Next-Item Prediction with LLMs: Measuring Latent Preferences and Household State from Purchase Histories

Massimiliano Luca [*] Ciro Beneduce [†] Simone Centellegher [‡]
Bruno Lepri [§] Erik Brynjolfsson[¶] Sophia Kazinnik[‖]

January 2026

Abstract

We study whether large language models (LLMs) can extract economically meaningful signals about household preferences, needs, and constraints from high-frequency purchase histories, and whether these inferred states help forecast what a household buys next. Using 1.85 million Amazon transactions linked to demographics and life events, we frame next-category prediction as a measurement problem: improved prediction implies better recovery of latent household conditions. We compare traditional statistical baselines to LLMs with three configurations: (i) purchase history only; (ii) history plus coarse demographics; and (iii) a structured agent that maintains memory, plans over a short horizon, and retrieves external signals including seasonality and product relationships. Off-the-shelf LLMs substantially outperform sequence-based baselines; demographic context provides further gains; and the agentified system performs best, raising top-1 accuracy from 0.53 to 0.70 and improving especially in sparse, event-driven categories. Accuracy gains align with self-reported life events, suggesting that structured LLM agents provide scalable, interpretable measures of household heterogeneity in evolving needs and constraints.

Keywords: consumption, household behavior, large language models, recommendation systems, sequential prediction, agentic architectures

JEL: D12; D91; C45; C53; C55; L86

[*]mluca@fbk.eu, Bruno Kessler Foundation
[†]cbeneduce@fbk.eu, Bruno Kessler Foundation; University of Trento
[‡]scentellegher@fbk.eu, Bruno Kessler Foundation
[§]lepri@fbk.eu, Bruno Kessler Foundation
[¶]Stanford University, Digital Economy Lab (HAI) and NBER.
[‖]kazinnik@stanford.edu, Stanford University, Digital Economy Lab (HAI).

1 Introduction

Understanding and forecasting household consumption is central to empirical work in economics and finance Deaton (1992); Attanasio and Weber (2010); Browning and Crossley (2001); Friedman (1957); Hall (1978). Consumption is a key state variable for household finance and asset pricing Bryzgalova and Julliard (2019, 2022), and it is also a core outcome for policy evaluation and welfare measurement Dea (2015); Meyer and Sullivan (2003); Attanasio and Pistaferri (2016); Meyer and Sullivan (2011). Yet observed purchases are an incomplete proxy for the latent drivers of consumption Hendel and Nevo (2006); Aguiar and Hurst (2005), which depend on internal state variables and are shaped by dynamic responses to price shocks, constraints, and life-cycle changes Erdem et al. (2003). Preferences, constraints, inventories, and discrete life events are only partially observed, while high-frequency transaction data generate long and heterogeneous purchase histories whose economic content is difficult to summarize in a scalable and interpretable way Einav and Levin (2014); Baker and Kueng (2022).

This paper treats this measurement challenge as a prediction problem. We ask a simple, testable question: can large language models (LLMs), given a household's past purchases, predict what the household buys next? We study *next-category prediction* using the e-commerce purchase dataset of Berke et al. (2024). We harmonize the dataset's original labels into a fixed outcome space of 30 Amazon-style macro categories and evaluate models at each purchase occasion. At each prediction point, we provide a truncated window of the most recent purchases (including product-title information for LLM specifications) and require models to output a ranked list of five candidate categories for the next purchase.

In our first experiment, we show that off-the-shelf LLMs used only at inference time (no fine-tuning) outperform classical sequence baselines in a strict zero-shot setting. We compare simple heuristics, a first-order Markov chain, and a recurrent neural network trained on category sequences. Relative to the strongest trained baseline (RNN), the best non-agentified LLM improves ACC@1 from 0.202 to 0.418 and macro-F1 from 0.183 to 0.402. Adding basic socio-demographic attributes (gender, age bin, and income) yields additional gains of roughly 2–3 percentage points in ACC@1 across models, indicating that LLMs can integrate heterogeneous context to refine predictions without any task-specific training.

However, the same comparison also clarifies the limitations of static prompting. A single-pass prompt compresses recent history into a one-shot prediction without a systematic mechanism for maintaining and updating an evolving household state (e.g., inventories, recurring needs, or temporally localized shocks). Static prompting also provides no explicit way to decide which signals are relevant for a given user at a given time, or to combine individual histories with population-level regularities such as seasonality and cross-category relationships. In short, the bottleneck is not only representation but the ability to update a coherent household state over time and to retrieve the right auxiliary structure when it matters.

Motivated by these limitations, we move from LLMs as predictors" to LLMs as decision systems" by introducing an *agentified* framework for next-category prediction. The key idea is to treat each household as a persistent entity with an evolving state,

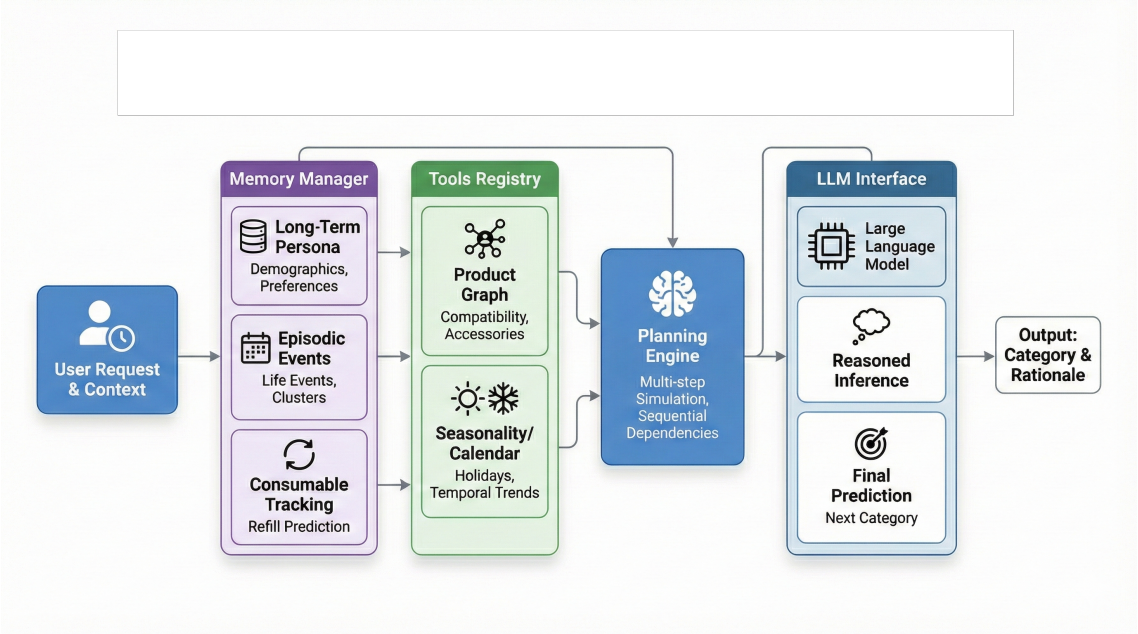


Figure 1: Overview of the agentified LLM framework. The architecture combines structured memory components and task-specific tools to enable iterative reasoning, planning, and dynamic information retrieval for next-category prediction.

and to let the LLM iteratively reason about that state before committing to a ranked prediction. Concretely, the agent interacts with a structured memory subsystem that separates relatively stable user characteristics from episodic signals and replenishment-like cycles, and it can query task-specific tools that expose population-level structure, including temporal regularities (seasonality) and relational product knowledge (cross-category associations). A planning component orchestrates what to retrieve and how to use it, so the model is not forced to rely on a single static prompt. This architecture is designed to exploit cross-household regularities without collapsing behavior into a single pooled process, while preserving household-specific, time-varying state.

Empirically, agentification yields a large additional improvement beyond static LLM prompting across multiple LLM families. The best non-agentified LLM reaches ACC@1 $= 0.418$ and macro-F1$= 0.402$, while the agentified system reaches ACC@1 up to 0.705 and macro-F1 up to 0.686. Mechanism evidence aligns with the design intent: removing tool access yields the largest performance drop, and removing seasonality produces the next-largest degradation, highlighting the role of explicit temporal grounding. Category-level results show broad-based improvements, with the largest relative gains concentrated in sparse, irregular, and preference-driven categories where individual histories provide weak short-run signals. Taken together, the results support a measurement interpretation: if accurate next-category prediction requires inferring latent needs and constraints, then structured LLM agents provide a scalable way to construct empirically useful summaries of household heterogeneity from transaction histories. Figure 1 provides an overview of the agentified architecture and the information sources it integrates.

This paper makes three contributions. First, we frame next-category prediction from purchase histories as a measurement problem in household demand: under a

fixed outcome space of 30 harmonized macro categories and a ranked top-5 evaluation protocol, the task asks how much of households' latent heterogeneity—preferences, needs, and time-varying constraints—can be recovered from observed transactions Kleinberg et al. (2015). Second, we provide a controlled comparison to standard empirical benchmarks for sequential choice (simple heuristics, Markov transitions, and an RNN fit to category sequences), and show that off-the-shelf LLMs used in a zero-shot setting deliver large improvements even when restricted to purchase-history inputs; incorporating basic demographic covariates yields further gains, consistent with these models using additional household context to refine predictions Mullainathan and Spiess (2017). Third, we develop and evaluate an agent-based framework that makes the prediction problem more explicitly economic by separating persistent household characteristics from episodic shocks and replenishment-like cycles, and by allowing the model to incorporate population-level regularities such as seasonality and cross-category complementarities. We show that this structured decision process produces substantial additional gains, and we support the interpretation with ablations and category-level patterns that align with the framework's economic intuition.

2 Data and Methods

2.1 Dataset

We use *Open E-commerce 1.0* (Berke et al., 2024), a consent-based dataset that links item-level Amazon purchase histories to a companion survey for U.S. participants. The purchase data span 2018–2022 and include over 1.8 million transactions contributed by 5,027 users. Each transaction record reports the order date, product title, product identifier (ASIN/ISBN), product category, quantity, purchase price, and the shipping-address state. Participants provided these data by exporting an Amazon order history report and, conditional on affirmative consent, uploading a scrubbed version that retains only non-identifying fields.

A key advantage of *Open E-commerce 1.0* is the structured survey information that accompanies the transaction histories. While the export was processing, participants answered questions on demographics (gender, age range, educational attainment, household income, race and ethnicity, sexual orientation, and state of residence), Amazon usage and household context (account sharing, perceived household size, and typical ordering frequency), and a small set of self-reported life events in 2021 (moving residence, job loss, pregnancy, childbirth, or divorce). The survey also includes additional items related to substance use and health conditions. This design enables structured analysis of heterogeneity in consumption behavior and of responses to salient shocks.

Our empirical analysis focuses on purchase sequences that can be cleanly attributed to an individual decision-maker. We therefore restrict attention to respondents who report that their Amazon account is not shared with others. We further focus on respondents who report gender as either "male" or "female" (97.5% of the dataset) to ensure stability in subgroup comparisons and interpretability of gender-based heterogeneity. We acknowledge that this restriction excludes other gender identities and limits the inclusivity of our conclusions. After applying these

Order date		Qty	State	Product title	Category	User ID
2018-01-21	23.1	1	OK	OtterBox Symmetry case for iPhone 8/7 Plus (FFP) ...	Cellular phone case	R_2zARigF
2018-02-06	16.3	1	OK	*Strength in Stillness: The Power of TM Meditation*	ABIS book	R_2zARigF
2018-04-03	6.4	1	OK	Square Reader (magstripe + headset jack)	Memory card reader	R_2zARigF
2018-06-11	5.3	1	OK	Dove Advanced Care antiperspirant ...	Body deodorant	R_2zARigF

Table 1: Illustrative transaction records from the historical purchases data. User ID is an anonymized respondent identifier.

filters, the analysis sample contains 3,449 users (1,781 female and 1,668 male), with an average of 272.05 historical purchases per user. Examples of raw purchase rows are shown in Table 1.

Although our filtered sample is small relative to the scale of e-commerce activity, *Open E-commerce 1.0* is attractive for its documentation and measurement quality. It is a peer-reviewed release with transparent collection procedures, linked survey metadata, and validation exercises reported by the dataset authors. These features make it a useful setting for evaluating how much observed heterogeneity in demographics and household context helps explain and anticipate subsequent purchase behavior.

Finally, to define a consistent outcome space for next-category analysis and to facilitate comparisons across specifications, we harmonize the dataset's original product categories into a fixed set of 30 official Amazon macro categories. This crosswalk was constructed manually by the research team, yielding a standardized taxonomy used throughout.

2.2 Baselines

We benchmark LLM predictions against simple sequence-based methods that use only each user's observed purchase history. These baseline models abstract from additional covariates (e.g., demographics or life events) and from external product knowledge. In Experiment 1, we use them to ask a narrow question: holding information constant and using a zero-shot prompt, can an LLM match standard sequence heuristics?

We consider four baselines. The **Most Frequent (Mode)** rule predicts, for each user, the category that appears most often in their past purchases. The **Repeat Last** rule predicts the category of the user's most recent purchase. These naive heuristics serve as lower bounds on predictive performance. The **Markov Chain (MC)** baseline estimates a first-order transition matrix on the training sample and predicts the modal next category conditional on the current category. Finally, we

train a **Recurrent Neural Network (RNN)** on sequences of purchase categories; the model takes category embeddings as inputs and returns a probability distribution over the 30 categories. The RNN is trained using cross-entropy loss with fixed hyperparameters and standard regularization; it operates solely on category labels and does not incorporate product text or covariates.

For the trained baselines (MC and RNN), each run randomly assigns 70% of users to the training set and uses their full purchase histories to estimate model parameters; the remaining 30% form the test set. We evaluate all methods on the same test users, including the heuristic rules and the LLMs. Reported results average over 3 independent train–test splits.

2.3 Large Language Models

We use large language models as prediction rules for next-category purchase behavior. Our main specifications use Claude Sonnet 3.5 Anthropic (2024), GPT-4.1 OpenAI (2024), and LLaMA 3.3 70B Meta AI (2024). In all cases, models are used off-the-shelf (inference only) and are not fine-tuned on the study sample.

In the first stage, each model receives a structured text summary of an individual's purchase history and is asked to predict the next purchase category, choosing exactly one of the 30 categories. For long purchase histories, inputs are truncated to fit within the model's token limit. We consider two information sets: (i) purchase history only, which matches the inputs available to the baselines; and (ii) purchase history plus socio-demographic attributes. Demographic information is included in a templated natural-language format (e.g., "The user is a 35–44 year old female with a household income of $75–100k..."). The resulting predictions are evaluated on the same held-out users and compared directly to the baseline methods.

In the second stage, we evaluate an *agentic* variant in which the model is allowed to call purpose-built tools during prediction. These tools implement simple retrieval and transformation functions—: for example, retrieving similar users, summarizing personal purchase cycles, or extracting shocks from the user's event history. This setting is intended to capture a more "assisted" prediction process: the model can iteratively select which information to emphasize before producing a single category prediction. We compare performance in this setting to both the baselines and the direct (non-agentic) LLM predictions to assess the incremental value of tool-assisted inference.

3 Prediction Task and Evaluation Metrics

3.1 Problem Definition

We study a simple forecasting problem in individual-level consumption: given a short window of a user's recent Amazon purchases, predict the *category* of the next purchase. Each transaction is mapped into one of $C = 30$ Amazon macro categories (Section 2.1), so the outcome space is fixed and common across all methods.

Unit of observation. An observation (evaluation instance) corresponds to a user i at a purchase occasion t for which at least K prior purchases are observed. The realized outcome is

$$y_{i,t+1} \in \{1, \ldots, C\},$$

the macro category of the next purchase.

Information set at prediction time. Let $H_{i,t}$ denote the user's most recent K purchases available at time t:

$$H_{i,t} = (y_{i,t-K+1}, \ldots, y_{i,t}), \qquad K = 20.$$

In specifications that incorporate survey information, we augment the history with a time-invariant demographic profile D_i. We write the predictor's information set as

$$x_{i,t} = \begin{cases} H_{i,t}, & \text{(history-only)} \\ (H_{i,t}, D_i), & \text{(history + demographics).} \end{cases}$$

For the statistical baselines, $H_{i,t}$ is represented as a sequence of category tokens. For LLM-based specifications, the same information is rendered into a structured text prompt that includes the product titles alongside their mapped macro categories.

Predictions as ranked choices. Because multiple categories may be plausible given limited history, we evaluate models on their ability to rank likely next categories. Each method must choose from the same set of $C = 30$ categories and return five ranked predictions:

$$\hat{\mathbf{y}}_{i,t} = (\hat{y}_{i,t,1}, \hat{y}_{i,t,2}, \hat{y}_{i,t,3}, \hat{y}_{i,t,4}, \hat{y}_{i,t,5}),$$

where $\hat{y}_{i,t,1}$ is the top-ranked category. To simplify notation in what follows, we index evaluation instances by $n \in \{1, \ldots, N\}$ and write (x_n, y_n) for $(x_{i,t}, y_{i,t+1})$.

3.2 Evaluation Metrics

Let $\hat{\mathbf{y}}_n$ denote the ranked list of five predicted categories for instance n.

Top-k accuracy (ACC@k). Top-k accuracy measures whether the realized next category appears among the model's first k ranked predictions:

$$\text{ACC@}k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[y_n \in \{\hat{y}_{n,1}, \ldots, \hat{y}_{n,k}\}], \qquad k \in \{1, 3, 5\}.$$

ACC@1 is strict "hit rate" for the single best guess, while ACC@3 and ACC@5 capture whether the model assigns high rank to the realized category even when it does not place it first.

Macro-F1 (top-1). Because some categories are much more common than others, accuracy can be dominated by performance on high-frequency categories. To summarize performance more evenly across the category distribution, we also compute the F1 score using the *top-1* prediction $\hat{y}_{n,1}$ as the model's single predicted label.

For each category $c \in \{1, \ldots, C\}$, define true positives, false positives, and false negatives:

$$\mathrm{TP}_c = \sum_{n=1}^{N} \mathbb{I}[y_n = c \wedge \hat{y}_{n,1} = c], \quad \mathrm{FP}_c = \sum_{n=1}^{N} \mathbb{I}[y_n \neq c \wedge \hat{y}_{n,1} = c], \quad \mathrm{FN}_c = \sum_{n=1}^{N} \mathbb{I}[y_n = c \wedge \hat{y}_{n,1} \neq c].$$

Precision and recall are

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c},$$

and the class-wise F1 is

$$\mathrm{F1}_c = \frac{2P_cR_c}{P_c + R_c}.$$

We report *macro-F1*, which weights each category equally:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c.$$

Macro-F1 is therefore more sensitive to performance on less frequent categories and complements the top-k accuracy measures.

4 Non-Agentic Next-Category Prediction

This experiment asks whether off-the-shelf LLMs can forecast an individual's *next purchase category* using only the information contained in that individual's observed purchase history (Radford et al., 2021; Bubeck et al., 2023). We evaluate the models in a *static* (non-agentic) and *zero-shot* setting: the LLM receives a structured summary of past purchases and returns a ranked list of candidate categories, without any task-specific training, fine-tuning, or tool use. We compare these predictions to standard sequence-based baselines and evaluate all methods on the same held-out users using the metrics in Section 3.

Table 2 summarizes performance for two information sets. Under *History Only*, each method conditions on the same $K = 20$ most recent purchases. Under *History + Demographics*, we augment the LLM prompt with a small set of survey covariates (gender, age bin, and income). The baseline methods are shown only in the History Only panel because they are defined to use purchase histories alone.

Baselines. The two heuristic benchmarks (Most Frequent and Repeat Last) perform poorly, indicating that neither long-run average category shares nor simple persistence capture much of the variation in next-category choices. More structured sequence models perform substantially better. The first-order Markov Chain and the RNN deliver similar accuracy across all metrics, suggesting that in this sample the incremental gains from a more flexible sequence model are limited, potentially due to the modest number of users available for training.

Panel A. History only

Model	ACC@1	ACC@3	ACC@5	Macro-F1
Baselines (sequence only)				
Most Frequent (Mode)	0.062±0.028	0.062±0.028	0.062±0.028	0.041±0.019
Repeat Last	0.033±0.015	0.089±0.024	0.128±0.031	0.028±0.012
Markov Chain (1st order)	0.198±0.037	0.312±0.041	0.387±0.044	0.176±0.033
RNN (sequence model)	0.202±0.018	0.328±0.023	0.401±0.026	0.183±0.016
LLMs (zero-shot)				
LLaMA 3.3 70B	0.384±0.079	0.521±0.068	0.598±0.061	0.362±0.074
Claude Sonnet 3.5	0.391±0.056	0.534±0.048	0.611±0.043	0.374±0.052
GPT-4.1	0.387±0.062	0.528±0.054	0.605±0.049	0.368±0.058

Panel B. History + demographics (LLMs only)

Model	ACC@1	ACC@3	ACC@5	Macro-F1
LLaMA 3.3 70B	0.408±0.026	0.547±0.022	0.623±0.019	0.389±0.024
Claude Sonnet 3.5	0.415±0.031	0.558±0.027	0.634±0.024	0.399±0.029
GPT-4.1	0.418±0.039	0.561±0.034	0.637±0.031	0.402±0.037

Table 2: **Non-agentic next-category prediction.** Baselines use purchase histories only. LLMs are evaluated zero-shot using either purchase history only or purchase history plus demographics (gender, age bin, income). Entries report mean±standard deviation across runs.

LLMs with purchase history only. When restricted to the same information set as the baselines (History Only), all three LLMs substantially outperform the strongest baseline. Relative to the RNN/Markov benchmarks, LLMs achieve large gains in ACC@1 and also improve ranked accuracy (ACC@3 and ACC@5) and macro-F1. In economic terms, the LLMs appear to extract additional structure from the *content* of purchases (via product titles and implied semantics) (Sun et al., 2022) that is not available to category-only sequence models, even though no task-specific estimation is performed.

Adding demographics. Providing a small set of demographic covariates further improves LLM performance. Across models, ACC@1 rises by roughly 0.02–0.03 when moving from History Only to History + Demographics, with commensurate improvements in ACC@5 and macro-F1 (largest for GPT-4.1). These gains suggest that even coarse socio-demographic information helps refine the mapping from a given purchase history to likely next-category choices, consistent with persistent heterogeneity in shopping patterns across individuals (Chiang et al., 2023).

5 Towards an Agentified LLM Framework

The results of the previous experiment indicate that Large Language Models (LLMs) can already serve as strong next-category predictors when used in a non-agentified, zero-shot setting. However, in that configuration, LLMs operate as static inference

engines: they receive a fixed prompt summarizing user information and produce a single-step prediction without the ability to explicitly reason about which information is most relevant, when to retrieve it, or how to combine heterogeneous signals. To address these limitations, we move to an agentified formulation that explicitly leverages the planning and reasoning capabilities of LLMs and allows them to interact dynamically with external memory components and task-specific tools.

The motivation behind this design choice is that next-category prediction is inherently a contextual and sequential decision-making problem. Purchasing behavior may depend on long-term user characteristics, recurring consumption patterns, recent behavioral changes, and salient life events, whose relevance varies across individuals and time. An agentified framework enables LLMs to reason iteratively about the prediction task, selectively retrieve information, and adapt the inference process to the specific context of each user.

Figure 1 presents an overview of the proposed agentic architecture. At its core, the framework is structured around two main components: a memory subsystem, which stores and organizes user-related information, and a set of domain-specific tools, which provide access to aggregated behavioral patterns and relational product knowledge. A planning engine orchestrates the interaction between these components, allowing the LLM to decompose the prediction task into intermediate reasoning steps before producing a final output.

The memory subsystem is composed of three complementary modules. The *Long-Term Persona* memory captures relatively stable user characteristics and preferences. Specifically, it stores vector embeddings representing all demographic attributes and preference-related information collected through the survey, such as age bin, gender, income level, and lifestyle indicators (e.g., smoking habits). These embeddings are stored in a queryable vector database, enabling similarity-based retrieval of users with comparable long-term profiles. This component allows the agent to ground its predictions in persistent individual traits rather than relying solely on recent behavior.

The *Episodic Events* memory focuses on temporally localized signals that may induce deviations from habitual purchasing patterns. It contains vector embeddings of episodic life events reported in the survey, such as changes in household composition, employment status, or other significant personal circumstances. Each event is associated with an explicit temporal component indicating when it occurred. This temporal annotation allows the agent to reason about the recency and potential impact of an event, for instance by assigning greater relevance to recent episodes or by identifying periods of behavioral transition.

The *Consumable Tracking* memory is designed to capture recursive and replenishment-driven purchasing behaviors. It encodes recurring consumption patterns inferred from historical purchase sequences, such as regular refills or cyclic buying habits. This component enables the agent to reason about expected future needs based on observed periodicity, complementing both long-term preferences and short-term events.

In parallel, the agent can access a set of task-specific tools that provide structured, population-level information. The *Product Graph* represents relationships between items as a weighted graph, where nodes correspond to products and edges encode

co-occurrence strength. An edge exists between two products if they have been purchased together within the same cart or sequentially across transactions, with edge weights reflecting the frequency of such occurrences. This structure allows the agent to reason about product compatibility, complementarities, and typical purchase transitions beyond individual-level histories.

The *Seasonality* tool captures temporal regularities at the product level. For each item, it stores statistics describing how often the product exhibits periodic purchasing behavior, such as being bought at regular intervals (e.g., every 30 days), as well as its association with specific temporal contexts, including holidays, seasons, or narrow time windows (e.g., within 15 days of a major holiday or during winter months). By querying this tool, the agent can incorporate temporal expectations into its reasoning process and adjust predictions according to calendar effects.

A central planning engine mediates access to memory and tools. Rather than consuming all available information at once, the LLM can iteratively decide which components to query and in which order. Although the internal reasoning steps are not directly observable, it is plausible that the agent alternates between forming hypotheses about the next likely purchase category, retrieving supporting evidence from memory and tools, and refining its prediction accordingly. This process culminates in a final prediction step, where the LLM outputs the predicted next category along with a rationale that reflects the information considered during reasoning.

When contrasted with non-agentified experiment, the benefits of the agentified formulation become particularly evident. In the non-agentified setting, even the strongest LLM configuration achieved ACC@1 of 0.391, despite already outperforming traditional baselines. In agentified experiment, ACC@1 increases to values between 0.681 and 0.705 across models, representing an absolute improvement of 28 percentage points. Similar gains are observed for ACC@3, ACC@5, and F1, with F1 scores from 0.374 to values close to 0.686. These improvements are substantially larger than the differences observed across LLM architectures within the same experimental setting, suggesting that the architectural shift toward agentic reasoning and structured information access plays a more decisive role than model choice alone.

Table 3 reports the quantitative results for the agentified LLMs. GPT-4.1 achieves the highest overall performance, followed closely by Claude Sonnet 3.5, while LLaMA 3.3 70B remains highly competitive despite being an open-weight model. The consistently high performance across models indicates that the proposed framework effectively unlocks the latent reasoning capabilities of LLMs, enabling them to integrate long-term traits, episodic signals, recurring consumption patterns, and population-level behavioral regularities into a unified prediction process.

Overall, these results demonstrate that agentification fundamentally changes how LLMs operate in recommendation tasks. By enabling iterative reasoning, planning, and targeted information retrieval, the proposed framework substantially outperforms both traditional baselines and non-agentified LLM predictors, highlighting the importance of structured agentic designs for complex, real-world sequential decision problems.

Model	ACC@1	ACC@3	ACC@5	F1
LLaMA 3.3 70B	0.681 ± 0.016	0.703 ± 0.014	0.715 ± 0.012	0.658 ± 0.015
Claude Sonnet 3.5	0.690 ± 0.021	0.710 ± 0.018	0.719 ± 0.016	0.669 ± 0.019
GPT-4.1	0.705 ± 0.014	0.716 ± 0.012	0.720 ± 0.011	0.686 ± 0.013

Table 3: Performance of agentified LLMs on next-category prediction. Results are reported as mean ± standard deviation over multiple runs. Compared to the non-agentified setting, the agentified framework yields large and consistent improvements across all metrics, highlighting the impact of structured memory, tool-based information access, and iterative reasoning.

5.1 Ablation Study: Contribution of Individual Tools

To better understand *why* the agentified framework improves performance, we conduct an ablation study aimed at isolating the contribution of the tool components. Concretely, we start from the standard agentified configuration (reported in Table 3) and selectively disable specific tools, keeping the rest of the pipeline unchanged. Results are reported in Table 4, focusing on ACC@3, which provides a stable view of ranking quality and is less brittle than top-1 accuracy in multi-class settings.

Across all three LLMs, the ablation results are remarkably consistent and reveal a clear hierarchy in the importance of the tools. First, removing *all* tools ("w/o tools") yields the largest degradation in performance for every model. For LLaMA 3.3 70B, ACC@3 decreases from 0.703 to 0.641; for Claude Sonnet 3.5, from 0.710 to 0.648; and for GPT-4.1, from 0.716 to 0.654. The absolute drop is therefore approximately 0.06 across models, substantially larger than the observed run-to-run variability. This indicates that the performance gains of the agentified framework are not merely due to prompt restructuring or additional reasoning steps, but are strongly tied to the availability of external, structured information accessed through the tools.

Removing the *Seasonality* tool produces the second-largest performance drop, again consistently across all models. ACC@3 decreases by approximately 0.05 for each LLM, highlighting the importance of explicitly modeling temporal regularities in purchase behavior. This result aligns well with the nature of the task, as many product categories exhibit strong periodicity or are closely associated with specific calendar windows, such as holidays or seasonal consumption cycles. Without explicit seasonality signals, the LLM must infer such patterns indirectly from sparse and noisy individual histories, which appears to be considerably less effective.

The *Product Graph* tool also contributes positively, though to a lesser extent. Disabling it leads to a smaller but still systematic decrease in ACC@3, on the order of 0.03–0.04 across models. The product graph encodes relational information derived from co-occurrence within carts and sequential purchasing behavior, enabling the agent to reason about product complementarities and typical transitions. The more moderate impact of this ablation may reflect partial redundancy with information already present in individual purchase histories, or the fact that category-level prediction is a coarser task where multiple related products map to the same category.

An interesting pattern emerges when comparing the "w/o tools" configuration with the single-tool ablations. While seasonality appears to be the most influential

Table 4: Ablation study on the contribution of individual tools within the agentified framework. Results are reported in terms of ACC@3 (mean ± standard deviation). "Standard" refers to the full agentified configuration; "w/o seasonality", "w/o product graph", and "w/o tools" denote configurations where the corresponding components are disabled.

Model	Configuration	ACC@3
LLaMA 3.3 70B	Standard	0.703 ± 0.014
	w/o seasonality	0.654 ± 0.016
	w/o tools	0.641 ± 0.015
	w/o product graph	0.668 ± 0.015
Claude Sonnet 3.5	Standard	0.710 ± 0.018
	w/o seasonality	0.662 ± 0.020
	w/o tools	0.648 ± 0.019
	w/o product graph	0.676 ± 0.019
GPT-4.1	Standard	0.716 ± 0.012
	w/o seasonality	0.668 ± 0.014
	w/o tools	0.654 ± 0.013
	w/o product graph	0.683 ± 0.013

individual tool, the complete removal of tools consistently results in the worst performance. This suggests that the agent benefits from combining multiple sources of structured evidence, and that the tools provide complementary rather than interchangeable information. The agentified framework therefore appears to exploit a compositional effect, where temporal signals, relational product knowledge, and other retrieved evidence jointly contribute to more accurate ranking of candidate categories.

Importantly, the effects observed in Table 4 are highly stable across LLaMA 3.3 70B, Claude Sonnet 3.5, and GPT-4.1. This cross-model consistency reinforces the conclusion that the observed improvements are primarily driven by the agentic design and tool-based retrieval mechanisms, rather than by idiosyncrasies of a specific LLM. Overall, the ablation study provides strong evidence that explicit access to structured tools—particularly those capturing temporal dynamics and population-level co-occurrence patterns—is a key factor underlying the strong performance of the agentified LLM framework.

5.2 Category-Level Analysis of Agentification Effects

To further understand the impact of the agentified framework, we analyze performance at the category level, comparing standard (non-agentified) LLM predictions against their agentified counterparts. Tables 5-7 report category-wise performance for all three models, together with relative improvements achieved through agentification. This analysis allows us to move beyond aggregate metrics and investigate which types of product categories benefit most from structured reasoning, memory access,

and tool-based retrieval.

A first striking observation is that *all categories benefit from agentification*, across all three LLMs, without exceptions. This consistency suggests that the agentified framework does not merely improve performance on a narrow subset of categories, but provides systematic advantages across the entire prediction space. Importantly, the relative improvements are also highly consistent across models, indicating that the effects are largely model-agnostic and driven by the framework rather than by idiosyncratic model behavior.

Categories characterized by *low baseline predictability* exhibit the largest relative gains. For instance, categories such as *Fine Art*, *Collectible Coins*, *Sports Collectibles*, and *Entertainment Collectibles* show relative improvements ranging from approximately 47% to over 60%. These categories are typically associated with sparse purchase histories, irregular buying patterns, and strong dependence on individual preferences or episodic events. In such settings, static sequence-based inference is particularly brittle, whereas the agentified framework can leverage profile similarity, episodic signals, and relational product knowledge to compensate for limited historical evidence. The magnitude of the improvements suggests that agentification is especially effective in reducing uncertainty where data sparsity is most severe.

Conversely, categories with *high baseline performance*, such as *Books*, *Beauty*, *Consumer Electronics*, and *Personal Computers*, exhibit more moderate relative improvements, typically in the range of 22% to 27%. While smaller in relative terms, these gains remain substantial in absolute terms and indicate that even well-structured, frequently purchased categories benefit from agentic reasoning. A plausible explanation is that these categories already contain strong sequential and semantic signals in the raw purchase histories, limiting the headroom for improvement, yet still benefiting from temporal grounding and relational constraints introduced by the agentic tools.

Several mid-frequency categories, including *Home & Garden*, *Office Products*, *Pet Supplies*, and *Grocery & Gourmet Foods*, display consistent relative improvements around 27–30%. These categories often combine recurring consumption patterns with moderate variability driven by seasonality or household composition. The observed gains align well with the design of the agentified framework, particularly the consumable tracking memory and the seasonality tool, which explicitly model replenishment cycles and calendar-driven demand.

Technology-related categories such as *Cell Phones & Accessories*, *Video Games*, and *Consumer Electronics* show stable and homogeneous improvements across models. This suggests that the product graph plays an important role in these domains by capturing co-occurrence and sequential purchase relationships, such as accessories following device purchases. The relatively uniform gains across models further reinforce the interpretation that relational product knowledge provides a robust and transferable signal.

An additional noteworthy pattern is the *remarkable consistency of relative improvements across LLMs*. For nearly all categories, differences in relative gains between LLaMA 3.3 70B, Claude Sonnet 3.5, and GPT-4.1 remain within a few percentage points. This observation strengthens the conclusion that the agentified framework systematically enhances the decision process itself, rather than exploiting

Category	LLaMA 3.3 70B	Claude Sonnet 3.5	GPT-4.1
Grocery	Std: 0.567 (0.032) Ag.: 0.723 (0.024) %: 27.5	Std: 0.579 (0.035) Ag.: 0.734 (0.027) %: 26.8	Std: 0.585 (0.030) Ag.: 0.740 (0.023) %: 26.5
Health & Care	Std: 0.601 (0.030) Ag.: 0.758 (0.023) %: 26.1	Std: 0.613 (0.033) Ag.: 0.770 (0.026) %: 25.6	Std: 0.619 (0.028) Ag.: 0.776 (0.022) %: 25.4
Beauty	Std: 0.589 (0.031) Ag.: 0.745 (0.023) %: 26.5	Std: 0.601 (0.034) Ag.: 0.755 (0.026) %: 25.6	Std: 0.607 (0.029) Ag.: 0.761 (0.022) %: 25.4
Pet Supplies	Std: 0.556 (0.032) Ag.: 0.712 (0.025) %: 28.1	Std: 0.568 (0.035) Ag.: 0.724 (0.028) %: 27.5	Std: 0.574 (0.030) Ag.: 0.730 (0.024) %: 27.2
Baby Products	Std: 0.512 (0.033) Ag.: 0.678 (0.025) %: 32.4	Std: 0.523 (0.036) Ag.: 0.687 (0.028) %: 31.4	Std: 0.529 (0.031) Ag.: 0.693 (0.024) %: 31.0

Table 5: Category-level comparison (ACC@1): **Replenishment / frequently purchased goods.** Entries report mean (standard deviation). % improvement is relative to the standard (non-agentified) setting.

specific strengths of a single model. In other words, the framework appears to act as a strong inductive bias that reshapes how different LLMs reason about the task.

Overall, the category-level analysis provides important qualitative insights that complement the aggregate results. Agentification appears particularly effective for categories with sparse, irregular, or preference-driven purchase behavior, while still delivering consistent gains in more structured and frequently purchased domains. These findings further support the claim that the proposed framework improves robustness and generalization, enabling LLMs to adapt their reasoning strategy to the specific characteristics of each category rather than relying on a one-size-fits-all inference process.

6 Discussion

The results presented in this work provide a coherent picture of how Large Language Models can be progressively transformed from static sequence predictors into structured, decision-making systems for next-category recommendation. Across all experiments, a clear hierarchy emerges: traditional baselines capture only limited regularities in purchasing behavior; non-agentified LLMs substantially outperform these baselines even in a zero-shot setting; and agentified LLMs achieve a further, dramatic improvement by explicitly leveraging reasoning, planning, and structured information retrieval.

A first important takeaway is that the strong performance of non-agentified LLMs already challenges the assumption that complex recommendation tasks necessarily require task-specific training. When provided with structured representations of purchase histories—and, optionally, socio-demographic information—LLMs are able to infer meaningful behavioral patterns that go beyond what classical sequence models achieve, even when those models are trained directly on the data. This suggests that LLMs encode powerful inductive biases for sequential and contextual reasoning that can generalize effectively to recommendation settings.

Category	LLaMA 3.3 70B	Claude Sonnet 3.5	GPT-4.1
	Std: 0.421 (0.038)	Std: 0.432 (0.041)	Std: 0.438 (0.035)
Automotive	Ag.: 0.589 (0.029)	Ag.: 0.598 (0.032)	Ag.: 0.604 (0.027)
	%: 39.9	%: 38.4	%: 37.9
	Std: 0.445 (0.036)	Std: 0.456 (0.039)	Std: 0.462 (0.033)
Camera & Photo	Ag.: 0.612 (0.028)	Ag.: 0.621 (0.031)	Ag.: 0.627 (0.026)
	%: 37.5	%: 36.2	%: 35.7
	Std: 0.598 (0.030)	Std: 0.610 (0.033)	Std: 0.616 (0.028)
Cell Phones & Acc.	Ag.: 0.756 (0.023)	Ag.: 0.767 (0.026)	Ag.: 0.773 (0.022)
	%: 26.4	%: 25.7	%: 25.5
	Std: 0.621 (0.029)	Std: 0.633 (0.032)	Std: 0.639 (0.027)
Consumer Electronics	Ag.: 0.774 (0.022)	Ag.: 0.786 (0.025)	Ag.: 0.792 (0.021)
	%: 24.6	%: 24.2	%: 23.9
	Std: 0.612 (0.030)	Std: 0.624 (0.033)	Std: 0.630 (0.028)
Personal Computers	Ag.: 0.767 (0.022)	Ag.: 0.779 (0.025)	Ag.: 0.785 (0.021)
	%: 25.3	%: 24.8	%: 24.6
	Std: 0.456 (0.035)	Std: 0.467 (0.038)	Std: 0.473 (0.033)
Major Appliances	Ag.: 0.623 (0.028)	Ag.: 0.634 (0.031)	Ag.: 0.640 (0.026)
	%: 36.6	%: 35.8	%: 35.3
	Std: 0.578 (0.031)	Std: 0.590 (0.034)	Std: 0.596 (0.029)
Home & Garden	Ag.: 0.734 (0.024)	Ag.: 0.746 (0.027)	Ag.: 0.752 (0.023)
	%: 27.0	%: 26.4	%: 26.2
	Std: 0.498 (0.034)	Std: 0.509 (0.037)	Std: 0.515 (0.032)
Tools & Home	Ag.: 0.664 (0.027)	Ag.: 0.675 (0.030)	Ag.: 0.681 (0.025)
	%: 33.3	%: 32.6	%: 32.2
	Std: 0.534 (0.032)	Std: 0.546 (0.035)	Std: 0.552 (0.030)
Office Products	Ag.: 0.698 (0.025)	Ag.: 0.710 (0.028)	Ag.: 0.716 (0.024)
	%: 30.7	%: 30.0	%: 29.7
	Std: 0.501 (0.034)	Std: 0.512 (0.037)	Std: 0.518 (0.032)
Software	Ag.: 0.667 (0.027)	Ag.: 0.678 (0.030)	Ag.: 0.684 (0.025)
	%: 33.1	%: 32.4	%: 32.0
	Std: 0.412 (0.038)	Std: 0.423 (0.041)	Std: 0.429 (0.035)
Industrial & Scientific	Ag.: 0.578 (0.030)	Ag.: 0.589 (0.033)	Ag.: 0.595 (0.028)
	%: 40.3	%: 39.2	%: 38.7

Table 6: Category-level comparison (ACC@1): **Durables / planned purchases (including tech and household).** Entries report mean (standard deviation). % improvement is relative to the standard (non-agentified) setting.

However, the comparison between non agentified and agentified experiments highlights the limitations of static prompting. While non-agentified LLMs outperform baselines, their predictions remain constrained by single-step inference and fixed input representations. The agentified framework fundamentally alters this interaction paradigm. By enabling iterative reasoning and selective access to memory and tools, LLMs are no longer forced to compress all relevant information into a single prompt. Instead, they can actively decide which signals are needed, retrieve them on demand, and integrate them into a coherent prediction process.

The ablation study provides strong evidence that the performance gains of agentification are not superficial. Removing access to tools results in large and consistent drops in performance across all models, indicating that structured external information plays a central role. Among the tools, seasonality emerges as the most influential component, underscoring the importance of explicitly modeling temporal regularities in purchasing behavior. At the same time, the product graph contributes complementary relational information that helps constrain predictions based on population-level co-occurrence and sequential patterns. The fact that these effects are stable across different LLM architectures reinforces the interpretation that the gains stem primarily from the agentic design rather than from model-specific quirks.

The category-level analysis further deepens this understanding. Categories characterized by sparse data, irregular purchase behavior, or strong dependence on personal taste—such as Fine Art, Collectibles, and Independent Design—benefit disproportionately from agentification. In these domains, historical sequences alone provide weak signals, and the ability to exploit profile similarity, episodic context, and structured relational knowledge becomes crucial. Conversely, categories with high baseline predictability, such as Books or Consumer Electronics, exhibit smaller relative improvements, suggesting diminishing returns where sequential patterns are already strong. Importantly, however, even these categories consistently benefit from agentification, indicating that the framework improves robustness rather than overfitting to specific domains.

Taken together, these findings suggest that agentification acts as a form of structural regularization. Rather than relying solely on implicit pattern matching, the agentified framework encourages LLMs to ground their predictions in explicit evidence drawn from multiple sources, reducing reliance on spurious correlations and improving generalization across categories and user profiles.

7 Conclusion

In this work, we explored the use of large language models for next-category prediction in e-commerce settings, progressively moving from traditional baselines to non-agentified LLMs and finally to a fully agentified framework. Our results show that while LLMs already provide strong performance in a zero-shot, static setting, their full potential is unlocked when they are embedded within an agentic architecture that supports reasoning, planning, and structured information access.

The proposed framework integrates heterogeneous signals through dedicated memory components and task-specific tools, enabling LLMs to reason about long-term user traits, episodic life events, recurring consumption patterns, temporal regularities, and relational product knowledge. Empirically, this design yields large and consistent improvements over both classical baselines and non-agentified LLMs, with particularly strong gains in challenging, sparse, and preference-driven categories.

Beyond raw performance, an important contribution of this work is methodological. Rather than treating LLMs as drop-in replacements for existing models, we demonstrate the value of designing interaction paradigms that align with their strengths. The results suggest that, for recommendation and other sequential decision problems, the question is not only which model to use, but how to structure the decision process around it.

Several directions for future work naturally emerge. First, a more fine-grained ablation of the memory components—such as isolating the contribution of long-term persona, episodic events, and consumable tracking—would provide deeper insights into how different forms of memory interact with reasoning and tool usage. Second, extending the framework to support learning-based tool selection or adaptive planning strategies could further improve efficiency and interpretability. Third, while this work focuses on category-level prediction, the same framework could be extended to item-level recommendation, bundle generation, or multi-step shopping trajectory forecasting. Finally, investigating human-in-the-loop or explainability-

oriented extensions, where agent reasoning traces are surfaced to users or analysts, represents a promising avenue for building transparent and trustworthy recommender systems.

Overall, this study highlights agentification as a powerful paradigm for applying LLMs to real-world recommendation problems. By moving beyond static prompting and embracing structured reasoning and evidence-driven decision making, agentified LLMs open new possibilities for building more adaptive, robust, and interpretable recommender systems.

Category	LLaMA 3.3 70B	Claude Sonnet 3.5	GPT-4.1
Books	Std: 0.634 (0.029) Ag.: 0.781 (0.021) %: 23.2	Std: 0.647 (0.032) Ag.: 0.792 (0.024) %: 22.4	Std: 0.653 (0.027) Ag.: 0.798 (0.020) %: 22.2
Music and DVD	Std: 0.523 (0.033) Ag.: 0.689 (0.026) %: 31.7	Std: 0.535 (0.036) Ag.: 0.700 (0.029) %: 30.8	Std: 0.541 (0.031) Ag.: 0.706 (0.025) %: 30.5
Video, DVD & Blu-ray	Std: 0.512 (0.033) Ag.: 0.678 (0.026) %: 32.4	Std: 0.523 (0.036) Ag.: 0.689 (0.029) %: 31.7	Std: 0.529 (0.031) Ag.: 0.695 (0.025) %: 31.4
Video Games	Std: 0.578 (0.031) Ag.: 0.734 (0.024) %: 27.0	Std: 0.590 (0.034) Ag.: 0.746 (0.027) %: 26.4	Std: 0.596 (0.029) Ag.: 0.752 (0.023) %: 26.2
Toys & Games	Std: 0.587 (0.031) Ag.: 0.743 (0.023) %: 26.6	Std: 0.599 (0.034) Ag.: 0.755 (0.026) %: 26.0	Std: 0.605 (0.029) Ag.: 0.761 (0.022) %: 25.8
Outdoors	Std: 0.489 (0.034) Ag.: 0.656 (0.027) %: 34.2	Std: 0.501 (0.037) Ag.: 0.668 (0.030) %: 33.3	Std: 0.507 (0.032) Ag.: 0.674 (0.026) %: 32.9
Sports	Std: 0.545 (0.032) Ag.: 0.701 (0.025) %: 28.6	Std: 0.557 (0.035) Ag.: 0.713 (0.028) %: 28.0	Std: 0.563 (0.030) Ag.: 0.719 (0.024) %: 27.7
Musical Instruments	Std: 0.398 (0.039) Ag.: 0.564 (0.032) %: 41.7	Std: 0.409 (0.042) Ag.: 0.575 (0.035) %: 40.6	Std: 0.415 (0.036) Ag.: 0.581 (0.029) %: 40.0
Independent Design	Std: 0.389 (0.040) Ag.: 0.556 (0.033) %: 42.9	Std: 0.401 (0.043) Ag.: 0.567 (0.036) %: 41.4	Std: 0.407 (0.037) Ag.: 0.573 (0.030) %: 40.8
Watches	Std: 0.423 (0.037) Ag.: 0.590 (0.029) %: 39.5	Std: 0.434 (0.040) Ag.: 0.601 (0.032) %: 38.5	Std: 0.440 (0.034) Ag.: 0.607 (0.027) %: 38.0
Collectible Coins	Std: 0.298 (0.045) Ag.: 0.467 (0.037) %: 56.7	Std: 0.309 (0.048) Ag.: 0.476 (0.040) %: 54.0	Std: 0.315 (0.042) Ag.: 0.482 (0.034) %: 53.0
Entertainment Collectibles	Std: 0.334 (0.043) Ag.: 0.501 (0.035) %: 50.0	Std: 0.345 (0.046) Ag.: 0.510 (0.038) %: 47.8	Std: 0.351 (0.040) Ag.: 0.516 (0.032) %: 47.0
Sports Collectibles	Std: 0.312 (0.044) Ag.: 0.478 (0.036) %: 53.2	Std: 0.323 (0.047) Ag.: 0.489 (0.039) %: 51.4	Std: 0.329 (0.041) Ag.: 0.495 (0.033) %: 50.5
Fine Art	Std: 0.276 (0.047) Ag.: 0.445 (0.038) %: 61.2	Std: 0.287 (0.050) Ag.: 0.454 (0.041) %: 58.2	Std: 0.293 (0.044) Ag.: 0.460 (0.035) %: 57.0

Table 7: Category-level comparison (ACC@1): **Leisure / taste goods and collectibles.** Entries report mean (standard deviation). % improvement is relative to the standard (non-agentified) setting.

References

The prize in economic sciences 2015: Advanced information — consumption, poverty and welfare. NobelPrize.org, 2015. URL https://www.nobelprize.org/prizes/economic-sciences/2015/advanced-information/. Accessed 2026-02-02.

Mark Aguiar and Erik Hurst. Consumption versus expenditure. *Journal of Political Economy*, 113(5):919–948, 2005.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com, 2024. Accessed: 2024.

Orazio P. Attanasio and Luigi Pistaferri. Consumption inequality. *Journal of Economic Perspectives*, 30(2):3–28, 2016. doi: 10.1257/jep.30.2.3.

Orazio P. Attanasio and Guglielmo Weber. Consumption and saving: Models of intertemporal allocation and their implications for public policy. *Journal of Economic Literature*, 48(3):693–751, 2010. doi: 10.1257/jel.48.3.693.

Scott R. Baker and Lorenz Kueng. Household financial transaction data. *Annual Review of Economics*, 14:47–67, 2022.

Alex Berke, Dan Calacci, Robert Mahari, Takahiro Yabe, Kent Larson, and Sandy Pentland. Open e-commerce 1.0, five years of crowdsourced us amazon purchase histories with user demographics. *Scientific Data*, 11(1):491, 2024.

Martin Browning and Thomas F. Crossley. The life-cycle model of consumption and saving. *Journal of Economic Perspectives*, 15(3):3–22, 2001. doi: 10.1257/jep.15.3.3.

Svetlana Bryzgalova and Christian Julliard. Consumption in asset returns. Working paper, London Business School, 2019.

Svetlana Bryzgalova and Christian Julliard. Consumption. Working paper, London Business School and London School of Economics, May 2022. URL https://csef.it/wp-content/uploads/Julliard.pdf.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Alex Chiang, Percy Liang, Jakob Macke, et al. Foundational models for decision making with heterogeneous agents. *arXiv preprint arXiv:2310.02639*, 2023.

Angus Deaton. *Understanding Consumption*. Clarendon Press (Oxford University Press), Oxford, 1992.

Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346(6210): 1243089, 2014. doi: 10.1126/science.1243089.

Tülin Erdem, Susumu Imai, and Michael P. Keane. Brand and quantity choice dynamics under price uncertainty. *Quantitative Marketing and Economics*, 1(1):5–64, 2003.

Milton Friedman. *A Theory of the Consumption Function*. Princeton University Press, Princeton, NJ, 1957. NBER General Series No. 63.

Robert E. Hall. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy*, 86(6):971–987, 1978.

Igal Hendel and Aviv Nevo. Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6):1637–1673, 2006. doi: 10.1111/j.1468-0262.2006.00721.x.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–495, 2015.

Meta AI. Llama 3.3: Open foundation and fine-tuned chat models. https://ai.meta.com/llama, 2024. Accessed: 2024.

Bruce D. Meyer and James X. Sullivan. Measuring the well-being of the poor using income and consumption. *Journal of Human Resources*, 38(Supplement):1180–1220, 2003. doi: 10.2307/3558985.

Bruce D. Meyer and James X. Sullivan. Further evidence on measuring the well-being of the poor using income and consumption. *Canadian Journal of Economics*, 44(1):52–87, 2011.

Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

OpenAI. Gpt-4.1 technical overview. https://openai.com, 2024. Accessed: 2024.

Alec Radford, Jong Wook Kim, Luke Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Haoyu Sun, Daniel Lee, Lingjiao Zhan, et al. Demonstration of text-based personalized product recommendations using pretrained transformers. *arXiv preprint arXiv:2205.09897*, 2022.