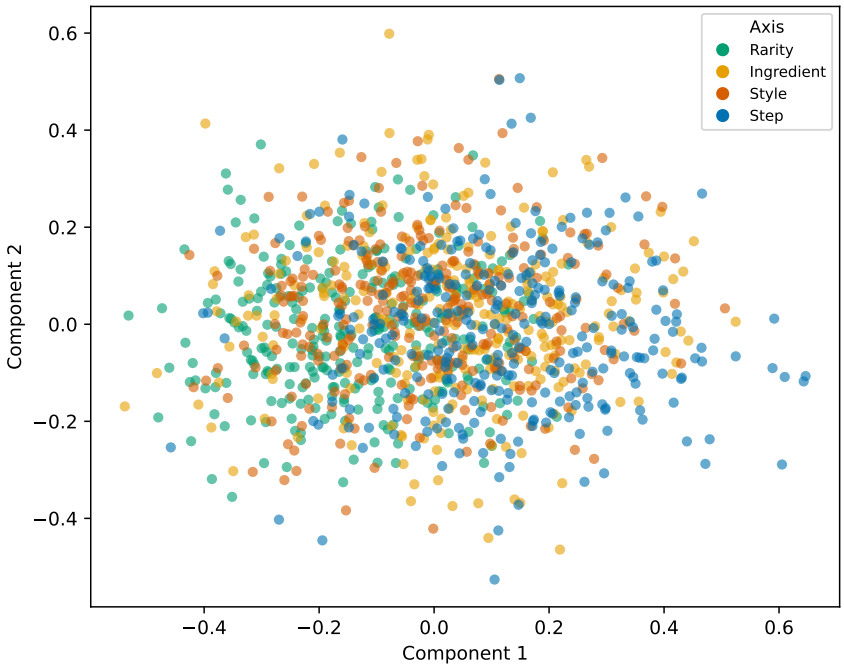

Figure 5: Low-dimensional embedding of response curves. Each point corresponds to a reference–axis pair, where response curves are represented as vectors over contrast bins and embedded based on pairwise similarity. Points are colored by axis. The embedding shows substantial overlap between axes, with no clear separation into a small number of distinct clusters, although weak axis-dependent shifts in the distribution can be observed.

For each reference procedure, we construct response curves under both the baseline condition and each intervention condition using the same set of candidate pairs under the step semantic axis. Each response curve is represented as a vector whose entries correspond to alignment probabilities over contrast bins.

To quantify changes in response curves, we define displacement as the distance between the baseline and intervention response curves, computed as $1-$ cosine similarity between these vectors. This provides a measure of change complementary to the similarity-based analysis used in R1.

For visualization, the embedding in Figure 6 is constructed using response curves under the step semantic axis. Each point corresponds to a response curve for a given reference procedure, and arrows indicate the shift in the centroid (mean position) of the distribution under each intervention condition.

Across conditions, the distributions of response curves remain substantially overlapping with the baseline, but exhibit consistent shifts in their centroid positions.

Note that the displacement values reported in Table 2 are computed in the original response-curve space, rather than in the low-dimensional embedding used for visualization.

Table 2 summarizes the magnitude of these changes as mean $\pm$ standard deviation of displacement across reference procedures under the step semantic axis. Displacement values are similar across interventions and consistently greater than zero, indicating non-trivial changes relative to the baseline condition.

These results suggest that contextual information is associated with systematic shifts in response patterns, while preserving substantial variability across reference procedures.

Summary Response curves provide a stable measurement of reuse behavior (R1), show axis-dependent differences in their dependence on contrast magnitude (R2), exhibit a largely overlapping distribution with

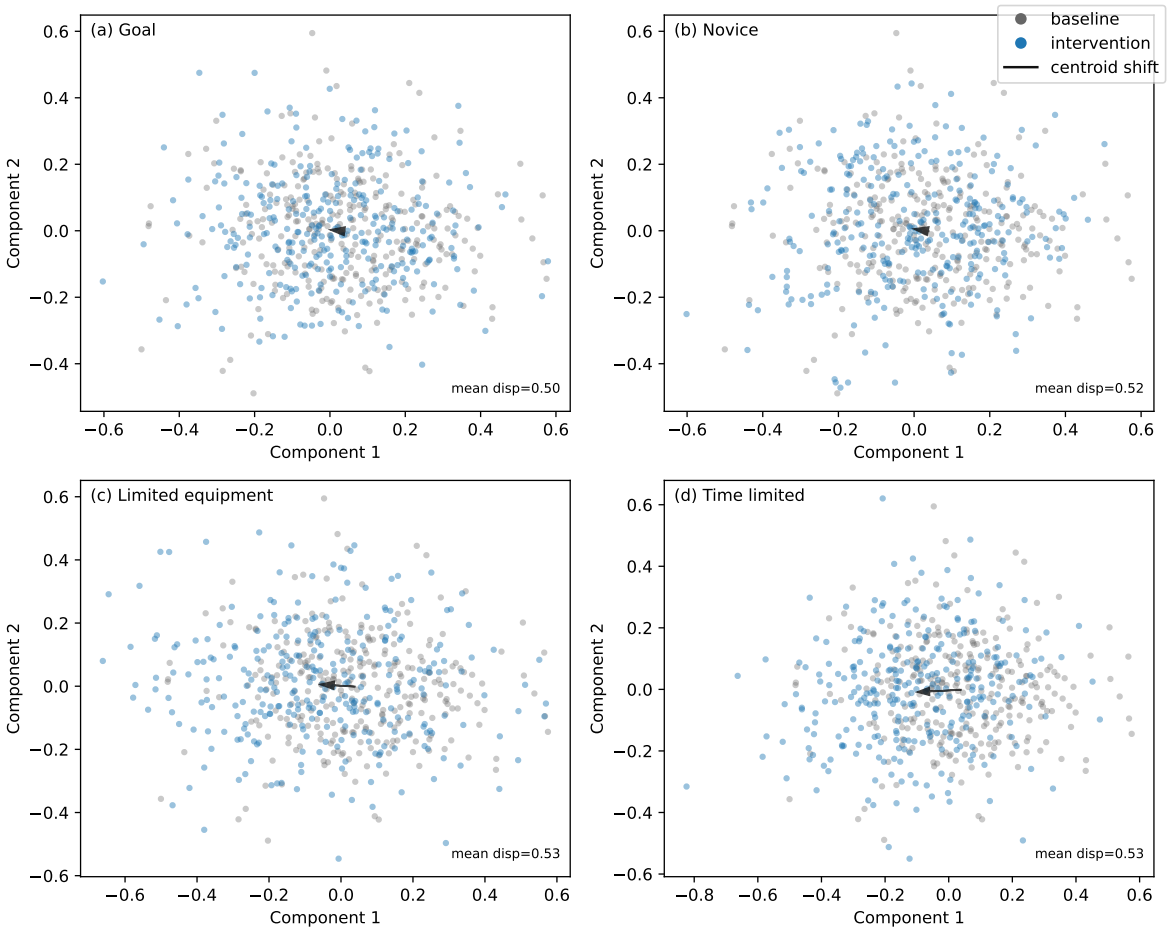

Figure 6: Shifts in response curves under contextual interventions. Each point corresponds to a response curve for a given reference procedure under the step semantic axis, embedded in a low-dimensional space constructed as in Figure 5. Gray points indicate baseline conditions, and blue points indicate intervention conditions. Arrows show the shift in the centroid (mean position) of the distribution from baseline to intervention. Each panel corresponds to a different intervention condition. The visualization illustrates that response curves shift in a consistent direction under each intervention, while retaining substantial overlap with the baseline distribution.

gradual variation in curve space (R3), and undergo consistent shifts under contextual interventions while retaining substantial variability (R4).

5 Interpreting Response-Curve Structure

The empirical results in Section 4 show that reuse judgments exhibit structured variation across reference procedures, structural axes, and contextual conditions. These patterns are captured by response curves that summarize alignment probability as a function of contrast magnitude.

In this section, we interpret what aspects of reuse behavior are reflected in this response-curve structure and clarify what can be inferred from these observations.

5.1 Reuse Judgment as a Response Function

The stability analysis (R1) shows that response curves are globally consistent across pairing seeds, despite local fluctuations at the level of individual bins. This suggests that response curves capture stable properties of the comparison structure rather than artifacts of a particular sampling configuration.

At the same time, the axis-dependent analysis (R2) shows that mean response curves differ across axes, while individual reference procedures exhibit substantial variability and often non-monotonic patterns.

Taken together, these results suggest that reuse judgments cannot be fully captured by a single global sensitivity parameter, and are more appropriately characterized as a mapping from contrast magnitude to alignment probability. The response curve provides a direct representation of this mapping.

5.2 Continuous Variation Rather than Discrete Types

The embedding analysis (R3) shows that response curves occupy a largely overlapping region in curve space, without clear separation into a small number of distinct clusters.

This suggests that variation in reuse judgments is better described as gradual variation in response curves, rather than as a small set of discrete response types, at least under the representation considered in this analysis.

However, this interpretation depends on the choice of representation and similarity metric, and should be viewed as an exploratory characterization rather than a definitive geometric claim.

5.3 Interpretation Gap as Response-Curve Shift

The intervention analysis (R4) shows that response curves constructed under different contextual conditions exhibit systematic shifts while remaining substantially overlapping with the baseline distribution.

Because these curves are constructed using identical candidate pairs, the observed differences are unlikely to be attributable to changes in the underlying comparisons. Instead, they are consistent with differences in how contrast magnitude is evaluated under different contexts.

Under this perspective, an interpretation gap can be operationalized as a shift in response-curve structure between conditions, measured as changes in alignment probability over contrast magnitude, providing a measurable characterization of how reuse judgments vary under context.

5.4 Implications for Evaluating Reuse Behavior

Traditional evaluations of procedural capability often focus on execution outcomes. The present framework instead isolates comparative reuse judgments under controlled structural variation.

The response-curve representation provides a way to evaluate how reuse judgments change under controlled conditions. In particular, shifts in response-curve structure can be used to quantify how contextual information or system interventions affect perceived applicability and sensitivity to structural variation.

By aggregating pairwise decisions across reference procedures, response curves provide a population-level view of how reuse judgments vary with structural contrast, enabling comparative analysis of different reuse-support mechanisms without requiring a predefined notion of optimal reuse behavior.

6 Relation to Prior Work

Our work relates to research on pairwise comparative evaluation, preference modeling, transferability estimation, and reuse in LLM-based systems. The key distinction of our approach is to treat pairwise judgments as a measurement process and to analyze how responses vary under controlled structural contrasts, rather than to aggregate comparisons into scalar scores or predictive models.

6.1 Pairwise Judgments as Measurement

Pairwise comparisons are widely used to evaluate model outputs and align systems with human preferences (Li et al., 2024; Zheng et al., 2023; Chiang et al., 2024). Classical models such as Thurstone, Bradley–Terry, and Elo aggregate such comparisons into latent scalar scores or rankings (Thurstone, 1927; Bradley & Terry, 1952; Elo, 1978).

In contrast, we do not aim to recover a latent score. Instead, we treat pairwise decisions as measurements and analyze how alignment probabilities vary as a function of contrast magnitude. Under this perspective, pairwise comparisons are used to estimate a contrast–response relationship, rather than to fit a ranking model. This perspective is related to measurement approaches in psychophysics, where response probabilities are analyzed as a function of controlled stimulus differences (Green & Swets, 1966; Wichmann & Hill, 2001).

6.2 From Preference Modeling to Response Characterization

Recent work on preference learning and reward modeling uses human preference data, often collected as rankings or comparisons between outputs, to infer latent reward functions or preference models, for example in reinforcement learning from human feedback (RLHF) settings (Christiano et al., 2017; Ouyang et al., 2022). In these approaches, preference data are typically aggregated into a scalar reward signal used for optimization.

In contrast, we do not aggregate comparisons into a latent function or use them as supervision. Instead, we treat pairwise decisions as measurements and analyze how responses vary as a function of controlled contrast magnitude. This enables direct characterization of response structure, which is not accessible through scalar reward modeling.

6.3 Task Similarity and Transferability

Prior work on transferability estimation studies how knowledge learned for one task generalizes to another, using representations and metrics such as Task2Vec, LEEP, and LogME (Achille et al., 2019; Nguyen et al., 2020; You et al., 2021).

These approaches aim to predict downstream transfer performance. In contrast, we do not predict transfer outcomes. Instead, we measure how reuse judgments vary under controlled structural contrasts. The axes in our framework are not predictive features but serve as experimental coordinates for probing response behavior.

6.4 Reuse in LLM-Based Systems

Recent work on LLM-based agents explores mechanisms for storing and reusing past experience, intermediate reasoning, or retrieved context (Wang et al., 2023; Shinn et al., 2023; Lewis et al., 2020).

These approaches aim to improve reuse capability. Our work is complementary: we provide a framework for measuring how reuse judgments respond to controlled variation, offering a diagnostic perspective for analyzing such systems.

7 Limitations and Scope

Our framework is designed to make reuse judgments observable under controlled variation. However, several limitations remain.

7.1 LLM-Based Evaluation and Scope

We use an LLM as the decision-maker. We do not assume that these judgments correspond to human decisions; rather, they provide a relatively consistent signal that can be systematically measured. Extending the framework to human evaluators remains an important direction.

The framework does not define an optimal reuse strategy. Instead, it measures how alignment probabilities vary under controlled contrast magnitude. Observed shifts should therefore be interpreted as changes in response behavior rather than improvements in reuse quality.

7.2 Axis Design and Dependencies

The structural axes are heuristic and not intended as validated predictors of transfer success. Their role is to provide controlled coordinates for measuring how alignment probabilities vary with contrast magnitude. Alternative axis definitions may reveal different response structures.

Candidate procedures are constructed using a shared-pairs design in which sampling is performed along a specific axis (step semantic similarity). As a result, the distribution of candidate pairs is not uniform across axes, and observed response patterns along other axes are conditioned on this sampling procedure.

This introduces a dependency between axis definition and data distribution. In particular, axis-dependent differences in response curves should be interpreted as measurements under the given sampling condition, rather than as intrinsic properties of the axes.

The structural axes are defined over a shared procedural representation and may exhibit correlations with one another. For example, procedures that are similar in step structure may also share ingredients or stylistic properties. As a result, variations along one axis may partially co-vary with others, making it difficult to attribute response changes to a single axis in isolation. Disentangling such dependencies would require jointly controlled or orthogonalized axis constructions.

7.3 Estimation and Discretization Effects

Response curves are constructed by grouping trials into bins based on contrast magnitude and estimating alignment probabilities within each bin. As a result, the shape of the response curves depends on the choice of binning scheme, including bin boundaries and the number of bins.

While our analysis uses a fixed binning strategy for consistency, alternative discretizations may yield different curve shapes, particularly in regions with sparse data. Exploring binning-free or adaptive estimation methods remains an important direction for future work.

7.4 Domain and Coverage Limitations

We focus on cooking recipes as a structured procedural domain. While the framework may generalize to other domains, such generalization is not evaluated here.

We also consider a limited set of contextual interventions. Future work may explore a broader space of constraints and how they reshape response curves.

8 Conclusion

We introduced a measurement framework for analyzing reuse judgments through response curves defined over contrast magnitude.

By using controlled pairwise comparisons, the framework treats reuse judgments as observable outcomes and characterizes how alignment probabilities vary with structural contrast along different axes.

Empirically, we find that response curves are stable under sampling variation, exhibit axis-dependent sensitivity, form structured but overlapping patterns in curve space, and shift systematically under contextual interventions.

These results indicate that reuse judgments cannot be fully captured by a single similarity score, and are more informatively described as structured variation in alignment probability under controlled contrast.

More broadly, this work proposes a measurement-oriented perspective on procedural reuse, complementing prediction-oriented approaches by enabling direct observation of how reuse judgments vary under controlled conditions.

Because the framework evaluates multiple structural axes over a shared set of trials, it enables controlled comparison of reuse behavior across axes while holding candidate sampling fixed.

We hope this perspective provides a foundation for studying procedural reasoning, designing reuse-aware systems, and developing evaluation methodologies for systems that generate or adapt procedures.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task Embedding for Meta-Learning, February 2019.

Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 0006-3444. doi: 10.2307/2334029.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4302–4310, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub, New York, 1978. ISBN 978-0-668-04721-0.

Food.com. Food.com recipes and reviews dataset. `https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews`. Accessed: 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. The Llama 3 Herd of Models, November 2024. URL `http://arxiv.org/abs/2407.21783`. arXiv:2407.21783 [cs].

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Signal Detection Theory and Psychophysics. John Wiley, Oxford, England, 1966.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. Mise en Place: Unsupervised Interpretation of Instructional Recipes. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 982–992, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1114.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 97–106, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2016.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods, December 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, February 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A New Measure to Evaluate Transferability of Learned Representations, August 2020.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019.

Yu-Zhe Shi, Haofei Hou, Zhangqian Bi, Fanxu Meng, Xiang Wei, Lecheng Ruan, and Qining Wang. AutoDSL: Automated domain-specific language design for structural representation of procedures with constraints. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12177–12214, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.659.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, October 2023.

Larisa N. Soldatova, Wayne Aubrey, Ross D. King, and Amanda Clare. The EXACT description of biomedical protocols. *Bioinformatics*, 24(13):i295–i303, July 2008. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btn156.

Larisa N Soldatova, Daniel Nadis, Ross D King, Piyali S Basu, Emma Haddi, Véronique Baumlé, Nigel J Saunders, Wolfgang Marwan, and Brian B Rudkin. EXACT2: The semantics of biomedical protocols. *BMC Bioinformatics*, 15(Suppl 14):S5, November 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-S14-S5.

L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927. ISSN 1939-1471. doi: 10.1037/h0070288.

Wil M. P. Van Der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-19344-6 978-3-642-19345-3. doi: 10.1007/978-3-642-19345-3.

Wil M. P. van der Aalst and Kees van Hee. *Workflow Management: Models, Methods, and Systems*. 2002. ISBN 978-0-262-28541-4. doi: 10.7551/mitpress/7301.001.0001.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models, October 2023.

Felix A. Wichmann and N. Jeremy Hill. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313, November 2001. ISSN 1532-5962. doi: 10.3758/BF03194544.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. https://arxiv.org/abs/2102.11005v3, February 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023.

A Experimental Setup

A.1 Dataset and Reference Sampling

The experiments use a recipe corpus derived from the Food.com dataset Food.com, which contains approximately 520,000 user-contributed recipes.

From this dataset, we construct a working set of approximately 20,000 recipes using cluster-based sampling to ensure diversity of procedural patterns.

From the sampled corpus, 300 recipes are randomly selected as reference procedures ($q_size = 300$) for the main experiments.

A.2 Candidate Pair Construction

Candidate pairs are constructed using a shared-pairs design so that the same set of trials is used across all structural axes.

For each reference procedure r_q, candidate recipes are drawn from the sampled corpus. In the main experiments, candidates are ranked according to their step-semantic contrast magnitude with respect to r_q. This axis is used only for constructing candidate pairs and does not affect the definition of other structural axes used in the analysis.

Candidates are grouped into bins according to their contrast magnitude values along the step-semantic axis, and candidate pairs are randomly sampled within each bin to form trials $\tau = (r_q, r_1, r_2)$.

In the shared-pairs experiments, approximately 40 candidate pairs are sampled per reference procedure ($pairs_per_reference = 40$). These same pairs are reused across all structural axes and experimental conditions.

This design ensures that all axes are evaluated on the same underlying set of trials. As a result, differences in response curves across axes reflect differences in reuse judgments, rather than differences in the candidate sampling process.

For diagnostic experiments (Appendix A.4), candidate pairs may instead be constructed separately for each axis by ranking candidates according to the corresponding axis. These axis-specific constructions are used to examine response patterns within each axis and to verify that observed trends are not artifacts of the shared-pairs design. They are not used for the main analysis.

A.3 Contrast Binning

Contrast-magnitude bins are defined using fixed band boundaries.

Step semantic, ingredient similarity, recipe style

$$\{0, 2, 4, 6, 8, 10\}$$

Ingredient rarity

$$\{-10, -6, -2, 0, 2, 6, 10\}$$

Bins with insufficient samples are excluded from analysis to ensure reliable estimation of response probabilities.

A.4 Per-Axis Diagnostic Experiments

To complement the shared-pairs design, we conduct axis-specific experiments in which candidate pairs are constructed separately for each axis.

In these experiments, candidates are ranked according to the corresponding axis, and approximately 50 candidate pairs are sampled per reference procedure ($pairs_per_reference = 50$, $q_size = 300$).

These experiments allow examination of response patterns within each axis independently, providing a complementary view of how reuse judgments vary with contrast magnitude.

Compared to the shared-pairs design, the resulting response curves exhibit differences in magnitude due to changes in the underlying distribution of candidate pairs. However, qualitative trends across axes (e.g., increasing or decreasing behavior with contrast magnitude) remain consistent.

Because candidate pairs differ across axes, comparisons across axes are not controlled for differences in the underlying trial set. For this reason, all main analyses involving cross-axis comparisons are based on the shared-pairs design.

A.5 LLM Judge Configuration

The primary experiments use the Llama-3.1-8B-Instruct model as the judge.

Inference settings are:

- temperature: 0.0
- top_p: 1.0
- max tokens: 64
- seed: 1

To reduce positional bias, the order of candidate procedures is randomly swapped with probability 0.5.

B Prompt Templates

Reuse judgments are elicited through pairwise comparison prompts presented to the LLM judge.

Each prompt consists of three components:

- an instruction describing the reuse task,
- a reference procedure (the "current recipe"),
- two candidate procedures labeled A and B.

In the experimental framework described in the main text, the choice between candidates corresponds to the decision variable $a(\tau) \in \{1, 2\}$.

B.1 Baseline Prompt

The baseline instruction used in the experiments is shown below.

> You will compare two candidate procedural texts (A and B).
>
> Below is the CURRENT RECIPE you are cooking (your own dish). You want to quickly borrow and adapt part of one candidate (a step, technique, or idea) to improve your CURRENT RECIPE.
>
> Which candidate feels more reusable for improving your CURRENT RECIPE?
>
> Output a one-line JSON object with keys "choice" and "confidence". "choice" must be either "A" or "B". "confidence" must be a number between 0.0 and 1.0.

Recipes are rendered as structured text consisting of procedural steps and, optionally, additional contextual information.

The final prompt presented to the judge has the following structure:

> [Instruction]
> — CURRENT RECIPE (your own dish you are trying to improve) — [Reference procedure text]
> — CANDIDATE A — [Candidate procedure A]
> — CANDIDATE B — [Candidate procedure B]

B.2 Prompt Modifications for Intervention Conditions

Contextual interventions described in Section 3 are implemented by modifying either the instruction or the information included in the rendered procedure text.

In particular, goal specification is introduced by adding a short goal description (derived from the recipe title) to the procedure text, while other interventions such as novice, time-limited, and limited-equipment conditions are implemented by augmenting the instruction with additional context describing the assumed scenario.

Example instruction modifications Interventions implemented through instruction changes are realized by prepending fixed templates that specify the assumed context. For example:

> **Novice:** You are a novice home cook. You prefer straightforward, low-risk adaptations.
>
> **Time-limited:** You have only 20 minutes to finish your current recipe. You must prioritize ideas that reduce total cooking time and avoid techniques requiring long preparation.

Other intervention conditions follow a similar pattern, using fixed instruction templates to express constraints or assumed capabilities.

All prompts preserve the same overall structure as the baseline, and the exact wording of each intervention template is kept fixed across all trials.

C Structural Axis Definitions

All structural axes instantiate the comparison function Δ_k defined in Section 2, and are computed over procedures drawn from the same corpus, ensuring consistency with the shared procedural space.

For all axes, $\Delta_k(r_i, r_j)$ is defined as a scalar contrast measure. Depending on the axis, it may represent either a symmetric distance or a signed difference, and is used consistently within each axis to construct contrast magnitude.

Step semantic axis Procedural steps are embedded using a SentenceTransformers model (`all-MiniLM-L6-v2`) (Reimers & Gurevych, 2019). Let $E(r)$ denote the set of step embeddings for procedure r. The comparison function is defined as:

$$\Delta_{\text{step}}(r_i, r_j) = \frac{1}{|E(r_i)|} \sum_{e \in E(r_i)} \min_{e' \in E(r_j)} \big(1 - \cos(e, e')\big).$$

This defines a directed semantic distance from r_i to r_j based on step-level nearest-neighbor matching. This computes, for each step in r_i, the distance to its nearest semantically similar step in r_j, and averages these distances across all steps.

Ingredient similarity axis Ingredient lists are converted into normalized tokens with stopword removal, and represented using a presence-based TF–IDF scheme (i.e., token occurrence without counting multiplicity). Let v_r denote the resulting TF–IDF vector for recipe r. The comparison function is defined as:

$$\Delta_{\text{ingredient}}(r_i, r_j) = 1 - \cos(v_{r_i}, v_{r_j}).$$

This defines a symmetric distance between ingredient sets.

Ingredient rarity axis Let $\tilde{I}(r)$ denote the set of deduplicated normalized ingredient items in recipe r. Ingredient rarity is defined using information content:

$$IC(i) = -\log p(i).$$

The rarity score of a recipe is defined as the mean information content over its ingredients:

$$R(r) = \frac{1}{|\tilde{I}(r)|} \sum_{i \in \tilde{I}(r)} IC(i).$$

The comparison function is defined as:

$$\Delta_{\text{rarity}}(q, c) = R(c) - R(q).$$

This defines a signed contrast reflecting relative ingredient rarity with respect to the reference (query) procedure.

Recipe style axis Each recipe is represented by a feature vector capturing coarse structural properties of procedural text, including counts and statistics over steps and ingredients, length-based features, and indicators such as numeric and temporal expressions. Let s_r denote the resulting feature vector for recipe r. The comparison function is defined as:

$$\Delta_{\text{style}}(r_i, r_j) = 1 - \cos(s_{r_i}, s_{r_j}).$$

This defines a symmetric distance over stylistic properties of procedural descriptions.

D Per-Axis Response Curves

To complement the shared-pairs analysis in R2, we visualize response curves obtained under axis-specific sampling.

In these experiments, candidate pairs are constructed independently for each axis (see Appendix A.4), resulting in axis-dependent trial distributions.

Figure 7 shows the resulting response curves. Each gray curve corresponds to a reference procedure, and the black curve indicates the mean.

Compared to the shared-pairs setting, the magnitude and shape of response curves vary due to differences in the underlying trial distributions. However, similar qualitative trends are observed within each axis.

These results provide a complementary view of response behavior, but are not used for cross-axis comparison due to the lack of control over the trial distribution.

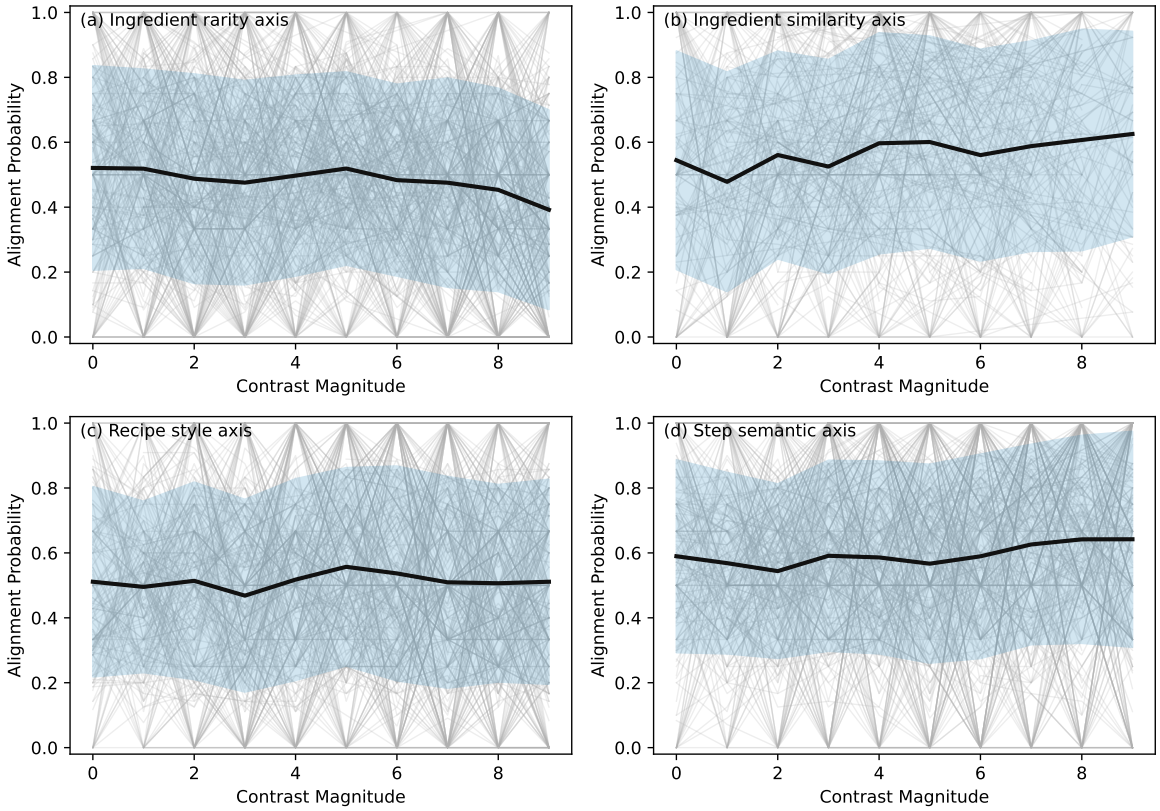

Figure 7: Per-axis response curves under axis-specific sampling. Each gray curve corresponds to a reference procedure, and the black curve indicates the mean across reference procedures.

E Robustness Analyses

E.1 Replication with an Additional Judge Model

To assess whether the observed response patterns depend on the choice of judge model, we replicate the experiments using Qwen2.5-7B-Instruct.

All experimental settings are kept identical to the main experiments, except for the choice of judge model, including the shared-pairs design, axis construction, contrast-magnitude binning, prompt templates, and the use of the step semantic axis for candidate pair construction.

Figure 8 shows baseline response curves obtained using Qwen. The overall structure closely resembles that observed with Llama, including relatively flat mean responses and substantial variation across reference procedures.

These results suggest that similar response structure is observed across models, and is not specific to a particular judge model.

E.2 Seed Stability

We evaluate the stability of response curves across different pairing seeds in the shared-pairs design.

Response curves are represented as discretized vectors over contrast-magnitude bins, and cosine similarity is computed between curves obtained from different seeds for the same reference procedure and axis.

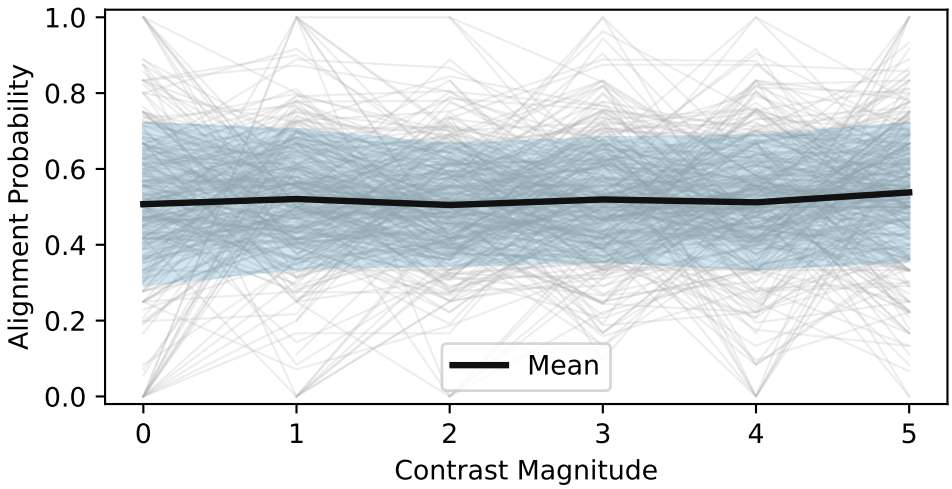

Figure 8: Baseline response curves using Qwen2.5-7B-Instruct as the judge model. Gray curves correspond to individual reference procedures, and the black curve indicates the mean response across reference procedures. The figure illustrates substantial variability at the level of individual reference procedures, with relatively weak trends at the population level.

Table 3: Seed-wise stability of response curves. Response curves are represented as discretized vectors over contrast-magnitude bins, and cosine similarity is computed between curves obtained from different pairing seeds for the same reference procedure and axis (mean ± standard deviation over reference–axis pairs).

Axis	Cosine similarity (mean ± std)	#References
Step semantic	0.853 ± 0.111	300
Ingredient similarity	0.825 ± 0.133	293
Ingredient rarity	0.874 ± 0.083	300
Recipe style	0.860 ± 0.095	300

Table 3 summarizes the results. Across all axes, response curves exhibit high similarity across seeds, indicating that the observed response patterns are stable with respect to candidate pair sampling.

E.3 Intervention Effects

We measure the displacement between baseline and intervention conditions in response-curve space to assess whether contextual modifications induce systematic changes in reuse judgments.

For each reference procedure and axis, response curves are constructed under both baseline and intervention conditions. Each curve is represented as a vector whose entries correspond to alignment probabilities over contrast-magnitude bins, and displacement is defined as $1-$ cosine similarity between the baseline and intervention vectors, consistent with the definition used in the main experiments.

Table 4 reports the magnitude of displacement across conditions (mean ± standard deviation over reference–axis pairs). Displacement values are similar across intervention types and consistently greater than zero, indicating non-trivial changes in response patterns under intervention.

The magnitude of displacement is comparable to that observed with Llama in the main experiments, suggesting that intervention effects are not strongly dependent on the choice of judge model.

Table 4: Displacement between baseline and intervention conditions. For each reference procedure and axis, displacement is defined as 1− cosine similarity between response-curve vectors (mean ± standard deviation over reference–axis pairs).

Condition	Displacement (mean ± std)	#References
Goal specification	0.509 ± 0.148	300
Novice profile	0.489 ± 0.160	300
Limited equipment	0.499 ± 0.165	300
Time limited	0.520 ± 0.167	300

F Embedding of Response Curves

Response curves are represented as vectors whose entries correspond to alignment probabilities over contrast-magnitude bins.

For visualization, we construct a single response-curve vector for each reference–axis pair by averaging across baseline runs. These vectors are stacked into a data matrix, and principal component analysis (PCA) is applied directly to this matrix.

The embedding is constructed using response curves from the baseline condition.

The first two principal components are used to obtain a two-dimensional projection for visualization. The embedding is used only for visualization purposes and does not affect any quantitative measurements reported in the main text.