

Predicting Collective Behavior with Large Language Models: The EBI Framework for Response-Side Inference

Yen Yu Shan
Academia Sinica
bubu3013@as.edu.tw

Li-Wun Chang
Academia Sinica
liwenc57@iis.sinica.edu.tw

Lun-Wei Ku
Academia Sinica
lwku@iis.sinica.edu.tw

Abstract

Collective behavior forecasting usually relies on historical outcome traces: past sales, past votes, or prior campaign performance. Such history-side records are useful when comparable cases exist, but they are brittle for new products, emerging political actors, or settings where historical patterns fail to generalize. We propose a complementary response-side approach: instead of only extrapolating from past aggregate outcomes, we prompt large language models as third-party domain analysts to infer collective outcomes from structured context and behavioral signals. We introduce EBI, an Exogenous–Behavioral–Inference framework that represents each task through external context, behavioral signals, and the target aggregate outcome. Across three real-world domains — health supplements, elections, and crowdfunding — we preliminarily evaluate whether response-side inference is competitive with history-side baselines, which are restricted to historical outcome data. We present the EBI system, a web-based tool designed for users with limited access to historical data, survey infrastructure, or analyst resources, allowing them to explore domain-specific EBI predictions, run live inference across multiple models, and ablate exogenous and behavioral features interactively to understand prediction drivers.

1 Introduction

The system addresses the problem of forecasting collective behavior when historical outcome records and survey infrastructure are sparse or absent. When a retailer wants to know how consumers will respond to a new product, surveys are the conventional approach but are expensive, time-consuming, and bounded by the data collected (Groves et al., 2009). Prior work has concentrated on modeling demand-side consumer behavior — preferences, purchase intent, and demand (Maier et al., 2025; Li et al., 2025; Caoui,

2026) — while predicting how a target audience would respond to a specific product remains comparatively underserved (Bansal et al., 2025; Zhu et al., 2025; Jannelli et al., 2025). Yet this gap carries measurable cost: decisions about new products, emerging political actors, or new crowdfunding campaigns are currently made without any forecasting signal — a gap this system directly addresses, with exploratory evidence of competitive performance against history-side baselines across all three domains (§5).

Most existing forecasting pipelines are outcome-history models. They predict tomorrow’s sales from yesterday’s sales, tomorrow’s vote share from past elections and polls, or tomorrow’s campaign success from previous platform records. In other words, they observe the world from the side of the seller, institution, or campaign — the side that records historical outcomes. This perspective is powerful when long comparable histories exist, but it becomes fragile when the forecaster faces a new product, a new political actor, a new market, or a target population whose future response has not yet been observed. Addressing this requires reasoning over exogenous and behavioral signals in place of missing outcome history.

We explore a complementary perspective: rather than extrapolating from historical outcomes, we ask whether an LLM can serve as a surrogate for a domain analyst — given structured context about the environment and observable behavioral signals — to forecast how a target population collectively responds. In consumer domains, this means predicting the monthly demand tertile; in elections, the vote-share distribution across parties; in crowdfunding, the monthly success rate — all without requiring population-scale data.

We propose EBI (Exogenous factors, Behavioral statistics, Inference target), a framework that decomposes collective behavior prediction tasks into three components: environmental context [E], ob-

servable behavioral signals [B], and the inference target [I]. An LLM is prompted as a third-party domain analyst to reason about how the target population would respond by drawing on its parametric world knowledge as a complementary prior, rather than relying on historical outcomes alone. This response-side approach remains applicable even when comparable historical outcome records are sparse or absent. We demonstrate EBI across three domains spanning three locales: Taiwan Health Supplements, Japan 2026 General Election, and US Kickstarter Crowdfunding Technology. Our contributions are twofold:

- **Framework.** We propose EBI for collective behavior prediction, in which LLMs act as domain-specific analysts provided with exogenous factors and contemporaneous behavioral signals, and show that it is preliminarily competitive with history-side baselines restricted to historical outcome data across all three evaluated domains.
- **System.** We present an interactive web-based demo of EBI prediction for users with limited access to historical records, survey infrastructure, or analyst resources. Users can see how different features impact the prediction through live inference (§4).

2 Related Work

Collective behavior prediction. Collective behavior prediction has been studied within individual domains, each requiring domain-specific modeling infrastructure. LLM-based time-series prompting (Gruver et al., 2023; Liu et al., 2024) improves zero-shot accuracy on historical outcome sequences such as monthly sales, an election’s vote share, or campaign success rates, but relies on the target series’ own historical values to produce a point forecast rather than a distribution over outcomes, leaving it unable to accommodate new entities with no comparable prior record. Agent-based LLM approaches (Park et al., 2023) target micro-level interaction dynamics rather than these macro-level outcomes. Survey-based forecasting aggregates stated preferences into population-representative estimates (Groves et al., 2009), but is bounded by the data collected and requires active respondent recruitment. Time-series and survey approaches provide no recourse when historical records or survey infrastructure are absent; agent-based approaches, by contrast, are limited by scope

rather than data availability — they do not directly yield macro-level distributional predictions.

Demand-side behavior modeling. Demand-side prediction has a long theoretical foundation in discrete choice modeling (McFadden, 1974). Recent work extends this to consumer contexts: Maier et al. (2025) show that LLMs reproduce human purchase-intent ratings via semantic elicitation, and Li et al. (2025) apply LLMs to consumer segmentation across product categories. Maier et al. (2025)’s preference distribution describes an existing product’s Likert ratings, and Li et al. (2025)’s segments describe an existing consumer base — both at a single point in time; neither describes a real-world outcome variable that must be forecast for a future period, such as next month’s sales tier, an election’s vote share, or a campaign’s funding success.

LLMs simulating human behavior. Argyle et al. (2023) and Sun et al. (2024) demonstrate that LLMs can simulate group-level behavioral tendencies from parametric knowledge, but both lack cross-domain generality, remaining confined to political opinion surveys rather than real-world collective outcomes such as monthly expenditure distributions, an election’s eventual vote share, or campaign success rates. EBI instead conditions the model on a single domain-analyst persona that reasons directly over population-level exogenous and behavioral signals, without generating or aggregating individual synthetic respondents.

EBI targets settings with limited or weakly comparable prior records, such as newly formed political parties, where response-side inference (Section 3) applies.

Integrating heterogeneous predictive signals. Prior forecasting work in electrical load forecasting shows that combining historical records with exogenous context can improve prediction accuracy (Christen et al., 2020). The relative importance of different signals is often not known a priori, motivating methods that learn how to combine multiple views (Xu et al., 2013; Urbanowicz et al., 2018). Our EBI system integrates both signal types into a single structured prompt and analyzes each feature’s contribution to prediction across the three real-world domains studied here.

3 The EBI System

EBI (Figure 1) is a framework instantiated as a web-based tool that leverages LLMs as synthetic

Domain	[E] Exogenous	[B] Behavioral	[I] Inference
HS TW	Epidemic index	Throw-in promotion	Sales tertile (Low/Mid/High)
JP Elec.	Cabinet approval	Party support dist.	7-party vote share (%)
KS Tech	S&P 500 return	% above 50% funded	Campaign success rate (%)

Table 1: Domains: HS TW = Taiwan Health Supplements; JP Elec. = Japan 2026 General Election; KS Tech = US Kickstarter Crowdfunding Technology. Full details in Appendix A.1.

domain-specific analysts to predict collective behavioral outcomes. It is domain-agnostic, instantiating the same Exogenous–Behavioral–Inference structure across three structurally diverse domains. EBI predicts sales tertile in Taiwan Health Supplements, vote share in Japan 2026 General Election, and success rate in US Kickstarter Crowdfunding Technology. Methodologically, EBI is agnostic to the model provider, supporting inference from both OpenAI and Anthropic models.

3.1 The EBI Framework

EBI operationalizes collective behavior prediction through three components:

- Exogenous factors [E]: environmental or demographic context that shapes collective behavior.
- Behavioral statistics [B]: observed or self-reported behavioral signals indicating current demand or preference.
- Inference [I]: the collective behavioral outcome of the target population.

Instantiation Table 1 illustrates the EBI instantiation across all three domains.

Prompt Design The prompt design of EBI assigns the model a domain-specific analyst role to predict collective behavioral outcomes from a third-party perspective. In this paper, we instantiate three roles: a market analyst for Taiwanese health supplements, an election analyst for the Japan 2026 general election, and a crowdfunding analyst for US Kickstarter campaign success. Each prompt presents [E], then [B], before eliciting the [I] prediction (Table 1). All experiments use $k=3$ few-shot examples drawn from the training window.

4 System Demo

The EBI system is publicly accessible at <https://ai-army-emnlp-demo-2026.vercel.app>. A video walkthrough is available at <https://youtu.be/H9Rr1fe6HUU>. Users supply their own API key, which is forwarded directly to the provider and never stored server-side. The system is released under the MIT License.

Each domain page presents performance metric cards (JSD, Spearman ρ , MAE), a prediction-versus-ground-truth breakdown, and an EBI-versus-baseline comparison chart switchable among the three metrics. Users select an inference model via dropdown; all visualizations update accordingly.

The system further provides three interactive components per domain: Guess It, Builder (both in Figure 2), and Try Your EBI (Figure 3). Guess It challenges users to identify the most impactful domain-specific feature, such as cabinet approval rating in the Japan 2026 General Election, before clicking to reveal a side-by-side comparison of the full EBI prediction and the ablated prediction, alongside the Δ JSD relative to the full EBI condition. Builder allows users to toggle exogenous factors and behavioral statistics blocks independently across four conditions: None, exogenous factors only, behavioral statistics only, and EBI. The predicted distribution and corresponding JSD, Spearman ρ , and MAE update in real time. Try Your EBI allows users to input exogenous factor values and behavioral statistics via domain-specific interactive controls, such as sliders, chips, and emoji-based ratings. By clicking “Run Prediction”, the system triggers a live API call and returns a three-way comparison of the user’s prediction, EBI, and the ground truth.

We conducted a small-scale user study with six participants spanning computer science and non-computer science backgrounds. After a brief tutorial with written explanations of the EBI system, we asked participants to freely explore the system’s Guess It, Builder, and Try Your EBI components. Participants explored at least one domain. Try Your EBI was the most recommended feature (4 of 6) and the most frequently cited as most useful (3 of 6). One participant said that it made it possible to “get the prediction and also see the reasons behind in a short time.” P1 worried that the numbers and terms shown were “beyond [my] knowledge” from the start of the session, while P5 found the volume of data, charts, and jargon overwhelming,

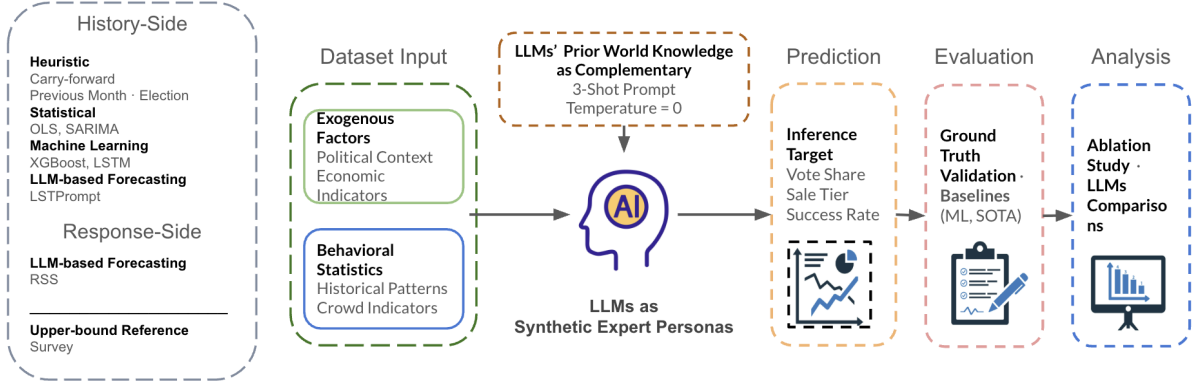


Figure 1: Left: History-side baselines. Right: EBI pipeline, which combines [E] and [B] with LLMs’ prior world knowledge to produce distributional predictions over domain-specific inference targets.

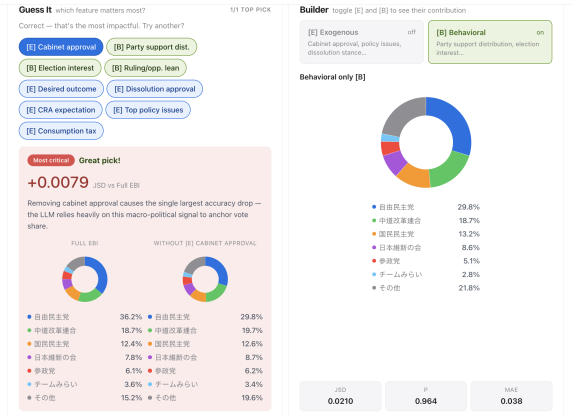


Figure 2: Guess It and Builder, Japan 2026 election

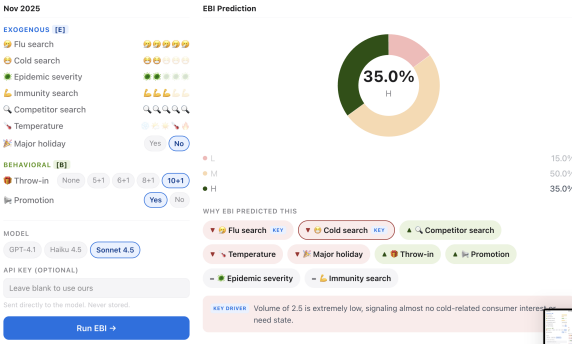


Figure 3: Try Your EBI, Taiwan Health Supplements.

unsure what to do next. Correspondingly, tutorials were the most requested addition (3 of 6). P2, one of the highest-scoring participants, reflected that the system let them “see exactly how these factors impact the performance using real figures instead of just guessing or speculating,” calling the data “persuasive.”

5 Evaluation

EBI is evaluated across the three domains (Table 3). Experimental setup is described in Section 5.1; results are in Section 5.2.

5.1 Experimental Setup

Baselines Baselines include naïve methods (Carry Forward, Moving Average), statistical methods (Ordinary Least Squares, OLS; Seasonal Autoregressive Integrated Moving Average, SARIMA), machine learning methods (eXtreme Gradient Boosting, XGBoost; Long Short-Term Memory, LSTM), and state-of-the-art methods (Long-Short-Term Prompting, LSTPrompt). In low-data settings, some baselines reduce to a single constant or near-constant prediction, for which ρ is undefined and reported as —. LSTPrompt serves as the SOTA baseline for Taiwan Health Supplements and US Kickstarter Crowdfunding Technology, since Liu et al. (2024) validates it on numeric time-series forecasting benchmarks. In practice, LSTPrompt decomposes forecasting into short-term and long-term sub-tasks via tailored prompts.

We compare EBI against history-side baselines that extrapolate from past aggregate outcomes and an upper-bound reference (Survey) included for context only. Survey represents pre-event polling averages, requiring large-scale real-time data collection unavailable in general deployment. Both EBI and history-side baselines only use data available before the prediction date. History-side methods have a structural blind spot: they cannot represent entities with no data in the historical record. For example, in the Japan 2026 election, Team-Mirai had no comparable prior record, so Carry

Method	Taiwan Health Supplements	Japan 2026 General Election	US KS Crowdfunding Tech.
Best EBI (JSD ↓)	0.166 (Sonnet)	0.006 (GPT-4.1)	0.0007 (GPT-4.1)
Best ML	0.333 (LSTM)	0.062** (XGBoost)	0.0022 (LSTM)
Best Statistical	—	0.093** (SARIMA)	0.0026 (SARIMA)
Best SOTA	0.662*** (LSTPrompt)	0.047** (RSS)	0.0027 (LSTPrompt)
Best Linear	0.333 (OLS)	0.090*** (OLS)	0.0020 (OLS)
Best Naive	0.639*** (MA-3)	0.067** (Carry Forward)	0.0021 (Carry Forward)

Table 2: Main experimental results (JSD ↓). EBI achieves the lowest JSD (**bold**) across all three domains. Significance vs. Best EBI: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Appendix A.3); unmarked = not significant. (—) indicates no non-degenerate baseline (Taiwan Health Supplements SARIMA predicts a constant class, JSD = 1.0). Tech. = Technology.

Domain	Target	CV	Δ
HS TW	Sales tertile	55.78%	+1 to +16
JP Elec.	Vote share (%)	74.50%	+7 to +20
KS Tech	Success rate (%)	3.35%	+1 to +21

Table 3: Domain abbreviations as in Table 1. CV = coefficient of variation of the target. Δ = target date — model cutoff in months, using the most conservative published cutoff for each model (GPT-4.1 knowledge cutoff 2024-06; Sonnet and Haiku training-data cutoff 2025-07). $\Delta > 0$: post-cutoff.

Forward, Moving Average, OLS, SARIMA, and XGBoost default to 0%. Across the evaluated domains, EBI also uses contemporaneous behavioral and exogenous signals that these baselines do not receive.

RSS serves as the SOTA baseline only for Japan 2026 General Election, since Sun et al. (2024) validates it on political opinion polls (ANES). By design, RSS simulates individual respondents by conditioning the same LLM on demographic distributions, without EBI’s structured context. EBI, as a response-side approach, represents each context through [E] exogenous factors and [B] behavioral statistics, allowing prediction even when prior outcome records do not exist.

Datasets To evaluate the EBI framework, we assemble prediction instances from diverse sources across multiple locales (Table 3).

Model Selection GPT-4.1’s knowledge cutoff is 2024-06, while Sonnet’s and Haiku’s training-data cutoffs are 2025-07. All models run at temperature = 0. We fix $k=3$ few-shot examples uniformly across all domains and models to provide prompt context while keeping prompts compact. To prevent data leakage, all models satisfy $t_{\text{cutoff}} < t_{\text{target}}$, where t_{target} is the prediction target date.

Time-Locked To ensure predictions are not confounded by memorized outcomes, we define $\Delta = t_{\text{target}} - t_{\text{cutoff}}$ (in months); $\Delta > 0$ indicates the target postdates parametric memory. All three domains are fully prospective ($\Delta > 0$) (Table 3).

5.2 Experimental Results

We evaluate the usefulness and domain-dependent generality of the EBI framework through exploratory experiments across all three domains. For each domain, we compared the EBI-enabled system against the baselines described in Section 5.1, all of which use only historical data available before the prediction date. Table 2 reports JSD across domains, with statistical significance testing detailed in the caption; full results including Spearman’s ρ and MAE are in Appendix Table 5.

EBI achieves the lowest JSD in every baseline category in all three domains (Table 2). In this experimental setting, EBI has access to limited historical records and contemporaneous behavioral signals, while most baselines are restricted to historical outcome data, except RSS, which follows a sample-and-aggregate approach by generating individual-level responses from an LLM conditioned on demographic distributions, whereas EBI directly produces population-level distributions via structured decomposition. Best EBI (Sonnet) achieves statistically significant gains over the SOTA and naive baselines in Taiwan Health Supplements. In Japan 2026 General Election, Best EBI (GPT-4.1) achieves statistically significant gains against all baselines. Moreover, Best EBI (GPT-4.1, JSD = 0.006065) is numerically lower than but statistically indistinguishable from the upper-bound reference, the pre-election survey (JSD = 0.006111, ns); this near-parity holds only for GPT-4.1, the best-performing model, as EBI Sonnet (JSD = 0.021427) and Haiku (JSD = 0.011080) trail the survey (Appendix Table 5). In US Kickstarter

Condition	HS TW	JP Elec.	KS Tech
Full EBI	0.166	0.006065	0.00075
None	0.645	0.018044***	0.00192
Exogenous only	0.368	0.002283 ***	0.00385
Behavioral only	0.166	0.020987***	0.00125
w/o best [B] feature	0.347	0.004317	0.00373

Table 4: Ablation ($k=3$ few-shot; JSD $\downarrow$; significance testing in Appendix A.3). **Bold** = best JSD per domain. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ vs. Full EBI; unmarked = not significant or not tested. Ablation models: Sonnet (HS TW, KS Tech), GPT-4.1 (JP Elec.). The best [B] feature: throw-in promotion (HS TW), party support distribution (JP Elec.), % above 50% funded (KS Tech).

Crowdfunding Technology, Best EBI (GPT-4.1) attains the lowest JSD in every baseline category.

In Taiwan Health Supplements, EBI’s JSD does not statistically significantly beat the best-performing ML (LSTM) and linear (OLS) baselines. The reason is only three test points, non-significant gains are expected (Cohen, 1988) given that training uses seven months. Correctly classifying all three months hinges on September, the one Low-tier month; August & October are High tier and covered by a constant-High guess. September had no promotion and recorded the lowest sales of the three held-out months. EBI classifies September correctly across all three models, while the history-side baselines (OLS, LSTM), which do not use behavioral promotion data, both predict High. This case-level pattern is consistent with the ablation study (§6), which finds promotion signals are this domain’s primary contributor. Given the limited test power behind the non-significance, the promotion-signal link, though consistent with domain-level ablation evidence, is only preliminary, single-case evidence.

In US Kickstarter Crowdfunding Technology, monthly success rates vary little (CV 3.35%; Table 3), so absolute differences between methods are small; paired month-level tests against the best baseline in each category do not reach statistical significance.

6 Discussion: Ablation Study

In Taiwan Health Supplements, the Behavioral-only condition matches Full EBI exactly on JSD, while Exogenous-only underperforms both. Feature-level ablation shows the same pattern: removing the throw-in promotion degrades JSD the most (Table 4). This raises the question of

what promotion information the Behavioral block contributes. Section 5.2 connects this pattern to the September test month, the one Low-tier month without a promotion in the test window. EBI predicted Low for September, while the non-behavioral baselines (OLS, LSTM) predicted High. This is preliminary evidence, based on a single test-month case. Behavioral-only and Full EBI match on JSD, but whether the Exogenous block is ignored or simply nets to zero remains unexplored.

In the Japan 2026 General Election, the Behavioral-only condition performs worse than both Full EBI and Exogenous-only. Feature-level ablation shows the same pattern: removing party support distribution improves JSD the most (Table 4). To identify which behavioral feature is most associated with this performance gap, we conducted a factorial ablation that keeps all exogenous factors fixed while varying three behavioral features: party support distribution, election interest, and voting tendency. Each feature is independently included or excluded, yielding all eight combinations. Across these combinations, Ishin, Sanseito, TeamMirai, and DPFP show strong positive correlations with higher JSD ($r = 0.86$ - 0.92), whereas LDP ($r = -0.80$) and CRA ($r = -0.92$) show negative correlations. Ishin, Sanseito, TeamMirai, and CRA’s predictions move closer to their party support percentage, as reflected in the model’s reasoning text. DPFP, however, moves away from its party support percentage, consistent with a few-shot example describing its 2024 result as exceeding its party support percentage. In contrast, LDP shows no such shift, consistent with the absence of party-support signals in its reasoning text. This explanation is preliminary, based on qualitative comparison of reasoning text; CRA’s negative correlation remains an unresolved contradiction despite its party-support-anchoring behavior matching the other three parties.

In US Kickstarter Crowdfunding Technology, EBI achieves the lowest JSD, ahead of None, Exogenous only, and Behavioral only. Among the three ablation conditions, Behavioral-only comes closest to Full EBI on JSD. Feature-level ablation echoes the same pattern: removing the proportion of campaigns reaching the 50% funding threshold degrades JSD the most (Table 4). Whether the Exogenous block adds meaningful signal beyond Behavioral statistics in this domain remains unclear: adding it lowers JSD relative to Behavioral only, yet the gap does not reach significance over

the eight test months.

The most impactful component differs by domain: Exogenous factors in the Japan 2026 General Election, Behavioral statistics in the remaining two. This is consistent with the domain-dependent, a priori unknown signal importance (§2). Evaluating both signal types through ablation makes these domain-specific differences explicit.

7 Conclusion

In this demonstration, we presented EBI, a system that integrates Exogenous factors and Behavioral statistics into a single prompt, and ablation results showed that their relative importance is domain-dependent across all three evaluated domains. In limited-scale experiments on real-world datasets, we found initial evidence that EBI can be a useful response-side alternative when history-side records are sparse or absent, remaining competitive with our evaluated history-side baselines, which are restricted to historical outcome data. In a small-scale user study, participants who found Try Your EBI most useful cited its ability to generate predictions from adjusted parameters, observe the model’s reasoning under different conditions, and experiment with inputs intuitively. In future work, we plan to test EBI’s robustness to wording-level variation in prompt phrasing and extend it to additional domains and locales.

Limitations

Experiments cover $n=3-8$ test months per domain across three locales; extending to additional domains, model families, and prompt configurations remains future work. Each domain’s history-side baselines are fit on the full historical training window, whereas EBI receives only $k=3$ few-shot historical examples in addition to contemporaneous exogenous and behavioral signals.

Acknowledgments

This work is partially supported by the National Science and Technology Council of Taiwan under Grant No. 114-2221-E-001-015-MY3, and by Academia Sinica.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate.

2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.

Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, Will Epperson, Tyler Payne, Jake M. Hofman, Brendan Lucier, Chinmay Singh, Markus Mobius, Akshay Nambi, Archana Yadav, Kevin Gao, David M. Rothschild, Aleksandrs Slivkins, Daniel G. Goldstein, Hussein Mozannar, Nicole Immorlica, and 5 others. 2025. [Magentic marketplace: An open-source environment for studying agentic markets](#). *Preprint*, arXiv:2510.25779.

El Hadi Caoui. 2026. [Estimating consumer preferences for llms: Evidence from Imarena](#). SSRN working paper 5970054.

Ramón Christen, Luca Mazzola, Alexander Denzler, and Edy Portmann. 2020. [Exogenous data for load forecasting: A review](#). In *Proceedings of the 12th International Joint Conference on Computational Intelligence (IJCCI)*, pages 489–500.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Robert M. Groves, Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*, 2 edition. Wiley Series in Survey Methodology. John Wiley & Sons, Hoboken, NJ.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. [Large language models are zero-shot time series forecasters](#). In *Advances in Neural Information Processing Systems*.

Valeria Jannelli, Stefan Schöpf, Matthias Bickel, Torbjørn Netland, and Alexandra Brintrup. 2025. [Agentic llms in the supply chain: towards autonomous multi-agent consensus-seeking](#). *International Journal of Production Research*.

Yinan Li, Ying Liu, and Muran Yu. 2025. [Consumer segmentation with large language models](#). *Journal of Retailing and Consumer Services*, 82:104078.

Haixin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B. Aditya Prakash. 2024. [LST-Prompt: Large language models as zero-shot time series forecasters by long-short-term prompting](#). *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7832–7840.

Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. 2025. [Llms reproduce human purchase intent via semantic similarity elicitation of likert ratings](#). *Preprint*, arXiv:2510.08338.

Daniel McFadden. 1974. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

- Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulation of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.
- Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.
- Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. 2018. [Relief-based feature selection: Introduction and review](#). *Journal of Biomedical Informatics*, 85:189–203.
- Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.
- Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. 2025. [The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets](#). In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 16–16, Suzhou, China. Association for Computational Linguistics.

Ethics Statement

This work uses aggregated retail sales data and publicly available survey, crowdfunding, and financial statistics solely for research. All inferences are population-level; no real individual is profiled. Participation in the user study was voluntary and with informed consent; participant feedback and demographics are reported only in anonymized form.

A Appendix

A.1 Full EBI Framework Categories and Data Sources

Taiwan Health Supplements

- [E] Temperature; epidemic index; holiday flags
- [B] Historical weekly sales; Google Trends (flu / cold / immunity / competitor); promotion activity
- [I] Monthly sales tertile (Low / Mid / High)
- Sources: PharmaDemand POS; Open-Meteo (open-meteo.com); Google Trends (trends.google.com); NCU RCTED (rcted.ncu.edu.tw)

Japan 2026 General Election

- [E] Cabinet approval; preferred election outcome; dissolution stance; coalition expectation; policy priority; consumption tax stance; unemployment rate
- [B] Party support distribution; election interest; voting tendency
- [I] 7-party proportional vote share
- Sources: Nishinippon Shimbun (nishinippon.co.jp/visual/243); MIC Labor Force Survey (stat.go.jp/data/roudou); official election results (soumu.go.jp)

US Kickstarter Crowdfunding Technology

- [E] Active campaign count; median funding goal; avg campaign duration; S&P 500 prior-month return; seasonal context
- [B] Avg early backers; % campaigns above 50% funded; avg pledged ratio; Google Trends “kickstarter”
- [I] Monthly campaign success rate
- Sources: WebRobots.io (webrobots.io/kickstarter-datasets); Yahoo Finance (finance.yahoo.com); Google Trends (trends.google.com)

A.2 Full Experimental Results

Table 5 expands the main experimental results (Table 2) with per-model JSD, Spearman’s ρ , and

MAE for EBI and every baseline across the three domains.

A.3 Significance Testing

Main results significance testing. Each baseline is compared against Best EBI: exact permutation test for US Kickstarter Crowdfunding Technology; bootstrap CI ($B=10,000$) otherwise.

Ablation significance testing. Each ablated condition is compared against Full EBI: paired bootstrap over test months ($B=10,000$) for Taiwan Health Supplements and US Kickstarter Crowdfunding Technology; multinomial ground-truth resampling ($B=10,000$) for Japan 2026 General Election. Feature-level conditions are not significance-tested except where noted.

Type	Approach	JSD↓	ρ ↑	MAE↓
<i>Taiwan Health Supplements</i> ($k=3$ few-shot, $n=3$ months; MAE on the ordinal 0/1/2 tertile scale)				
EBI	gpt-4.1-2025-04-14	0.306	0.866	0.333
	claude-haiku-4-5-20251001	0.263	1.000	0.000
	claude-sonnet-4-5-20250929	0.166	1.000	0.000
Statistical	OLS	0.333	—	0.667
	SARIMA	1.000	—	1.000
ML	XGBoost	1.000	—	1.000
	LSTM	0.333	—	0.667
SOTA	LSTPrompt	0.662	—	1.000
Naïve	Carry Forward	1.000	−0.866	1.667
	MA (3-month)	0.639	−0.500	1.333
<i>Japan 2026 General Election</i> ($k=3$ few-shot, $n=1$ election, 7 parties)				
EBI	gpt-4.1-2025-04-14	0.006065	1.0000	0.016446
	claude-haiku-4-5-20251001	0.011080	1.0000	0.026536
	claude-sonnet-4-5-20250929	0.021427	0.9643	0.033679
Statistical	OLS	0.090407	0.9550	0.057547
	SARIMA	0.093092	0.9550	0.061276
ML	XGBoost	0.062081	0.9286	0.058977
	LSTM	0.066837	0.8929	0.066754
SOTA (RSS)	gpt-4.1-2025-04-14	0.047438	0.6786	0.056900
	claude-haiku-4-5-20251001	0.052440	0.8571	0.063386
	claude-sonnet-4-5-20250929	0.049890	0.8571	0.061957
Naïve	Carry Forward	0.067399	0.8929	0.059039
	MA (3-election)	0.090047	0.8571	0.064520
Upper Bound	Survey	0.006111	1.0000	0.019671
<i>US Kickstarter Crowdfunding Technology</i> ($k=3$ few-shot, $n=8$ months)				
EBI	gpt-4.1-2025-04-14	0.000735	0.8623	0.015663
	claude-haiku-4-5-20251001	0.001926	0.6826	0.027037
	claude-sonnet-4-5-20250929	0.000751	0.8095	0.016637
Statistical	OLS	0.001999	−0.4048	0.026788
	SARIMA	0.002609	—	0.032060
ML	XGBoost	0.002269	0.5714	0.026865
	LSTM	0.002156	—	0.028932
SOTA	LSTPrompt	0.002735	0.0714	0.033920
Naïve	Carry Forward	0.002150	0.3810	0.025563
	MA (3-month)	0.002774	−0.3333	0.032917

Table 5: Per-model results across three domains. **Bold** = best EBI per column per domain. Upper Bound Survey included for reference only. — in the ρ column: the method predicts a constant or near-constant class, so ρ is undefined.