Teacher Model-Guided Bayesian Decision Support for Structured Choice Prediction

Anonymous Authors

Abstract

Structured choice prediction is central to decision support systems in economics, transportation, and recommendation. Classical discrete choice models (DCM) provide interpretable utilities and well-defined likelihoods, but they often rely on handcrafted attributes that miss semantic context. Recent large language models (LLMs) can provide low-cost preference signals from textual descriptions, yet such signals are noisy, miscalibrated, and difficult to use directly as decision rules. In this paper, we propose TG-BC, a Teacher LLM-Guided Bayesian decision support framework for structured Choice prediction. The framework treats the LLM as an teacher semantic scorer and integrates its signal into a Bayesian discrete choice model through calibrated utility guidance and lightweight latent steering. Posterior inference then produces calibrated choice probabilities, credible intervals, and action-level uncertainty for downstream policy simulation. We instantiate the framework in an urban mobility mode-choice setting, where a decision maker compares interventions such as fare discounts, service-frequency improvements, parking surcharges, and informational nudges. The experimental protocol evaluates prediction, calibration, decision quality, interpretability, and robustness against DCM-only, LLM-only, naive-fusion, and deterministic-steering baselines.

1 Introduction

Structured choice prediction underpins decision support in transportation, economics, and recommendation (McFadden 1974; Ben-Akiva and Lerman 1985; Train 2009). Discrete choice models (DCMs) are the standard instrument in these settings not because they predict best, but because they deliver three things at once: an interpretable utility for every alternative, a well-defined likelihood over the choice set, and a simulator for interventions that were never observed (Danaf et al. 2019). An analyst who asks how travel mode shares move when fares fall or parking prices rise needs all three: a number, an interval around it, and a reason that can be audited by someone who did not fit the model. The price of that guarantee is that everything the model may consider must first be written into a hand-specified attribute table.

Language models make that table look narrow, and three lines of work have tried to widen it. The first keeps the choice model and enriches the context it is allowed to see: context effects are learned directly from observed choices (Seshadri, Peysakhovich, and Ugander 2019; Tomlinson and Benson 2021), or contextual variables are mapped into occasion-specific shifts of the preference parameters (Łukawska, Jensen, and Rodrigues 2025a). The second replaces the choice model, scoring alternatives directly from a textual description of the decision case (Nishida, Ishigaki, and Onishi 2025a; Mo et al. 2026), or asking the language model to propose the utility specification itself (Sfeir et al. 2025). The third keeps both and mixes them, combining the probabilities of a fitted utility model with probabilities read off the language model under a single weight tuned on validation data. Different as they are, all three follow the same sequence: decide once how much the language model counts, then apply that decision to every case.

Despite their effectiveness, these methods share an implicit assumption that the reliability of the language model is a property of the model, and therefore a constant across decision cases. Mode choice is where the assumption breaks. Some cases are settled by numbers: a two-hour walk against a ten-minute train ride leaves nothing for semantics to add, and a semantic term can only inject noise into a margin the attributes have already resolved. Other cases are settled by exactly what the attribute table omits — a late-night leg described as unsafe, a transfer described as unreliable, a service description a traveler reads as a reason to stay away. A single weight is wrong in both directions at once: set high enough to help the semantic cases, it over-trusts the numeric ones; set low enough to be safe on the numeric ones, it under-uses the only cases where the language model was ever going to contribute.

Our preliminary analysis (Figure 1) reveals three observations. First, teacher accuracy measured case by case is dispersed rather than flat: on each benchmark the teacher is near-oracle on one part of the test set and near-chance on another, so reliability is a profile over cases rather than a number attached to the model. Second, this profile can be read off the teacher itself: how sharply it spreads its scores within a choice set moves together with whether it is about to be right, which means the weight a case deserves is already available at scoring time, without any additional supervision. Third, the damage from a wrong teacher is not local: under a teacher biased toward one alternative, a mixture whose weight was tuned on clean validation data loses accuracy and calibration across the board, including on the numeric cases where the semantic signal was never needed.

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.

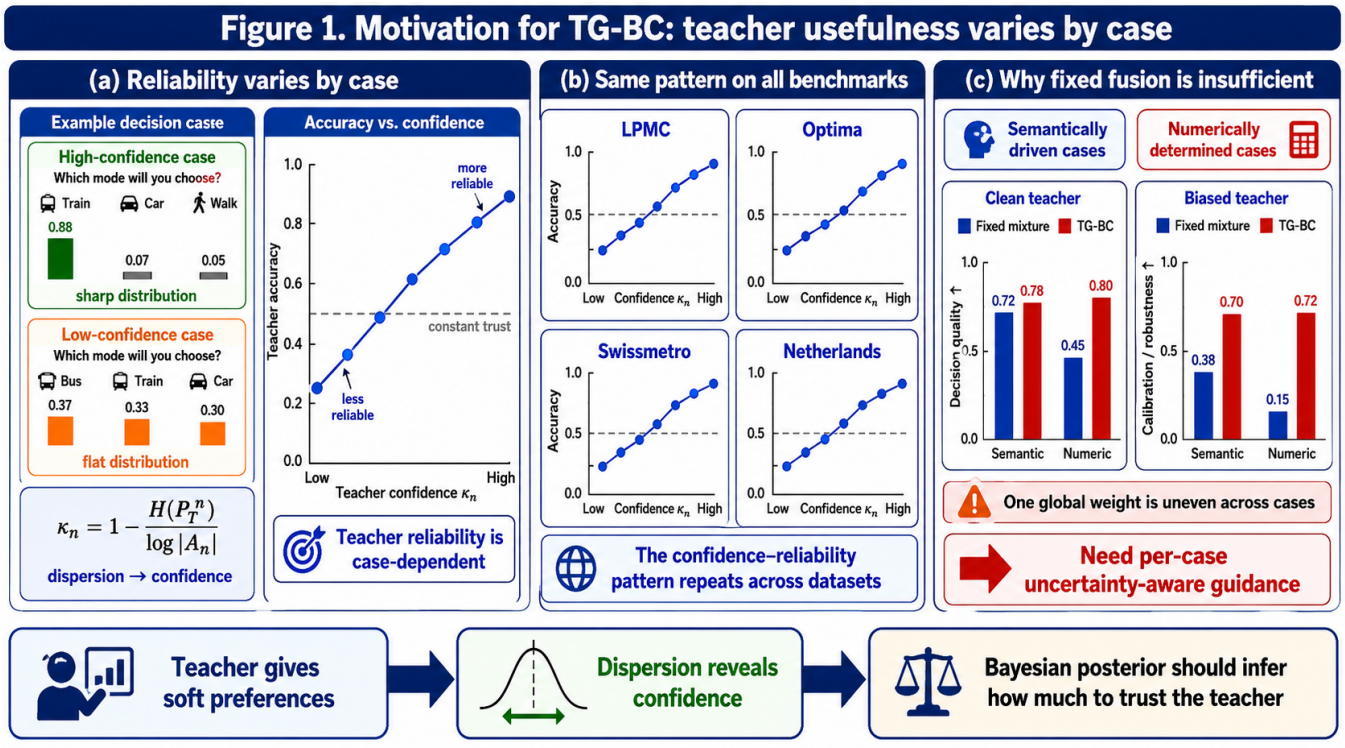

Figure 1: Motivation for TG-BC: teacher usefulness varies across decision cases. (a) Teacher confidence derived from score dispersion tracks case-level reliability. (b) The confidence–reliability pattern repeats across four mode-choice benchmarks. (c) A single fixed fusion weight is uneven across semantically driven and numerically determined cases, motivating per-case uncertainty-aware guidance.

These observations suggest that the key is not how much to trust the language model, but where it has earned that trust. What the standard pipeline discards is precisely what the teacher already reveals about itself. Instead of collapsing the teacher output into a predicted alternative or into a probability that enters under a fixed weight, we keep the full score distribution and use its own dispersion as a soft, per-case weight, so that unconfident cases are down-weighted rather than dropped — and so that the overall strength of the semantic term becomes a quantity the observed choices can revise, rather than one the analyst must choose in advance.

This leads to the question we study: *How can a semantic signal whose reliability varies from case to case enter an interpretable utility model as evidence whose weight is inferred from the observed choices rather than fixed in advance, so that predicted probabilities stay calibrated and candidate interventions stay rankable even when the signal is wrong?* We answer it with TG-BC, a two-stage framework. The first stage turns teacher output into per-case evidence: it scores each alternative with a frozen teacher, keeps the resulting soft preference distribution on the same simplex as the choice model, and derives a per-case weight from the dispersion of that distribution — addressing observations (i) and (ii) with no distillation weight, temperature, or top-k cutoff to tune. The second stage infers the utility coefficients and a global trust level jointly from the observed choices, so that consistent teacher evidence gains influence and inconsistent evidence is shrunk away — addressing observation (iii) — and then propagates the result into action values with credible intervals for downstream policy comparison.

Our contributions are summarized as follows:

- To the best of our knowledge, we are among the first to formulate language-model preference signals for structured choice as per-case soft evidence whose weight is itself inferred, rather than as labels, features, or a fixed-weight forecast (Section 3).
- We develop TG-BC, which reads a per-case reliability signal from the teacher's own score dispersion and thus admits it as evidence whose strength is inferred jointly with the utility coefficients — a two-stage procedure that keeps the teacher frozen, introduces no distillation weight to tune, and queries the language model only at training time (Algorithms 1–2).
- On four public mode-choice benchmarks, TG-BC improves calibration and counterfactual action ranking over Bayesian, context-aware, neural, and fusion baselines (Tables 1–2), and degrades gracefully under a biased teacher, where the inferred guidance strength shrinks by roughly two thirds (Table 3); the gains concentrate where semantic context is present and recede where choices are numerically determined (Appendix, Limitations).

2 Related Work

2.1 Context in Discrete Choice Models

Random utility maximization gives choice modeling its interpretive contract. McFadden made qualitative choice tractable through a likelihood over alternatives (McFadden 1974). Ben-Akiva and Lerman turned that likelihood into the standard workflow for reading attributes as policy levers (Ben-Akiva and Lerman 1985). Train showed how simulation-based estimation carries preference heterogeneity into the same framework (Train 2009). The contract, in all three, is that a fixed table of attributes explains the choice. Work on context effects loosens it. Seshadri et al. discovered context effects directly from raw choice data (Seshadri, Peysakhovich, and Ugander 2019). Tomlinson and Benson made feature-level context effects interpretable and statistically testable rather than hand-specified (Tomlinson and Benson 2021). Rosenfeld et al. aggregated over variable alternative sets so that the composition of the set informs the prediction (Rosenfeld, Oshiba, and Singer 2020a). Sifringer et al. split utility into a knowledge-driven and a data-driven component (Sifringer, Lurkin, and Alahi 2020). Wong and Farooq embedded residual layers inside a logit structure to capture nonlinear cross-effects (Wong and Farooq 2021). Han et al. learned taste heterogeneity through a neural module while preserving behaviorally meaningful coefficients (Han et al. 2022). Lukawska et al. moved the same idea into Bayesian territory, mapping context into occasion-specific preference shifts (Łukawska, Jensen, and Rodrigues 2025a). Van Cranenburgh et al. reviewed the tension these methods manage between predictive flexibility and behavioral identifiability (van Cranenburgh et al. 2022). These works establish that preference is context-dependent and that the dependence can be learned rather than assumed. They differ from ours in where context is permitted to live: each reads context out of structured variables or out of the composition of the choice set, so a description that never becomes a column can never enter the utility. TG-BC differs by treating free-text semantics as a source of context that enters the model as evidence rather than as a feature.

2.2 Language Models Meeting Choice Modeling

A second line puts a language model inside the choice pipeline, and the four ways of doing so differ mainly in the role assigned to it. Nishida et al. used the language model as the predictor, scoring alternatives from text-form choice cases across several alternative sets (Nishida, Ishigaki, and Onishi 2025a). Mo et al. used it as a feature extractor, passing generated representations to a supervised classifier for travel behavior (Mo et al. 2026). Sfeir et al. used it as a specification generator, asking it to propose a utility specification that is then estimated on the data (Sfeir et al. 2025). Goli and Chintagunta treated it as a respondent and documented where synthetic preferences part from human ones (Goli and Singh 2024). Voltes-Dorta et al. reached a similar caution in a transport setting, finding language-model answers useful as hypotheses but unsafe as ground truth (Voltes-Dorta and Suau-Sanchez 2025). These works establish that language models carry usable preference information about decision cases, and that the information degrades when it is taken at face value. They differ from ours in when the role is fixed: predictor, feature, specification, or respondent is chosen before any evidence about whether the role is warranted, and the choice then applies uniformly to every case. TG-BC differs by leaving the weight of that role open — the teacher supplies evidence, and how much evidence each case contributes is decided by the observed choices.

2.3 Distillation and Generalized Bayesian Updating

A third line supplies the mechanism we borrow. Bucila et al. trained a compact model on the predictions of a large ensemble (Bucila, Caruana, and Niculescu-Mizil 2006). Hinton et al. formalized this as distillation from softened teacher probabilities (Hinton, Vinyals, and Dean 2015). Kim and Rush extended it to sequence-level outputs (Kim and Rush 2016). Sun et al. supervised a student from intermediate teacher layers rather than final logits alone (Sun et al. 2019). Sanh et al. and Jiao et al. moved distillation into pretraining and into a two-stage transfer procedure (Sanh et al. 2019; Jiao et al. 2020). Gou et al. surveyed the resulting design space of knowledge types and student architectures (Gou et al. 2021). In the language-model era the teacher signal became richer: Wang et al. bootstrapped instruction data from the teacher's own generations (Wang et al. 2023), Hsieh et al. transferred rationales as auxiliary supervision (Hsieh et al. 2023), Mukherjee et al. learned from explanation traces (Mukherjee and others 2023), Gu et al. replaced the forward divergence with a reverse-divergence objective suited to generative students (Gu et al. 2024), and Agarwal et al. corrected the train-inference mismatch by distilling on the student's own samples (Agarwal et al. 2024). Gudibande et al. then showed the limit of the whole family: imitating teacher outputs improves style while leaving capability gaps the imitation data never covered (Gudibande et al. 2024). Two works give the mechanism its formal footing. Fang et al. cast teacher feedback as a teacher-informed prior, connecting distillation to posterior uncertainty rather than to a training loss (Fang et al. 2024). Bissiri et al. provided the general construction behind it, updating beliefs coherently with a loss in place of a likelihood (Bissiri, Holmes, and Walker 2016). These works establish that a teacher is transferred by weighting a student's own likelihood, and that the weight, not the architecture, is what carries the transfer. We import this construction into choice modeling: teacher scores become an evidence mass on the same simplex as the choice probabilities, the dispersion of those scores becomes the per-case exponent on that evidence, and generalized Bayesian updating supplies the lens under which a single trust level is inferred instead of tuned. Because the import must earn its complexity, two simpler members of the same family — a validation-tuned probability mixture and a maximum-a-posteriori fit of the same utility — are retained as baselines in Section 4.

2.4 Decision Support Under Posterior Uncertainty

The last line explains why the student is Bayesian rather than merely regularized. Allenby and Rossi argued for hierarchical priors as the practical route to individual-level preferences from sparse choice data (Allenby and Rossi 1998). Rossi et

Algorithm 1 Teacher Evidence Construction

Require: Data $\mathcal{D}$; teacher F_T; prior $p_0(\theta)$
Ensure: Evidence $\mathcal{G}_T = \{P_T^n, z_{nj}^T, \kappa_n, m_{nj}\}_{n,j}$

$q^{(0)}(\theta) \leftarrow \textsc{BayesDCM}(\mathcal{D}, p_0)$
for $n = 1$ to N **do**
  $\bar{P}_0^n(j) \leftarrow \mathbb{E}_{q^{(0)}}[P_\theta(j \mid A_n, c_n)],\ j \in A_n$
  $r_T^n \leftarrow F_T(\textsc{Prompt}(A_n, x_n, c_n, \bar{P}_0^n))$
  $P_T^n \leftarrow \mathrm{softmax}(r_T^n);\ \kappa_n \leftarrow 1 - H(P_T^n)/\log|A_n|$
  for $j \in A_n$ **do**
    $z_{nj}^T \leftarrow r_{Tj}^n - |A_n|^{-1}\sum_{k\in A_n} r_{Tk}^n$
    $m_{nj} \leftarrow \kappa_n P_T^n(j)$
  end for
end for
return $\mathcal{G}_T$

Algorithm 2 Teacher-Guided Bayesian DCM Learning and Decision Support

Require: $\mathcal{D}$, $\mathcal{G}_T$, priors $p_0(\theta), p_0(\rho)$, actions $\mathcal{B}$
Require: Intervention T_a and rewards $r_{nj}^{(a)}$
Ensure: Model $\mathcal{M}_{\text{TG-BC}}$, values and intervals, action $a^\star$

Set $\theta = (\beta, \gamma)$ and inferable teacher reliability ρ
$U_\theta(n, j) \leftarrow \beta^\top x_{nj} + \gamma z_{nj}^T$
$P_\theta(j \mid A_n, c_n) \leftarrow \mathrm{softmax}_{j\in A_n}(U_\theta(n, j))$
Define the teacher-informed posterior

$$\pi(\theta, \rho) \propto p_0(\theta)p_0(\rho)\prod_{n=1}^{N} P_\theta(y_n \mid A_n, c_n) \times \prod_{n=1}^{N}\prod_{j\in A_n} P_\theta(j \mid A_n, c_n)^{\rho m_{nj}}$$

$q^\star(\theta, \rho) \leftarrow \textsc{Infer}(\pi)$
$\mathcal{M}_{\text{TG-BC}} \leftarrow (q^\star, U_\theta, P_\theta)$
for $a \in \mathcal{B}$ **do**
  $x_{nj}^{(a)} \leftarrow T_a(x_{nj})$, for all n, j
  $S_a(\theta) \leftarrow \sum_n \sum_{j\in A_n} r_{nj}^{(a)} P_\theta(j \mid A_n^{(a)}, c_n)$
  $V(a) \leftarrow \mathbb{E}_{q^\star}[S_a(\theta)];\ I(a) \leftarrow \mathrm{CI}_{q^\star}[S_a(\theta)]$
end for
$a^\star \leftarrow \arg\max_{a\in\mathcal{B}} V(a)$
return $\mathcal{M}_{\text{TG-BC}}, \{V(a), I(a)\}_{a\in\mathcal{B}}, a^\star$

al. gave the unified Bayesian treatment, including decision-oriented interpretation of the posterior (Rossi, Allenby, and McCulloch 2005). Becker et al. separated inter- from intra-consumer heterogeneity in hierarchical Bayes logit mixtures (Becker et al. 2018). Berry et al. used random-coefficients demand to make equilibrium counterfactuals estimable under price and product interventions (Berry, Levinsohn, and Pakes 1995). Heckman marked the boundary of such exercises, showing how selection and endogeneity break the causal reading of discrete outcomes (Heckman 1978). Downstream, Van Ryzin and Mahajan linked choice probabilities to assortment-level optimization (Van Ryzin and Mahajan 1999), Kok and Fisher estimated substitution-aware demand for the same purpose (Kök and Fisher 2007), Jiang et al. balanced relevance against diversity through nested logit probabilities (Jiang, Qi, and Sun 2014), and Danaf et al. deployed choice models that update inside interactive recommendation (Danaf et al. 2019). On the evaluation side, Swaminathan and Joachims formulated counterfactual risk minimization from logged feedback (Swaminathan and Joachims 2015), London and Sandler recast it in posterior terms (London and Sandler 2019), and Villarraga and Daziano showed that approximate Bayesian inference improves uncertainty coverage in deep choice models (Villarraga and Daziano 2025). These works establish that a posterior over preferences, not a point estimate, is what makes an intervention comparison honest — and that the honesty is fragile as soon as model-implied shifts are read as causal effects. They are complementary to ours: they ask what to do with the uncertainty a choice model already owns, whereas we ask how an outside signal may add to that model without corrupting the uncertainty it reports. We inherit their reporting discipline, presenting action values with credible intervals and stopping short of causal claims from a model-based simulator.

3 Method

3.1 Problem statement.

We study structured choice prediction with an external LLM teacher. Let

$$\mathcal{D} = \{(A_n, \{x_{nj}\}_{j\in A_n}, c_n, y_n)\}_{n=1}^N$$

denote the observed choice data, where A_n is the choice set for instance n, $x_{nj} \in \mathbb{R}^p$ is the structured attribute vector of alternative j, c_n is the optional context, and $y_n \in A_n$ is the observed human choice. Let $J_n = |A_n|$. The goal is to learn a probabilistic choice model that predicts the selected alternative and supports downstream decision simulation.

We use a Bayesian DCM as the student model. Let θ denote the student parameters. For each alternative $j \in A_n$, the student assigns a latent utility $U_\theta(n, j)$ and induces a choice probability

$$P_\theta(j \mid A_n, c_n) = \frac{\exp(U_\theta(n, j))}{\sum_{k\in A_n}\exp(U_\theta(n, k))}.$$

Without teacher information, a standard Bayesian DCM learns $p(\theta \mid \mathcal{D})$ from the human choice likelihood. In this work, we instead learn a teacher-informed posterior

$$p(\theta, \rho \mid \mathcal{D}, T),$$

where T denotes the LLM teacher evidence and $\rho \geq 0$ is a learned global teacher reliability variable.

The LLM teacher is used as a semantic scorer rather than a direct decision rule. For each choice instance, the teacher produces an alternative-level soft preference distribution

$$P_T^n(\cdot) \in \Delta^{J_n-1},$$

where $P_T^n(j)$ represents the teacher's preference belief for alternative j. This distribution is not treated as ground truth. It is treated as uncertain semantic evidence. To avoid manually tuning a teacher weight, we compute an instance-level teacher confidence from the entropy of the teacher distribution:

$$\kappa_n = 1 - \frac{H(P_T^n)}{\log J_n},$$

where $\kappa_n \in [0, 1]$. A confident teacher distribution has low entropy and a larger κ_n, while an uninformative distribution has high entropy and a smaller κ_n.

The teacher distribution is converted into two objects. First, we define a teacher semantic signal z_{nj}^T for utility guidance. Second, we define a teacher evidence mass

$$m_{nj} = \kappa_n P_T^n(j),$$

which measures how much soft evidence the teacher assigns to alternative j. The teacher-guided utility is written as

$$U_\theta(n, j) = \beta^\top x_{nj} + \gamma z_{nj}^T,$$

where β keeps the interpretable DCM coefficients and γ controls how the teacher semantic signal enters the latent utility. Both are included in θ. The teacher signal therefore guides the latent utility, but its strength is learned through posterior inference rather than fixed manually.

We define the teacher-informed posterior as

$$p(\theta, \rho \mid \mathcal{D}, T) \propto p_0(\theta) p_0(\rho) \prod_{n=1}^{N} P_\theta(y_n \mid A_n, c_n) \prod_{n=1}^{N} \prod_{j \in A_n} P_\theta(j \mid A_n, c_n)^{\rho m_{nj}}.$$

The first likelihood term is the observed human choice evidence. The second term is the LLM teacher evidence. This formulation treats LLM outputs as soft evidence over the choice simplex, rather than as hard labels or as a replacement for the DCM. The Bayesian posterior jointly learns the utility parameters and the teacher reliability, allowing the model to use informative teacher signals while reducing reliance on uncertain ones.

3.2 Method overview.

The method has two stages. Algorithm 1 converts LLM scores into the evidence package $\mathcal{G}_T = \{P_T^n, z_{nj}^T, \kappa_n, m_{nj}\}_{n,j}$. Algorithm 2 then infers the teacher-guided posterior, evaluates candidate actions under posterior uncertainty, and recommends the action with the largest expected value.

3.3 Teacher Evidence Construction

Algorithm 1 fits a human-choice-only Bayesian DCM and uses its posterior prediction to form a student-aware teacher prompt. The normalized teacher scores define case-level confidence κ_n, centered utility guidance z_{nj}^T, and soft evidence mass m_{nj}.

3.4 Teacher-Guided Bayesian DCM Learning and Decision Support

Algorithm 2 combines the human-choice likelihood with the soft teacher factor. The learned reliability ρ controls the teacher's posterior influence. Posterior draws are then propagated to action values and credible intervals.

4 Experiments

We evaluate TG-BC in an urban mobility mode-choice decision-support setting. The experiments are designed to answer three questions: (i) whether LLM teacher evidence improves held-out choice prediction beyond Bayesian discrete choice models (DCMs) and neural choice baselines, (ii) whether the improvement is calibrated enough for decision support, and (iii) whether the Bayesian posterior reduces reliance on the teacher signal when the teacher is noisy or biased.

4.1 Experimental Setup

Datasets. We use four public mode-choice datasets that have been used in recent LLM-based mode-choice evaluation: LPMC, Optima, Swissmetro, and Netherlands (Nishida, Ishigaki, and Onishi 2025b). Each observation contains a traveler, a finite set of available travel modes, alternative-specific structured attributes, and the observed chosen mode. Structured attributes include travel time, monetary cost, access or wait time, service frequency, availability, and reliability when available. Following the text-conversion protocol of prior LLM mode-choice work, each choice occasion is converted into a short natural-language description listing the traveler context and the attributes of each available mode. The chosen label is never shown to the LLM teacher. We use a 70/15/15 train/validation/test split and average over five random splits.

Teacher signal. The LLM teacher receives the textual choice case and returns one score in $[0, 100]$ for each alternative, together with a one-sentence explanation. Scores are normalized within the choice set and calibrated on the validation split before entering the TG-BC utility. The teacher is queried for the observed choice case and for counterfactual actions that change text or alternative attributes; all teacher outputs are cached. Unless otherwise stated, all LLM calls use temperature zero.

Baselines. We keep the comparison set compact and include only strong representatives from three families. *MNL* is a classical multinomial logit using structured attributes only. *Bayesian DCM* is a hierarchical Bayesian mixed logit with random coefficients for travel time, cost, and wait time. *C-MMNL* is a context-aware Bayesian mixed multinomial logit baseline that maps context into preference shifts (Łukawska, Jensen, and Rodrigues 2025b). *SDA* is a set-dependent neural choice model that captures choice-set effects without a Bayesian random-utility posterior (Rosenfeld, Oshiba, and Singer 2020b). *LLM-only* predicts with a softmax over calibrated LLM scores (Nishida, Ishigaki, and Onishi 2025b). *Naive fusion* linearly combines Bayesian DCM probabilities and LLM-only probabilities, with the mixing weight tuned on validation data. TG-BC uses the same structured features and the same cached teacher scores as these baselines.

Metrics. For prediction, we report accuracy and negative log-likelihood (NLL). For calibration, we report Brier score and expected calibration error (ECE) with 15 equal-frequency bins. For decision support, we report normalized expected action value, posterior risk, and regret against a held-out semi-synthetic oracle simulator. The oracle is a high-capacity mixed-logit generator with nonlinear context effects fit on the training split and evaluated on counterfactual actions. All uncertainty numbers are computed from 256 posterior samples. Reported error bars are standard errors over five random splits.

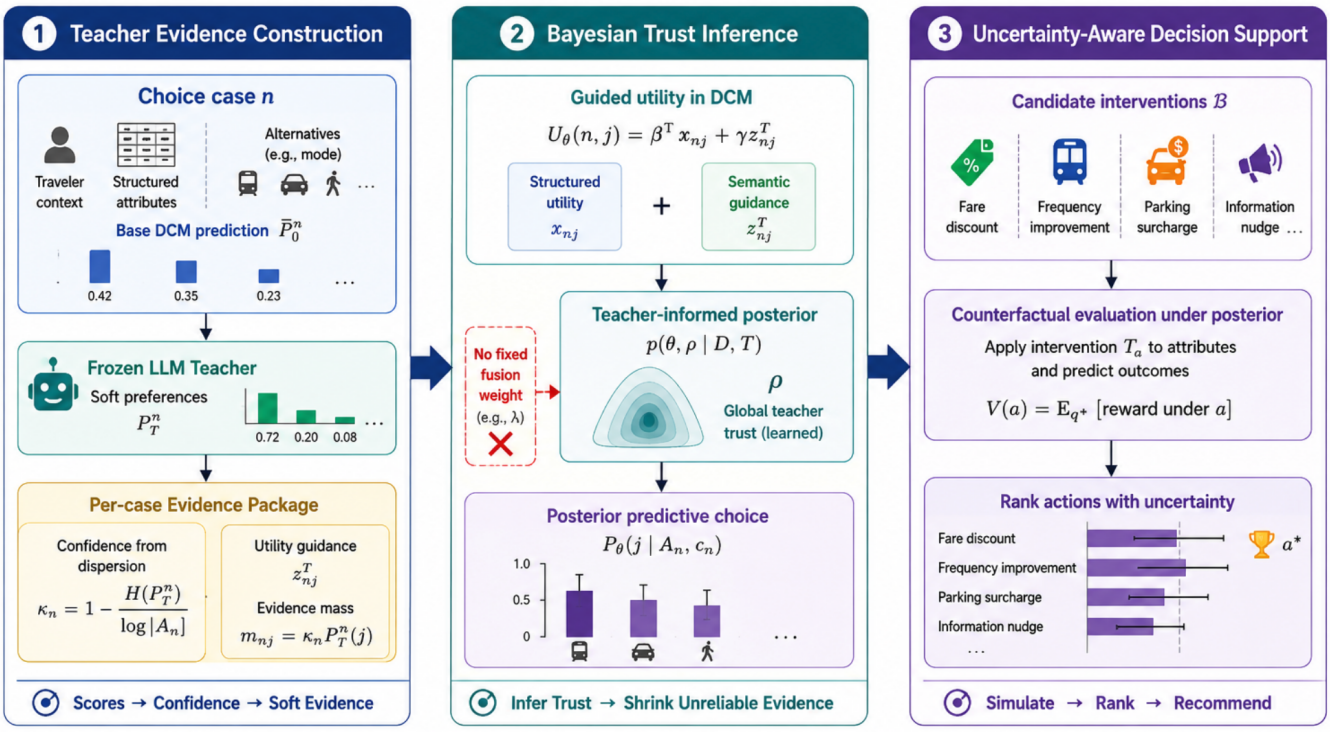

Figure 2: Overview of TG-BC. (1) A frozen LLM teacher converts each choice case into confidence-weighted soft evidence. (2) Bayesian trust inference jointly learns structured utility and global teacher trust, shrinking unreliable evidence and producing posterior predictive choices. (3) Posterior counterfactual simulation evaluates candidate interventions and ranks actions with uncertainty.

4.2 Prediction and Calibration

Table 1 reports the main held-out results. TG-BC achieves the best average accuracy and the lowest probabilistic error. The gain over C-MMNL is modest in accuracy but larger in NLL and ECE, which is important because the downstream decision layer depends on calibrated posterior probabilities rather than only top-1 prediction. LLM-only performs competitively in accuracy, especially on Swissmetro and Netherlands, but its ECE is substantially worse. This confirms that the teacher signal is informative but should not be used as a direct decision rule. Naive fusion improves over both individual components, but it remains less calibrated than TG-BC because it does not propagate uncertainty about when the LLM should be trusted.

4.3 Decision-Support Quality

We next evaluate the posterior decision simulator with four feasible actions: transit fare discount, service-frequency improvement, parking surcharge, and information nudge, plus a no-intervention baseline. The reward is sustainable-mode adoption minus normalized implementation cost. Table 2 reports the top action selected by each method. TG-BC selects the service-frequency intervention with the highest expected value and the lowest regret. The deterministic and LLM-only baselines either overvalue the low-cost information nudge or underestimate uncertainty under parking-related shifts. Bayesian DCM and C-MMNL provide uncertainty, but their expected values are lower because they do not use the LLM teacher's semantic assessment of reliability and service descriptions. This result supports the main decision-support claim: the LLM signal is most useful when it is integrated as shrinkable posterior evidence rather than as a standalone choice rule.

4.4 Ablation and Robustness

Table 3 isolates the components of TG-BC. Removing either scalar utility guidance or latent steering reduces accuracy and increases regret, showing that the two LLM-integration paths are complementary. Removing calibration mainly hurts ECE, while replacing the posterior with a MAP estimate keeps accuracy close but substantially increases ECE and decision regret. The last two rows stress-test the teacher by adding a fixed car-favoring bias to the LLM scores. Naive fusion degrades sharply because it treats the corrupted teacher as another calibrated predictor. TG-BC degrades less because the posterior shrinks the teacher coefficient γ from 0.42 to 0.13, reducing the influence of the biased signal. This posterior shrinkage is the key robustness mechanism of the proposed framework.

Table 1: Held-out prediction and calibration results. Accuracy is reported separately by dataset; NLL, ECE, and Brier are averaged across datasets. Higher accuracy is better; lower NLL, ECE, and Brier are better.

Method	LPMC	Optima	Swiss.	Neth.	Avg. Acc.	NLL	Brier	ECE
MNL	0.642	0.694	0.665	0.736	0.684	0.806	0.201	0.090
Bayesian DCM	0.655	0.704	0.676	0.744	0.695	0.782	0.191	0.053
C-MMNL	0.672	0.715	0.688	0.750	0.707	0.760	0.184	0.047
SDA	0.681	0.721	0.695	0.752	0.712	0.754	0.186	0.065
LLM-only	0.675	0.726	0.711	0.759	0.718	0.789	0.198	0.122
Naive fusion	0.694	0.736	0.724	0.766	0.730	0.724	0.173	0.071
TG-BC	**0.718**	**0.752**	**0.739**	**0.778**	**0.747**	**0.684**	**0.156**	**0.031**

Table 2: Counterfactual decision-support results. Value is normalized sustainable-mode reward minus action cost. Regret is computed against the oracle-best action on held-out counterfactuals. Posterior risk is unavailable for deterministic baselines.

Method	Recommended action	Value ↑	Risk ↓	Regret ↓
MNL	Parking surcharge	0.071	–	0.064
Bayesian DCM	Parking surcharge	0.085	0.33	0.047
C-MMNL	Service frequency	0.104	0.21	0.026
SDA	Parking surcharge	0.092	–	0.041
LLM-only	Information nudge	0.065	–	0.089
Naive fusion	Service frequency	0.116	–	0.020
TG-BC	**Service frequency**	**0.124**	**0.08**	**0.000**

Table 3: Ablation and teacher-stress results averaged over the four datasets. γ denotes the posterior mean of the scalar LLM-guidance coefficient.

Variant	Avg. Acc. ↑	ECE ↓	Regret ↓	$\mathbb{E}[\gamma]$
TG-BC full	**0.747**	**0.031**	**0.000**	0.42
No LLM-score calibration	0.737	0.084	0.017	0.47
No scalar guidance γs^{cal}	0.739	0.039	0.015	0.00
No latent steering Wz^{LLM}	0.733	0.046	0.024	0.36
MAP instead of posterior	0.742	0.067	0.029	–
Naive fusion, biased teacher	0.681	0.153	0.121	–
TG-BC, biased teacher	0.716 ± 0.05	0.052 ± 0.03	0.041 ± 0.06	0.13 ± 0.04

5 Conclusion

We presented TG-BC, a teacher-guided Bayesian framework for structured choice prediction and decision support. Rather than using an LLM as an opaque end-to-end decision maker, TG-BC treats its outputs as uncertain semantic evidence and integrates them into an interpretable Bayesian discrete choice model. This design preserves the structured utility representation of classical choice models while allowing semantic context to inform both scalar utility guidance and latent preference steering. Across the evaluated mode-choice benchmarks, TG-BC consistently improves the quality and calibration of predictive probabilities, and its posterior decision layer translates these gains into lower-risk intervention rankings. The ablation and teacher-bias experiments further show that posterior inference is important not only for uncertainty quantification, but also for shrinking unreliable teacher signals when they conflict with observed choices. Together, these results support a complementary role for LLMs in structured decision support: they can enrich interpretable models with semantic information without replacing the probabilistic structure needed for auditing and counterfactual analysis.

6 Limitations and Future Works

A key limitation is that our experiments have not yet been extended to substantially larger-scale teacher LLMs. Consequently, the current results do not establish how TG-BC scales across model sizes or whether the additional semantic capability of frontier-scale teachers produces commensurate improvements in choice prediction, calibration, and decision quality. Larger teachers may provide more nuanced contextual judgments, but they also introduce higher API cost and latency, greater dependence on proprietary and changing model versions, and additional reproducibility constraints. Future work should therefore conduct controlled comparisons across teacher scales and model families, reporting both statistical performance and the full cost–quality tradeoff. It should also investigate whether teacher-score caching, smaller open models, or distillation can retain most of the benefits of larger teachers while preserving the efficiency and auditability required for practical decision-support systems.

Teacher evidence is also sensitive to prompt wording, decoding settings, model updates, and the verbalization of structured records. Although posterior shrinkage reduces conflicting signals, it cannot guarantee robustness under joint shifts in traveler populations, alternatives, and semantic descriptions. The reported action values are likewise model-based scenario rankings rather than causal effect estimates, and TG-BC may inherit social or geographic biases from the teacher. Practical deployment therefore requires versioned prompts, subgroup calibration and fairness audits, re-estimation after pipeline changes, causal validation, and human review.

Future Work includes larger scale of Teacher model and more varies of Sutdent Discrete Choice Model Establishments.

References

Agarwal, R.; Vieillard, N.; Zhou, Y.; Stanczyk, P.; Ramos Garea, S.; Geist, M.; and Bachem, O. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Allenby, G. M., and Rossi, P. E. 1998. Marketing models of consumer heterogeneity. *Journal of Econometrics* 89(1–2):57–78.

Becker, F.; Danaf, M.; Song, X.; Atasoy, B.; and Ben-Akiva, M. 2018. Bayesian estimator for logit mixtures with inter- and intra-consumer heterogeneity. *Transportation Research Part B: Methodological* 117:1–17.

Ben-Akiva, M. E., and Lerman, S. R. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press.

Berry, S.; Levinsohn, J.; and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Bissiri, P. G.; Holmes, C. C.; and Walker, S. G. 2016. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(5):1103–1130.

Bucila, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Danaf, M.; Becker, F.; Song, X.; Atasoy, B.; and Ben-Akiva, M. 2019. Online discrete choice models: Applications in personalized recommendations. *Decision Support Systems* 119:35–45.

Fang, L.; Chen, Y.; Zhong, W.; and Ma, P. 2024. Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 12935–12956. PMLR.

Goli, A., and Singh, A. 2024. Frontiers: Can large language models capture human preferences? *Marketing Science* 43(4):709–722.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Gudibande, A.; Wallace, E.; Snell, C.; Geng, X.; Liu, H.; Abbeel, P.; Levine, S.; and Song, D. 2024. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations*.

Han, Y.; Pereira, F. C.; Ben-Akiva, M.; and Zegras, C. 2022. A neural-embedded discrete choice model: Learning taste representation with strengthened interpretability. *Transportation Research Part B: Methodological* 163:166–186.

Heckman, J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46(4):931–959.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017. Association for Computational Linguistics.

Jiang, H.; Qi, X.; and Sun, H. 2014. Choice-based recommender systems: A unified approach to achieving relevancy and diversity. *Operations Research* 62(5):973–993.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Kim, Y., and Rush, A. M. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kök, A. G., and Fisher, M. L. 2007. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* 55(6):1001–1021.

London, B., and Sandler, T. 2019. Bayesian counterfactual risk minimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 4125–4133. PMLR.

Łukawska, M.; Jensen, A. F.; and Rodrigues, F. 2025a. Context-aware bayesian mixed multinomial logit model. *Journal of Choice Modelling* 54:100536.

Łukawska, M.; Jensen, A. F.; and Rodrigues, F. 2025b. Context-aware bayesian mixed multinomial logit model. *Journal of Choice Modelling* 54:100536.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., ed., *Frontiers in Econometrics*. New York: Academic Press. 105–142.

Mo, B.; Xu, H.; Ma, R.; Cho, J.-H.; Zhuang, D.; Guo, X.; and Zhao, J. 2026. Large language models for travel behavior prediction. *Transportation Research Interdisciplinary Perspectives* 38:102124.

Mukherjee, S., et al. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint*.

Nishida, R.; Ishigaki, T.; and Onishi, M. 2025a. Large language models predict transportation mode choice behavior for a variety of alternative sets. *Transportation Research Record: Journal of the Transportation Research Board* 2679(12).

Nishida, R.; Ishigaki, T.; and Onishi, M. 2025b. Large language models predict transportation mode choice behavior for a variety of alternative sets. *Transportation Research Record: Journal of the Transportation Research Board* 2679(12):249–264.

Rosenfeld, N.; Oshiba, K.; and Singer, Y. 2020a. Predicting choice with set-dependent aggregation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8220–8229. PMLR.

Rosenfeld, N.; Oshiba, K.; and Singer, Y. 2020b. Predicting choice with set-dependent aggregation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8220–8229. PMLR.

Rossi, P. E.; Allenby, G. M.; and McCulloch, R. E. 2005. *Bayesian Statistics and Marketing*. John Wiley & Sons.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS Workshop: Energy Efficient Machine Learning and Cognitive Computing*.

Seshadri, A.; Peysakhovich, A.; and Ugander, J. 2019. Discovering context effects from raw choice data. In Chaudhuri, K., and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5660–5669. PMLR.

Sfeir, G.; Nova, G.; Hess, S.; and van Cranenburgh, S. 2025. Can large language models assist choice modelling? insights into prompting strategies and current models capabilities. *arXiv preprint arXiv:2507.21790*.

Sifringer, B.; Lurkin, V.; and Alahi, A. 2020. Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Swaminathan, A., and Joachims, T. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 814–823.

Tomlinson, K., and Benson, A. R. 2021. Learning interpretable feature context effects in discrete choice. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1582–1592. Association for Computing Machinery.

Train, K. E. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition.

van Cranenburgh, S.; Wang, S.; Vij, A.; Pereira, F.; and Walker, J. 2022. Choice modelling in the age of machine learning—discussion paper. *Journal of Choice Modelling* 42:100340.

Van Ryzin, G., and Mahajan, S. 1999. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science* 45(11):1496–1509.

Villarraga, D. F., and Daziano, R. A. 2025. Bayesian deep learning for discrete choice. *arXiv preprint arXiv:2505.18077*.

Voltes-Dorta, A., and Suau-Sanchez, P. 2025. Can large language models mimic airline passenger preferences? *Artificial Intelligence for Transportation* 3–4:1–18.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Association for Computational Linguistics.

Wong, M., and Farooq, B. 2021. Reslogit: A residual neural network logit model for data-driven choice modelling. *Transportation Research Part C: Emerging Technologies* 126:103050.