# LLM Exposure and Funnel Erosion in Consumer Journeys: Measuring Journey Compression and Reordering in Passive Behavioral Logs

**Abstract**

LLM assistants increasingly mediate product discovery, yet most journey analytics and attribution systems still assume that discovery begins with search, retailer browsing, or direct brand visits. That assumption misstates the observed journey. When evaluation moves into conversational interfaces, passive behavioral logs undercount early-stage activity and standard journey reconstruction can resemble funnel erosion. We develop a measurement framework that treats LLM usage as a first-class touchpoint in de-identified passive event logs linked to a post-trace survey. The framework detects observed LLM exposure from domain and app signals, operationalizes a non-linear stage codebook (consideration, intent, conversion-proximate), and computes journey metrics for compression, timing, channel mix, and transition structure without imposing a canonical path. A production-scale, no-lookahead labeling system combines deterministic cues, model-assisted classification, locked holdouts, and targeted adjudication audits to deliver reproducible measurement in noisy passive logs. In a one-month observation window, the most stable empirical result is a shift in visible pre-intent discovery: LLM-exposed journeys show substantially higher observable LLM share and higher review/UGC share, while unadjusted compression and time-to-intent differences shrink after adjustment. Conversational discovery changes what marketing analytics systems observe, and journey measurement must be updated accordingly.

**Keywords:** customer journey analytics; clickstream analysis; marketing attribution; generative AI; conversational search; funnel measurement

## 1 Introduction

Customer-journey analytics depends on observing discovery and evaluation in the order consumers encounter them. Measurement choices shape resource allocation decisions (Iacobucci *et al*, 2019), yet digital analytics can also produce analytics myopia when visible behaviors are mistaken for strategically meaningful stages of the customer decision journey (Vollrath and Villegas, 2022). That risk has intensified as LLM assistants (e.g., ChatGPT, Claude, Gemini, Perplexity, Copilot) become entry points for product research, comparison, and narrowing a choice set. When the first touchpoint is a conversation rather than a query, standard funnel reconstructions can misstate where discovery began and how much evaluation occurred before downstream visits. Analytics dashboards may not simply miss a channel; they may misclassify the strategic meaning of the visible path. What appears to be a shorter or more efficient funnel may instead reflect the migration of early-stage consideration into a less observable interface. This matters because touchpoint identification and optimization frameworks in digital retail depend on accurate observation of sales-relevant interactions (Zimmermann and Auinger, 2023). We therefore treat LLM exposure as an observability-boundary problem for marketing analytics systems, not simply as an AI-use issue.

| **Traditional (search-led) trace** | **LLM-mediated trace** |
|---|---|
| 1. Search engine query for budget ANC headphones | 1. LLM prompt for budget ANC headphones prioritizing comfort |
| 2. Review site(s) and comparison articles | 2. Shortlist generated inside LLM |
| 3. Marketplace category browsing | 3. Direct jumps to a few product pages (PDPs) |
| 4. Multiple PDP views across brands | 4. Price/shipping checks for the shortlist |
| 5. Cart / checkout | 5. Cart / checkout |

Figure 1: Illustrative journeys. When evaluation moves into an LLM interface, passive logs may show fewer early-stage touches and shorter time-to-intent, even if the consumer explored many options conversationally.

Figure 1 contrasts a familiar search-led trace with an LLM-mediated trace for a shopper seeking noise-canceling headphones under $200. In the search-led path, discovery unfolds across queries, reviews, category browsing, product detail pages, and checkout. In the LLM-mediated path, the shopper may begin with a prompt that generates a shortlist and then move directly to a small set of product and price-check pages. Passive logs can therefore show fewer early-stage touches, fewer distinct domains, and shorter time-to-intent even when substantial evaluation occurred inside the conversational interface.

This paper establishes LLM usage as a distinct touchpoint class and shows how conventional funnel reconstruction breaks when conversational discovery is omitted from the trace. Rather than imposing a linear funnel, we analyze path- and time-based signals that capture (i) **compression** (a smaller observable interval from consideration to intent), (ii) **reordering** (more direct observable transitions to intent or conversion-proximate behavior), and (iii) **discovery mix shifts** (visible substitution away from classic search toward conversational agents or direct retailer visits).

These patterns can arise through several concrete mechanisms. LLMs can front-load shortlist formation before any external browsing begins. They can move comparison steps that previously occurred across multiple review sites into one conversational interface. They can also redirect downstream validation toward a smaller set of price, shipping, or retailer pages. Each mechanism simplifies the *visible* journey. The measurement implication is straightforward: journey analytics must account for consideration that occurs inside an interface recorded as a single touchpoint.

The paper makes four measurement contributions:

- **A reproducible LLM detection registry for passive logs.** User-level observed exposure is identified from domain/app signals with first-observed timing.
- **Journey diagnostics that expose hidden consideration.** We operationalize compression, transition structure, and discovery-mix measures without forcing a legacy search-led funnel.
- **A production-scale labeling and audit stack.** The measurement system combines deterministic rules, model-assisted classification, no-lookahead labeling, locked holdouts, and targeted audits to turn noisy multi-source traces into reproducible journey data.
- **A guardrail design for interpretation.** Survey-linked adjustment and the $G0/G1/G2$ decomposition separate user exposure from journey augmentation and prevent over-attribution.

## 2 Background and related work

### 2.1 Funnels, journeys, and non-linearity

Marketing theory continues to inform journey measurement through staged models that move from awareness to action, including hierarchy-of-effects formulations (Lavidge and Steiner, 1961; Wijaya, 2012) and DAGMAR-style managerial objectives (Colley, 1961). Observed decision sequences, however, are rarely linear. Reviews of advertising effects emphasize indirect and context-dependent pathways (Vakratsas and Ambler, 1999), customer journey research conceptualizes experience as distributed across interactions among consumers, firms, partners, and external actors (Lemon and Verhoef, 2016), and practitioner frameworks describe iterative exploration and evaluation rather than one-way funnels (McKinsey and Company, 2009; Rennie and Protheroe, 2020).

For passive trace data, stages are best treated as an operational codebook rather than as a literal behavioral script. Our operational funnel therefore uses three stages—consideration, intent, and conversion-proximate behavior—because awareness is rarely observable in passive clickstreams.

Touchpoint systems must also remain aligned with how customers are actually observed. Conversion rate optimization research formalizes the need to identify sales-impacting touchpoints along the customer journey accurately (Zimmermann and Auinger, 2023), while touchpoint research shows that different touchpoints contribute differently to brand consideration and resource-allocation decisions (Baxendale *et al*, 2015). Broader analytics work also warns against tool-centric measurement that drifts away from decision-journey strategy (Vollrath and Villegas, 2022). We build on that logic by treating LLM usage as a first-class touchpoint and by showing when apparent funnel erosion reflects changed observability rather than changed behavior.

### 2.2 Clickstream-based journey and attribution modeling

Event-level logs are routinely used to reconstruct journeys and estimate how touchpoints relate to conversion. Graph-based attribution modeling shows that adjacency and order carry information about pre-conversion structure (Anderl *et al*, 2016). Multichannel models further demonstrate that paths can be treated as dynamic sequences with measurable effects on downstream outcomes (Li and Kannan, 2014). Touchpoint studies also show that different channels contribute differently to consideration and choice (Baxendale *et al*, 2015).

Our work stays within this tradition but updates a key measurement assumption. Discovery and evaluation may now occur inside an LLM interface that appears in passive logs as a single domain/app event. A conversational assistant can operate simultaneously as information source, comparison interface, recommendation agent, and redirect mechanism. The theoretical issue is therefore not only that LLMs add another touchpoint to the journey, but that they may consolidate several formerly separate touchpoints into one conversational interaction. If that touchpoint is omitted or misclassified, downstream product pages, retailer visits, or price-checking pages may appear to initiate the journey, leading analytics systems to over-credit lower-funnel touchpoints. That shift can change inferred funnel length, timing, and channel contribution even when underlying cognitive effort is unchanged.

### 2.3 Conversational AI and product evaluation

Conversational agents have been shown to influence purchase intention, perceived information value, and engagement in commerce contexts (Lo Presti *et al*, 2021; Le, 2023; Luo *et al*, 2023). Recent work comparing generative-AI chatbots with search engines finds meaningful differences in how consumers evaluate products and disclose information (Kim and Priluck, 2025). These findings

indicate that LLM-mediated discovery should be measured as its own touchpoint class rather than absorbed into search.

The key behavioral link for this paper is not that LLMs eliminate evaluation, but that they can relocate evaluation and change how much of it remains externally visible. A parsimonious consumer-search account is that conversational interfaces can front-load option screening, compress visible exploration, and redirect validation activity toward a smaller set of downstream pages. In decision-aid terms, the technology may change the distribution of visible effort across interfaces rather than simply changing final choices. Because our data do not observe the conversational content itself, we treat these pathways as theory-consistent interpretations of the journey patterns rather than as identified mechanisms.

# 3 Data

## 3.1 Multi-source passive behavioral logs

The empirical setting combines multi-source event logs from a third-party passive measurement panel with survey data linked at the user level (`caseid`). The trace data record event-level touchpoints across web and app contexts, including domain/app bundle, URL fields when available, timestamps, and auxiliary metadata. The survey contributes demographics, media-attitudinal measures (news frequency items `NEWS1a-f`, media consumption/distrust items `CMEDIA1-4`, digital media self-efficacy items `MDSE1-3`, and political/news interest and ideology), and survey weights.

The provider applies de-identification and aggressive PII reduction before delivery. All analyses therefore operate on de-identified traces, and we do not perform additional PII scrubbing beyond the provider's process. The trace window spans **October 4, 2024–November 5, 2024** (based on file timestamps), and the survey field period is **November 21, 2024–December 6, 2024**.

## 3.2 Unit of analysis

The analysis uses two linked units. We construct (i) a user-day panel for first observed exposure timing and within-user comparisons and (ii) journey-level sequences (paths) for transition-structure and timing analyses. A "journey" is defined as an ordered sequence of events within a rolling window, segmented by a fixed inactivity gap and maximum window length; robustness checks vary the gap/cap choices. The unit of statistical inference is the user (`caseid`), and uncertainty is clustered accordingly.

# 4 Measurement and Methods

## 4.1 Detecting LLM exposure

LLM exposure is detected from domain/app bundle signals and light text cues when available. Examples include domains such as `chat.openai.com`, `claude.ai`, `gemini.google.com`, and `perplexity.ai`, and corresponding mobile app bundles. Detection uses only current-event fields (never future context), which preserves a strict no-lookahead rule and avoids label leakage.

We define:

- **User-level observed exposure ($E_i$):** whether user $i$ has any detected LLM touchpoint in the observation window.

- **First observed exposure timing:** the first observed day/week with an LLM touchpoint (first observed, not necessarily first-ever use).
- **LLM intensity (continuous):** LLM touches and/or minutes per week when duration is available.

### 4.2 Exposure taxonomy: user exposure versus journey augmentation

The measurement design separates a broad user-level exposure registry from a narrower journey-level augmentation signal. User-level observed exposure is $E_i \in \{0, 1\}$ as defined above and follows the paper's primary AI/LLM detection registry, which can include canonical LLM domains/apps as well as adjacent conversational surfaces detectable in passive logs. Journey-level augmentation is stricter and is based only on canonical in-journey LLM channel matches before first intent:

$$A_{pre,j} = 1 \text{ iff journey } j \text{ contains at least one LLM touchpoint before first intent.}$$

If a journey has no observed intent event, the full journey is treated as the pre-intent window for this indicator. We enforce hierarchy $A_{pre,j} \leq E_i$ in all reporting artifacts so that journey augmentation cannot be positive for users without observed exposure.

We report a three-group decomposition in Results to separate exposure propensity from journey augmentation:

- $G0$: $E = 0$, $A_{pre} = 0$ (non-exposed users),
- $G1$: $E = 1$, $A_{pre} = 0$ (exposed users, non-augmented journeys),
- $G2$: $E = 1$, $A_{pre} = 1$ (exposed users, pre-intent LLM-augmented journeys).

This decomposition separates user-level exposure propensity from within-journey augmentation and avoids treating all exposed-user journeys as LLM-augmented.

### 4.3 Stage codebook for trace events

Awareness is not reliably observable in clickstreams, so we operationalize a three-stage funnel. Stages are assigned at the event level, and non-linear paths (repeats, backtracking, and out-of-order transitions) are allowed (Table 1).

### 4.4 Journey metrics

We compute several journey-level metrics to represent distinct forms of "funnel erosion" without imposing a canonical path.

**Funnel compression.** Let $N_{j,s}$ be the number of events in journey $j$ labeled as stage $s$. A count-based compression metric is

$$C_j^{(count)} = \frac{N_{j,\text{intent}} + N_{j,\text{conv}}}{N_{j,\text{total}}}. \tag{1}$$

We also compute time-based variants using dwell-time or inter-event time when available.

Table 1: Operational funnel stages and illustrative trace signals.

| Stage | Behavioral signals | Examples in traces |
|---|---|---|
| Consideration | Information seeking, browsing, reviews, comparisons, category exploration | Review articles, "best-of" lists, YouTube reviews, retailer category pages, general brand-site browsing, LLM prompts asking for recommendations |
| Intent | Preference formation, price/shipping checks, store locator, repeated product views, cart additions | "price" or "shipping" pages, "where to buy," cart events, repeated visits to a small set of PDPs, LLM prompts like "compare X vs Y" |
| Conversion-proximate | Checkout/payment flows, order confirmation, account/order tracking, subscription start | Checkout URL patterns, payment domains, order confirmation pages, app events consistent with purchase completion |

*Note:* PDP = product detail page.

**Stage timing.** We compute time-to-event outcomes such as:

$$\Delta t_j^{(\text{intent})} = t_{j,\text{first intent}} - t_{j,\text{first consid}}, \tag{2}$$

and analogous measures for first conversion-proximate behavior, with right-censoring when not observed.

**Discovery mix.** For pre-intent events, we compute touchpoint shares across channel classes: LLM, classic search, retailer/marketplace, direct brand, review/UGC, and other. We also compute any-touch indicators by channel class.

**Transition structure.** We compute stage-to-stage transition frequencies and transition probabilities

$$P(a \to b) = \frac{\#(a \to b)}{\sum_{b'} \#(a \to b')}, \tag{3}$$

including direct consideration→conversion-proximate jumps and looping intensity (e.g., consideration↔intent).

**Consideration breadth.** When entity extraction coverage permits, we measure breadth as the number of unique sources visited before first intent.

### 4.5 Hypotheses

Directional predictions are used only where the measurement logic is reasonably clear; elsewhere, tests are framed as difference hypotheses. H1–H5 and H7 are journey-level outcomes; H6 is user-level profiling.

- **H1 (Compression):** Journey compression differs by observed LLM exposure state, with decomposition by journey augmentation state ($G0/G1/G2$).

- **H2a (Shorter observable interval):** Observed LLM exposure is associated with a shorter trace-level interval from first consideration to first intent.
- **H2b (Longer observable interval):** In some settings, observed LLM exposure is associated with a longer trace-level interval to intent if LLM-mediated comparison induces additional validation loops.
- **H3 (Discovery mix shift):** Pre-intent channel shares differ by observed exposure state, with explicit decomposition for LLM-augmented versus non-augmented journeys among exposed users.
- **H4 (Transition structure):** Stage-transition patterns differ by observed exposure state, including patterns consistent with more direct jumps or fewer loops, interpreted with the $G0/G1/G2$ decomposition.
- **H5 (Breadth):** Breadth of sources explored prior to intent differs by observed exposure state, with decomposition by journey augmentation state.
- **H6 (Survey alignment):** User-level observed LLM exposure is systematically related to survey measures of media use and digital media self-efficacy, indicating consistent profiling rather than random noise.
- **H7 (Moderation):** Tech affinity/involvement moderates associations between observed LLM exposure and funnel outcomes (when available).

### 4.6 Labeling pipeline (conceptual overview)

The labeling pipeline is a hybrid system designed for production-scale measurement rather than one-off annotation. Deterministic behavioral cues handle semantically crisp trace patterns, model-assisted classification handles stage and entity cases that require context, and targeted manual audits test the boundaries where failure risk is highest. The pipeline enforces a strict no-lookahead rule, preserves auditable intermediate artifacts, and supports high-agreement subset analyses without rewriting the underlying trace. The result is a reproducible, repeatable journey measurement system for noisy multi-source behavioral logs.

Low-level implementation details (token budgets, price-verification tolerances, and infrastructure settings) are provided in Appendix B.

### 4.7 Empirical strategy and statistical inference

Our primary estimands are differences in journey-level outcomes between LLM-exposed and non-exposed observations.[1] Because event logs contain repeated measures within users, the unit of statistical inference is the user (caseid) rather than individual events. We therefore report uncertainty using standard errors clustered at the caseid level for regression models and caseid-level bootstrapping for non-parametric estimands such as differences in transition probabilities (Cameron and Miller, 2015; MacKinnon *et al*, 2023). The empirical design follows the measurement architecture: user-level inference, clustered uncertainty, and separate exposure-versus-augmentation estimands.

To keep constructs distinct, we report (i) exposure-state contrasts using $E_i$ and (ii) journey-augmentation decomposition using $G0/G1/G2$. The first estimand addresses whether users with

[1] The full specification registry (outcomes, estimands, covariate sets, optimization defaults, weights, and inference settings) is provided in Appendix D.

observed LLM exposure exhibit different journey traces overall. The second estimand addresses whether differences are concentrated in journeys that are themselves LLM-augmented ($A_{pre} = 1$) versus non-augmented journeys among exposed users ($A_{pre} = 0$). We treat this decomposition as a core identification guardrail against user-level selection being mistaken for journey-level augmentation effects.

We report both unadjusted and adjusted associations. Adjusted models include a pre-specified set of covariates: demographics, digital media self-efficacy (MDSE), media consumption/distrust (CMEDIA), news frequency (NEWS), and baseline trace intensity (total events and active days observed). Where survey weights are available, we report weighted estimates as a robustness check alongside unweighted estimates.

For fractional outcomes such as count-based compression and pre-intent channel shares, we estimate fractional response models (Papke and Wooldridge, 1996) with caseid-clustered standard errors. For time-to-event outcomes (time-to-intent and time-to-conversion-proximate), we estimate Kaplan–Meier curves and Cox proportional hazards models (Kaplan and Meier, 1958; Cox, 1972), using robust variance clustered by caseid. For transition-structure differences, we compute stage-to-stage transition probabilities and obtain 95% confidence intervals for differences via nonparametric bootstrapping over caseids.

Because observed LLM exposure is not randomly assigned, we supplement regression adjustment with propensity-based reweighting. Specifically, we estimate a propensity model for observed exposure using pre-specified covariates and compute inverse-probability weights; covariate balance is assessed using standardized mean differences. These analyses are intended to reduce observable differences between groups rather than to claim causal identification.

### 4.8 Technique selection rationale and parameterization details

Estimator choice follows outcome support and censoring structure rather than preferred findings. Fractional-response GLMs (logit link) are used for bounded outcomes in $[0, 1]$ (compression shares and discovery-mix shares). Cox proportional hazards models are used for time-to-event outcomes with right censoring. Poisson GLMs are used for non-negative count breadth outcomes. Survey-alignment models are estimated with weighted least squares because those outcomes are case-level survey means with provided sample weights.

Specification is pre-declared as unadjusted (outcome $\sim$ LLM) and adjusted (outcome $\sim$ LLM + demographics + MDSE + CMEDIA + NEWS + baseline activity). Baseline activity is total events and active days. Demographics enter as available indicator dummies after coding harmonization. We do not use user-supplied starting values; model fitting uses library defaults for initialization and optimizer settings.

Uncertainty is cluster-robust at the caseid level for regression models. Transition-difference uncertainty uses caseid bootstrap resampling (main specification: 500 draws). For Cox models, convergence fallback uses a penalizer ladder (0.00, 0.01, 0.10, 1.00) when needed. Propensity weighting uses unstabilized inverse-probability weights from a logistic observed-exposure model with the same pre-specified covariates used in adjustment models.

Timing outcomes are defined in minutes from first consideration within journey. Right-censored records are retained when the target event is not observed. Negative observed lags are treated as invalid ordering artifacts and recast to censored observations; non-positive durations are clipped to a small positive constant for Cox/KM compatibility. Baseline journey segmentation for the main analysis uses a 24-hour inactivity gap and 168-hour cap; robustness checks vary gap and cap as reported in Table 5.

Adjusted models use complete-case estimation for required covariates, which produces model-specific sample sizes. We therefore report caseid/journey counts for each hypothesis family in the table notes to make missingness and effective sample changes explicit. As an additional robustness design, we emphasize within-user comparisons among exposure-positive users who have both $A_{pre} = 0$ and $A_{pre} = 1$ journeys, so user-level stable differences are differenced out when interpreting augmentation-specific patterns.

# 5 Validation

The measurement system is evaluated in three layers. First, we use a consolidated expert-label pool with fixed dev/holdout splits for stage and price, plus held-out guided-choice audits for product name, brand, and category. Second, we retain a targeted manual-adjudication packet focused on the known failure boundaries (stage boundaries, entity ambiguity, price semantics, and exposure substrings) and treat it as a stress test rather than a representative performance estimate. Third, one domain expert hand-reviewed 100 rows from each final inference layer (stage classification, entity extraction, price semantics, and observed LLM exposure detection). Together, these layers characterize both held-out quality and deployed-output behavior after remediation; because the key metrics are already stated directly below, the manuscript keeps the validation evidence in prose rather than as separate standalone tables.

On the reaudited expert pool, stage classification reaches accuracy 0.914 and macro-F1 0.808 ($\kappa = 0.858$) on dev and accuracy 0.879 and macro-F1 0.834 ($\kappa = 0.808$) on holdout. Residual uncertainty is now concentrated in a small number of consideration-versus-intent boundary rows rather than in systematic conversion failure. The final 100-row production audit also shows exact agreement 1.000 ($\kappa = 1.000$), which places the stage layer above the paper's ship gate for both held-out performance and deployed-output agreement.

Entity extraction is now reported as three audited fields rather than as a single all-fields exact-match construct. On held-out guided-choice audits, micro-F1 is 0.844 for product name, 0.906 for brand, and 0.875 for category. In the 100-row expert audit, exact agreement is 0.880/0.970/1.000 and $\kappa$ is 0.874/0.968/1.000 for product name, brand, and category, respectively. The targeted entity packet shows that the remaining error is dominated by product-title over-specification and normalization rather than brand or category confusion.

Price semantics is evaluated on the consolidated expert pool and the final thresholded price layer. The current price pass reaches positive-class F1 1.000 and $\kappa = 1.000$ on both dev and holdout splits. In the 100-row final-output audit, price-present agreement is 0.970, $\kappa = 0.917$, precision is 1.000, recall is 0.880, and F1 is 0.936. The targeted price packet was useful because it isolated semantic false positives from non-price dollar mentions; those cases motivated the stricter thresholding now used in the final stack.

Observed-LLM exposure detection is evaluated with the targeted 48-row boundary packet and a 100-row final-output audit. The packet localized the remaining risk to ambiguous substring/search cases rather than canonical domain or app hits, and the final detector reaches exact agreement 0.960, $\kappa = 0.919$, precision 0.911, recall 1.000, and F1 0.953. This is the basis for the stricter canonical-exposure robustness check reported later.

# 6 Results

## 6.1 Sample composition and group sizes

The final analysis dataset comprises 587 distinct users (caseids), 3,360 journeys, and 2,917,655 trace events within the observation window. Of these users, 185 are exposure-positive (at least one observed LLM touchpoint), and 402 have no observed LLM touchpoint. At the journey level, 1,102 journeys belong to exposure-positive users and 2,258 belong to non-exposed users. Table 2 reports the $G0/G1/G2$ decomposition, which separates user-level exposure from journey-level augmentation; 91 exposure-positive users contribute both augmented and non-augmented journeys, allowing within-user contrasts. Table 3 summarizes sample characteristics by exposure group, and Table 4 reports unadjusted and adjusted model estimates with caseid-clustered uncertainty.

Table 2: Exposure versus augmentation decomposition (guardrail against construct conflation).

| **Group definition** | **Users** | **Journeys** | **Journey share** |
|---|---|---|---|
| G0: $E = 0$, $A_{pre} = 0$ (non-exposed users) | 402 | 2,258 | 67.20% |
| G1: $E = 1$, $A_{pre} = 0$ (exposed users; non-augmented journeys) | 180 | 890 | 26.49% |
| G2: $E = 1$, $A_{pre} = 1$ (exposed users; pre-intent LLM-augmented journeys) | 96 | 212 | 6.31% |

*Notes:* $E$ = user-level observed LLM exposure (any AI/LLM touch in-window under the primary registry). $A_{pre}$ = journey-level pre-intent LLM augmentation indicator under the stricter canonical in-journey channel mapping. Exposed users with both $A_{pre} = 0$ and $A_{pre} = 1$ journeys: 91 of 185. Raw parser exceptions ($E = 0$, raw pre-intent LLM signal=1) before hierarchy gating: 2 journeys.

Table 3: Sample descriptives by exposure group.

| Metric | LLM-exposed | Non-exposed |
|---|---|---|
| N caseids | 185 | 402 |
| N journeys | 1,102 | 2,258 |
| N events | 1,691,708 | 1,225,947 |
| Median journeys per user | 5.000 | 5.000 |
| MDSE mean | 4.937 | 4.796 |
| CMEDIA mean | 4.160 | 4.152 |
| NEWS mean | 4.059 | 3.915 |
| Total events per user | 9,144.368 | 3,049.619 |
| Active days | 25.676 | 21.965 |

Notes: Counts are integers with comma grouping. MDSE = digital media self-efficacy; CMEDIA = media consumption/distrust; NEWS = news-frequency scale.

## 6.2 H1 Compression

Compression differences are not robust after adjustment. Descriptively, LLM-exposed journeys show slightly higher count compression (0.0236 vs 0.0201) and higher time compression (0.498 vs 0.370). The unadjusted count-compression AME is +0.003 (95% CI [-0.002, 0.009], $p = 0.227$)

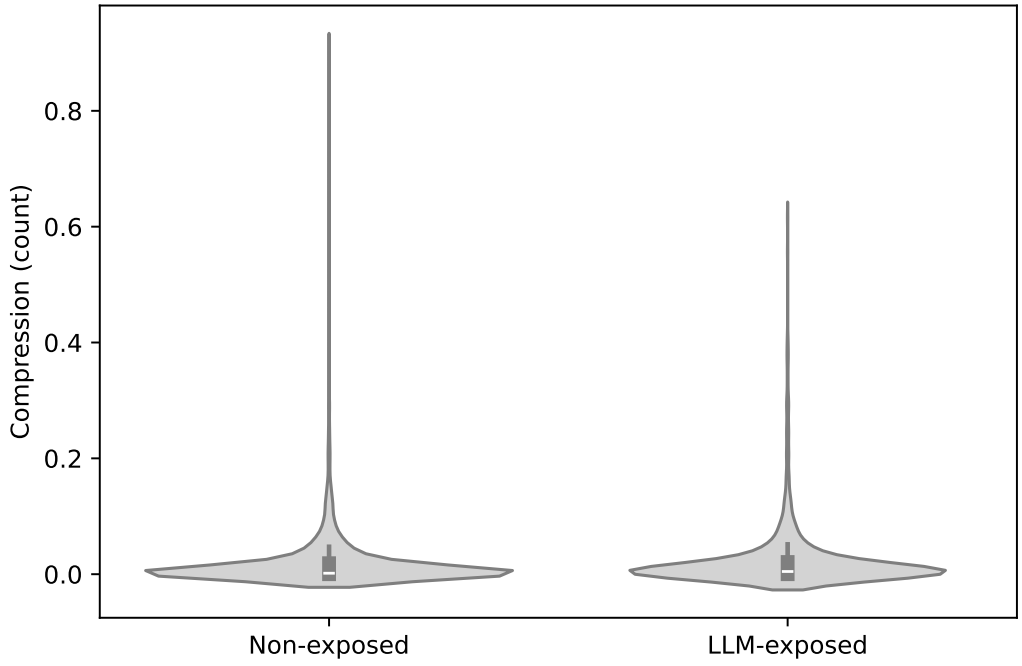


Figure 2: Compression (count) distribution by exposure group.

and the adjusted AME is +0.003 (95% CI [-0.004, 0.010], $p = 0.379$). Time-based compression is positive before adjustment (AME +0.126, 95% CI [0.073, 0.179], $p < 0.001$) and returns to zero after adjustment (AME -0.010, 95% CI [-0.069, 0.049], $p = 0.740$). H1 therefore does not provide adjusted support for a headline acceleration result.

### 6.3 H2 Timing

Timing estimates point in the same raw direction as compression but remain imprecise after adjustment. Time-to-intent is shorter for exposure-positive journeys in the censored survival frame (mean follow-up 3,501 vs 3,756 minutes), but the Cox contrasts do not separate clearly: HR=1.192 unadjusted (95% CI [0.974, 1.458], $p = 0.089$) and HR=1.203 adjusted (95% CI [0.956, 1.513], $p = 0.116$). The data therefore do not establish faster progression to intent once covariates are included.

For conversion-proximate timing, no observed events remain in the survival frame after censoring and negative-lag recoding, so we do not report a Cox contrast or Kaplan–Meier panel for this endpoint. Combined with the weaker validation of the conversion-proximate class, this outcome remains outside the paper's primary interpretive results in the current window.

### 6.4 H3 Discovery mix

Discovery mix is the most stable empirical result in the paper. Pre-intent LLM share is higher in exposure-positive journeys descriptively (0.00546 vs 0.00002) and in both unadjusted and adjusted models (adjusted AME +0.009, 95% CI [0.005, 0.014], $p < 0.001$). Review/UGC share is also higher after adjustment (AME +0.028, 95% CI [0.003, 0.053], $p = 0.027$). Classic-search share is positive but not distinguishable from zero after adjustment (AME +0.017, $p = 0.292$),

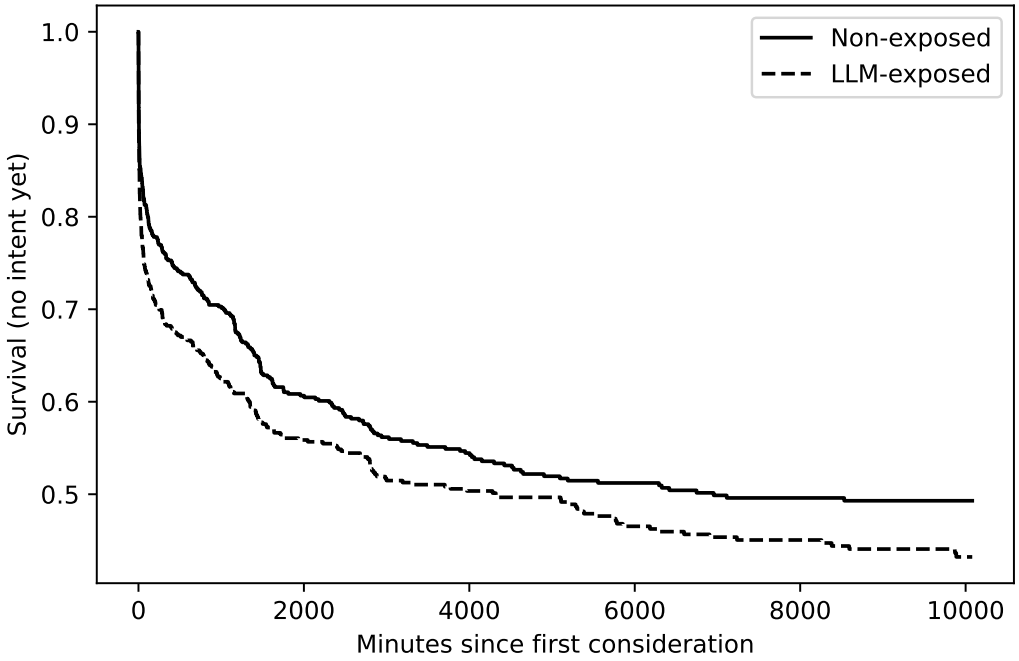


Figure 3: Kaplan–Meier survival curves for time-to-intent (solid: non-exposed; dashed: LLM-exposed).

and retailer/marketplace and direct-brand shares remain null. Figure 4 shows the corresponding compositional shift: the visible pre-intent mix changes by adding detectable LLM and review/UGC activity, not by collapsing retailer or direct-brand traces.

### 6.5 H4 Transition structure

Transition estimates are directionally consistent with more direct trace-level progression, but they are not precise. Consideration→intent transitions are more common among exposure-positive journeys (0.252 vs 0.199), and consideration→consideration loops are less common (0.748 vs 0.801). The bootstrap difference for $C \to I$ is +0.053 (95% CI [-0.070, 0.188]) and the $C \to C$ difference is -0.055 (95% CI [-0.185, 0.074]); direct $C \to V$ transitions are zero in both groups in the current run. The direction is stable, but the interval width keeps H4 secondary to the discovery-mix result in the current window.

### 6.6 H5 Breadth

Breadth shows no adjusted difference. Measured as the number of unique pre-intent sources, the Poisson model yields an adjusted incidence rate ratio of 0.980 (95% CI [0.911, 1.054], $p = 0.589$). Because breadth depends on the secondary entity-extraction layer, we use it as a descriptive cross-check rather than a core test. On that standard, breadth differences are negligible in this observation window.

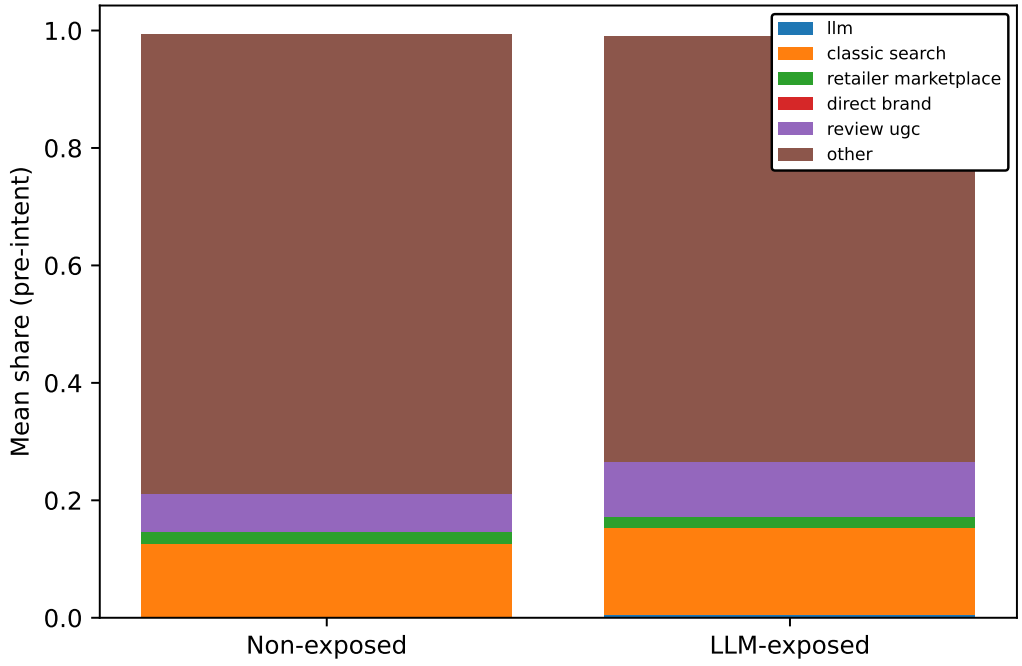


Figure 4: Pre-intent channel shares by exposure group (stacked bars labeled by channel class).

### 6.7 H6 Survey alignment

Survey alignment is limited to NEWS scores. LLM exposure is associated with higher NEWS scores in adjusted models (0.215, 95% CI [0.030, 0.400], $p$=0.023), while MDSE and CMEDIA differences are not statistically distinguishable from zero. The results indicate a measurable link to news frequency, not a broad split in digital media self-efficacy or media distrust in this sample.

### 6.8 H7 Moderation

Moderation is null in the current sample. The interaction between LLM exposure and MDSE in predicting compression is effectively null (interaction coefficient 0.010, 95% CI [-0.241, 0.261], $p = 0.939$) and does not warrant a main-text figure.[2]

Robustness checks across exposure definitions, reweighting, label restrictions, within-user augmentation contrasts, and journey segmentation are summarized in Table 5. Full segmentation and within-user tables are provided in Appendix E.

Table 5 leaves the main interpretation unchanged. First, at the reported precision shown in the table, the IPW sensitivity row is numerically unchanged from the main adjusted model, which indicates that reweighting does not alter the reported compression, timing, or pre-intent LLM-share conclusions in this sample. The stricter canonical exposure definition weakens compression and timing further (count-compression AME +0.002, $p = 0.538$; intent HR 1.142, $p = 0.274$) but strengthens the pre-intent LLM-share effect (AME +0.039, $p < 0.001$). Second, aggressive high-agreement restrictions do not rescue a stronger acceleration story: they preserve only a noisy compression estimate and do not yield an interpretable timing contrast. Third, the within-user

[2]The fitted marginal-effects profile is nearly flat: predicted compression changes little across low versus high MDSE, and the exposed/non-exposed lines remain close to parallel.

Table 4: Main model estimates with caseid-clustered uncertainty (unadjusted and adjusted), Panel A.

| Outcome | Model | Unadjusted | | | Adjusted | | |
|---|---|---|---|---|---|---|---|
| | | Estimate (SE) | 95% CI | $p$ | Estimate (SE) | 95% CI | $p$ |
| H1: Compression (count) | Frac. logit (AME) | 0.003 (0.003) | [-0.002, 0.009] | 0.227 | 0.003 (0.003) | [-0.004, 0.010] | 0.379 |
| H1: Compression (time) | Frac. logit (AME) | 0.126 (0.027) | [0.073, 0.179] | < 0.001 | -0.010 (0.030) | [-0.069, 0.049] | 0.740 |
| H2: Time to intent | Cox PH (HR) | 1.192 (—) | [0.974, 1.458] | 0.089 | 1.203 (—) | [0.956, 1.513] | 0.116 |
| H2: Time to conversion-proximate | Cox PH (HR) | — | — | — | — | — | — |
| H3: Pre-intent LLM share | Frac. logit (AME) | 0.010 (0.002) | [0.005, 0.015] | < 0.001 | 0.009 (0.002) | [0.005, 0.014] | < 0.001 |
| H3: Pre-intent classic-search share | Frac. logit (AME) | 0.022 (0.013) | [-0.004, 0.048] | 0.094 | 0.017 (0.016) | [-0.015, 0.049] | 0.292 |
| H3: Pre-intent retailer/marketplace share | Frac. logit (AME) | -0.003 (0.006) | [-0.014, 0.009] | 0.638 | -0.004 (0.008) | [-0.019, 0.011] | 0.622 |
| H3: Pre-intent direct-brand share | Frac. logit (AME) | -0.000 (0.000) | [-0.000, 0.000] | 1.000 | -0.000 (0.000) | [-0.000, 0.000] | 1.000 |
| H3: Pre-intent review/UGC share | Frac. logit (AME) | 0.029 (0.011) | [0.008, 0.051] | 0.008 | 0.028 (0.013) | [0.003, 0.053] | 0.027 |
| H3: Pre-intent other share | Frac. logit (AME) | -0.057 (0.017) | [-0.090, -0.023] | < 0.001 | -0.048 (0.021) | [-0.090, -0.006] | 0.024 |

Notes: AME = average marginal effect; HR = hazard ratio; IRR = incidence rate ratio; WLS = weighted least squares; CI = confidence interval; SE = standard error. Em dashes indicate not estimated for that specification. Caseids/journeys (unadjusted; adjusted): H1/H3/H5 = 587/3,360; 395/2,216. H2 = 346/1,270; 234/870. H4 = 587/3,360; not adjusted. H6 MDSE = 491/491; 491/491. H6 CMEDIA = 456/456; 456/456. H6 NEWS = 528/528; 528/528. H7 = not unadjusted; 491/2,815.

$G2 - G1$ contrast among the 91 exposure-positive users who contribute both journey types is especially decisive: augmented journeys show higher visible LLM share ($\Delta = +0.030$, 95% CI [0.022, 0.038]) and slightly *lower* compression ($\Delta = -0.006$, 95% CI [-0.015, 0.000]). Together with the segmentation grid, where H1 AMEs remain small (+0.000 to +0.002), H2 intent HRs remain only modestly above 1.0 (1.114 to 1.185), and $C \to I$ differences remain tightly clustered (+0.052 to +0.053), the stable result is the visible pre-intent LLM-share shift. A broad acceleration claim does not survive the more restrictive specifications.

Table 4 (continued): Main model estimates with caseid-clustered uncertainty (unadjusted and adjusted), Panel B.

| | | Unadjusted | | | Adjusted | | |
|---|---|---|---|---|---|---|---|
| Outcome | Model | Estimate (SE) | 95% CI | $p$ | Estimate (SE) | 95% CI | $p$ |
| H4: Transition $C \to I$ | Bootstrap diff. | 0.053 (—) | [-0.070, 0.188] | — | — | — | — |
| H4: Transition $C \to V$ | Bootstrap diff. | 0.000 (—) | [0.000, 0.000] | — | — | — | — |
| H4: Transition $C \to C$ | Bootstrap diff. | -0.055 (—) | [-0.185, 0.074] | — | — | — | — |
| H5: Breadth (IRR) | Poisson (IRR) | 0.994 (—) | [0.936, 1.056] | 0.855 | 0.980 (—) | [0.911, 1.054] | 0.589 |
| H6: MDSE score | WLS | 0.098 (0.092) | [-0.081, 0.278] | 0.283 | 0.088 (0.096) | [-0.101, 0.277] | 0.362 |
| H6: CMEDIA score | WLS | -0.002 (0.106) | [-0.210, 0.207] | 0.988 | 0.056 (0.112) | [-0.163, 0.276] | 0.614 |
| H6: NEWS score | WLS | 0.164 (0.088) | [-0.009, 0.338] | 0.063 | 0.212 (0.094) | [0.028, 0.397] | 0.025 |
| H7: MDSE interaction | Frac. logit (interact.) | — | — | — | 0.010 (0.128) | [-0.241, 0.261] | 0.939 |

Notes: AME = average marginal effect; HR = hazard ratio; IRR = incidence rate ratio; WLS = weighted least squares; CI = confidence interval; SE = standard error. Em dashes indicate not estimated for that specification. Caseids/journeys (unadjusted; adjusted): H1/H3/H5 = 587/3,360; 395/2,216. H2 = 346/1,270; 234/870. H4 = 587/3,360; not adjusted. H6 MDSE = 491/491; 491/491. H6 CMEDIA = 456/456; 456/456. H6 NEWS = 528/528; 528/528. H7 = not unadjusted; 491/2,815.

Table 5: Structured sensitivity checks (IPW, exposure definition, label restriction, within-user augmentation, and segmentation summary).

| Specification | Compression | Time to intent | Pre-intent LLM share | Transition $C \to I$ |
|---|---|---|---|---|
| Main adjusted model (primary broad exposure registry) | AME 0.003, $p$=0.379 | HR 1.203 [0.956, 1.513], $p$=0.116 | AME 0.009, $p$=< 0.001 | 0.053 [-0.070, 0.188] |
| IPW-adjusted sensitivity (unchanged at reported precision) | AME 0.003, $p$=0.379 | HR 1.203 [0.956, 1.513], $p$=0.116 | AME 0.009, $p$=< 0.001 | — |
| Strict canonical LLM exposure definition | AME 0.002, $p$=0.538 | HR 1.142 [0.900, 1.449], $p$=0.274 | AME 0.039, $p$=< 0.001 | 0.062 [-0.063, 0.207] |
| High-agreement journeys only ($unclear = 0$; unadjusted subset) | AME 0.065, $p$=0.539 | — | AME -0.000, $p$=1.000 | — |
| Within-user augmentation contrast (paired $G2 - G1$ among both-state exposed users) | $\Delta$ -0.006 [-0.015, 0.000] | — | $\Delta$ 0.030 [0.022, 0.038] | — |
| Segmentation grid (9 gap/cap combinations; unadjusted range) | AME range [0.000, 0.002] | HR range [1.114, 1.185] | — | $\Delta$ range [0.052, 0.053] |

Notes: IPW = inverse-probability weighting. At the displayed precision, the IPW row is numerically identical to the main adjusted model. $C \to I$ denotes the transition from consideration to intent. The main/IPW/strict rows report model-based estimates; the high-agreement row reports an unadjusted subset analysis with $unclear = 0$; the within-user row reports paired user-level mean differences comparing $G2$ and $G1$; the segmentation row summarizes the full grid reported in Appendix E. Em dashes indicate not reported or not estimable for that specification.

# 7 Discussion

LLM assistants have changed the observable structure of discovery. If analysts continue to assume search-led journeys, they will read missing early-stage touches as higher funnel efficiency or weaker consideration when those activities have shifted into a conversational interface. The contribution of this paper is to treat LLM usage as a first-class touchpoint and to provide journey diagnostics that keep attribution and funnel reporting aligned with the current observability boundary.

The stage codebook is deliberately pragmatic. Classic hierarchy-of-effects and goal-based funnel models (Lavidge and Steiner, 1961; Colley, 1961; Wijaya, 2012) remain useful as managerial language, but clickstream traces rarely provide reliable signals for awareness. Consistent with customer journey research on non-linear movement and looping (Vakratsas and Ambler, 1999; Lemon and Verhoef, 2016) and practitioner frameworks that emphasize iterative exploration (McKinsey and Company, 2009; Rennie and Protheroe, 2020), we operationalize three stages (consideration, intent, conversion-proximate) and analyze path structure and timing rather than impose a single canonical sequence.

The observed patterns are consistent with several mechanisms. LLMs can front-load shortlist formation, move comparison activity into one conversational interface, and redirect downstream validation toward a smaller set of price, shipping, retailer, or checkout-adjacent pages. In passive logs, each pathway changes the *visible* journey even when deliberation continues.

A useful distinction is between behavioral compression and observable compression. Behavioral compression would imply that LLMs reduce the actual amount of consumer deliberation or decision effort. Observable compression means that the traceable portion of the journey becomes shorter because some discovery and evaluation activities move into less visible environments. The current study is designed to measure observable compression. Passive behavioral logs can reveal changes in the structure of observed touchpoints, but they cannot fully capture cognitive effort, conversational search depth, or within-interface comparison.

The empirical result is correspondingly narrow and clear. The stable finding is a shift in visible pre-intent discovery, not a universal acceleration effect. Pre-intent LLM share rises sharply,

review/UGC share also rises, and that pattern survives adjustment and robustness checks. Compression and timing move in the expected raw direction, but they do not survive adjustment as headline behavioral results. The within-user $G2 - G1$ contrast sharpens the interpretation: augmented journeys show higher visible LLM share without higher compression. That pattern is consistent with observability change. The paper therefore makes a measurement claim rather than a behavioral acceleration claim: once conversational discovery enters the path, standard journey analytics understate early-stage activity and overstate the apparent role of downstream visible touchpoints.

This interpretation is consistent with prior evidence that conversational agents affect perceived information value and purchase-related intentions in commerce contexts (Lo Presti *et al*, 2021; Le, 2023; Luo *et al*, 2023) and with work showing that generative-AI chatbots and search engines produce different evaluation behaviors (Kim and Priluck, 2025). The manuscript measures observable journey structure, not internal cognitive effort, and the interpretation remains bounded to that distinction.

### 7.1 Managerial implications

In path-based marketing analytics, adjacency and order are treated as informative signals about channel roles (Li and Kannan, 2014; Anderl *et al*, 2016) and touchpoint contributions to consideration (Baxendale *et al*, 2015). When conversational discovery becomes a common first touch, two failure modes follow. First, path-based models may attribute upstream influence to the first observable non-LLM touchpoint (for example, a retailer product detail page), inflating the apparent marginal contribution of lower-funnel interactions. Second, common "funnel health" diagnostics that rely on counts of early-stage touches (for example, the number of review sites visited) can register reduced upper-funnel engagement when those behaviors have migrated into the LLM interface.

This issue is operational rather than hypothetical. Marketplace advertisers increasingly manage and allocate spend using stage-specific funnel metrics (e.g., awareness, consideration, purchase) at scale (Qin, Pauwels and Zhou, 2024). If conversational discovery moves consideration into an LLM interface that collapses multiple steps into a single observed event, then early-stage activity can be systematically undercounted. A sudden increase in direct product-page visits, faster apparent movement to intent, or fewer visible review-site visits may appear to indicate improved funnel efficiency. Among LLM-exposed users, however, those same patterns may signal that consideration has moved into conversational interfaces. That undercount can inflate the apparent marginal contribution of lower-funnel touchpoints and steer budget reallocation toward tactics that merely appear to start the journey in clickstream logs. The guardrail metrics proposed here are intended to identify that measurement drift before it affects spend decisions. This complements calls to keep analytics aligned with decision-journey strategy rather than optimizing only what is most visible in dashboards (Vollrath and Villegas, 2022; Iacobucci *et al*, 2019).

**Operational controls for journey dashboards.** For teams running journey analytics in production, the following controls can be implemented as routine diagnostics:

1. **Treat LLM as a separate channel class.** Maintain a versioned list of LLM domains/app bundles and monitor user-level first observed exposure timing.

2. **Track observability alarms, not just performance KPIs.** Add time-to-intent distributions, direct transition probabilities ($C \to I$, $C \to V$), and pre-intent LLM-touch share as measurement-health indicators.

3. **Stratify attribution outputs by LLM exposure.** Report channel contributions separately for LLM-exposed and non-exposed journeys to reduce first-observed-touch misattribution.

4. **Audit conversion-proximate evidence routinely.** Run recurring targeted checks on checkout/cart/retailer-link signals and escalate drift before model refreshes.

These measures function as observability guardrails. Compression and transition directness indicate how much of the journey may have become unobservable under conventional channel taxonomies, while transition matrices retain the non-linear structure of browsing. In practice, they complement attribution models by indicating when the measurement environment has changed. A shift toward higher direct transitions and shorter time-to-intent may reflect increased exposure to conversational intermediaries rather than a genuine improvement in media efficiency. Journey dashboards should therefore treat unusually direct consideration-to-intent movement as an observability alarm, not automatic evidence of improved media performance. In this setting, the limiting factor is often not model complexity but whether the measurement layer still reflects how discovery is actually observed.

### 7.2 Limitations and future research

This paper makes a descriptive measurement claim, and the evidence is matched to that claim. It does not claim causal identification. Users self-select into LLM usage, and the exposure-timing construct is first observed usage within the observation window rather than first-ever adoption. Passive logs do not observe conversational content, within-interface comparison, or cognitive effort, so the analysis is about observable journey structure. The main inferential claims therefore rest on the best-validated layers: stage for consideration and intent, and canonical exposure detection. Conversion-proximate classification, entity breadth, and price semantics remain secondary audited layers. The observation window is one month and should be replicated across longer windows, categories, and platforms.

Future work can extend this framework in three directions. First, replication across longer windows and additional datasets would allow more stable estimation of channel substitution in the pre-intent discovery mix. Second, quasi-experimental exposure-timing designs (for example, within-user event studies around first observed LLM usage) could better separate behavioral change from selection. Third, richer instrumentation, including consented capture of conversational intent signals, could connect trace-level compression to consumer-perceived effort and satisfaction.

## 8 Conclusion

LLM assistants are changing the observable structure of consumer discovery. Journey analytics and attribution systems that continue to assume search-led discovery will misstate funnel length, timing, and channel roles once consideration shifts into conversational interfaces. This paper provides a measurement correction: it treats LLM usage as a first-class touchpoint detectable in passive logs, operationalizes a non-linear stage codebook suited to clickstream evidence, and introduces journey-level metrics for compression, timing, discovery mix, and transition structure. The empirical conclusion is equally bounded. Observed LLM exposure reshapes the visible pre-intent discovery mix, and that result survives adjustment and robustness checks. Broad acceleration claims do not. The practical response is to update measurement, channel taxonomies, and dashboard diagnostics so that hidden consideration is not mistaken for genuine funnel improvement.

## Declarations

Funding: No external funding was received.

Competing interests: The authors declare no competing interests.

Ethics approval: The study uses de-identified behavioral trace data provided by a third-party passive measurement panel provider. The protocol received exempt institutional review board approval.

Consent to participate: Informed consent was obtained under the approved exempt protocol.

Consent for publication: Not applicable.

Data availability: The data analyzed in this study are proprietary and subject to third-party restrictions. Aggregated results and analysis code availability are described below.

Code availability: Analysis code is available upon request; repository link available on request.

Use of AI tools: Large language models were used to assist with event labeling (stage and entity extraction) under a structured schema and were evaluated with targeted human audits. The models were not used as authors and are not responsible for the manuscript's claims.

Author contributions: All authors contributed to conceptualization, methodology, analysis, and writing.

## References

Anderl, E., Becker, I., von Wangenheim, F. and Schumann, J.H. (2016) 'Mapping the customer journey: lessons learned from graph-based online attribution modeling', *International Journal of Research in Marketing*. Available at: https://doi.org/10.1016/j.ijresmar.2016.03.001.

Baxendale, S., Macdonald, E.K. and Wilson, H.N. (2015) 'The impact of different touchpoints on brand consideration', *Journal of Retailing*. Available at: https://doi.org/10.1016/j.jretai.2014.12.008.

Cameron, A.C. and Miller, D.L. (2015) 'A practitioner's guide to cluster-robust inference', *Journal of Human Resources*, 50(2), pp. 317–372. Available at: https://doi.org/10.3368/jhr.50.2.317.

Colley, R.H. (1961) *Defining Advertising Goals for Measured Advertising Results (DAGMAR)*. New York, NY: Association of National Advertisers.

Cox, D.R. (1972) 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), pp. 187–220. Available at: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Iacobucci, D., Petrescu, M., Krishen, A. and Bendixen, M. (2019) 'The state of marketing analytics in research and practice', *Journal of Marketing Analytics*, 7, pp. 152–181. Available at: https://doi.org/10.1057/s41270-019-00059-2.

Kaplan, E.L. and Meier, P. (1958) 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, 53(282), pp. 457–481. Available at: https://doi.org/10.1080/01621459.1958.10501452.

Kim, S. and Priluck, R. (2025) 'Consumer responses to generative AI chatbots versus search engines for product evaluation', *Journal of Theoretical and Applied Electronic Commerce Research*, 20(2), 93. Available at: https://doi.org/10.3390/jtaer20020093.

Lavidge, R.J. and Steiner, G.A. (1961) 'A model for predictive measurements of advertising effectiveness', *Journal of Marketing*, 25(6), pp. 59–62. Available at: https://doi.org/10.2307/1248516.

Le, X.C. (2023) 'Inducing AI-powered chatbot use for customer purchase: the role of information value and innovative technology', *Journal of Systems and Information Technology*, 25(2), pp. 219–241. Available at: https://doi.org/10.1108/JSIT-09-2021-0206.

Lemon, K.N. and Verhoef, P.C. (2016) 'Understanding customer experience throughout the customer journey', *Journal of Marketing*, 80(6), pp. 69–96. Available at: https://doi.org/10.1509/jm.15.0420.

Li, H. and Kannan, P.K. (2014) 'Attributing conversions in a multichannel online marketing environment: an empirical model and a field experiment', *Journal of Marketing Research*, 51(1), pp. 40–56. Available at: https://doi.org/10.1509/jmr.13.0050.

Lo Presti, L., Maggiore, G. and Marino, V. (2021) 'The role of the chatbot on customer purchase intention: toward digital relational sales', *Italian Journal of Marketing*. Available at: https://doi.org/10.1007/s43039-021-00029-6.

Luo, X. *et al* (2023) 'Chatbots in e-commerce: the effect of chatbot language style on customers' continuance usage intention and attitude toward brand', *Journal of Retailing and Consumer Services*, 71, 103209. Available at: https://doi.org/10.1016/j.jretconser.2022.103209.

MacKinnon, J.G., Nielsen, M.O. and Webb, M.D. (2023) 'Cluster-robust inference: A guide to empirical practice', *Journal of Econometrics*, 232(2), pp. 272–299. Available at: https://doi.org/10.1016/j.jeconom.2022.04.001.

McKinsey and Company (2009) 'The consumer decision journey'. Available at: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-consumer-decision-journey (Accessed: 29 January 2026).

Papke, L.E. and Wooldridge, J.M. (1996) 'Econometric methods for fractional response variables with an application to 401(k) plan participation rates', *Journal of Applied Econometrics*, 11(6), pp. 619–632. Available at: https://econpapers.repec.org/article/jaejapmet/v_3a11_3ay_3a1996_3ai_3a6_3ap_3a619-32.htm.

Qin, V., Pauwels, K. and Zhou, B. (2024) 'Data-driven budget allocation of retail media by ad product, funnel metric, and brand size', *Journal of Marketing Analytics*, 12, pp. 235–249. Available at: https://doi.org/10.1057/s41270-024-00294-2.

Rennie, A. and Protheroe, J. (2020) 'How people decide what to buy lies in the "messy middle" of the purchase journey', *Think with Google*. Available at: https://business.google.com/au nz/think/consumer-insights/navigating-purchase-behavior-and-decision-making/ (Accessed: 29 January 2026).

Vakratsas, D. and Ambler, T. (1999) 'How advertising works: what do we really know?', *Journal of Marketing*, 63(1), pp. 26–43. Available at: https://www.jstor.org/stable/1251992 (Accessed: 29 January 2026).

Vollrath, M.D. and Villegas, S.G. (2022) 'Avoiding digital marketing analytics myopia: revisiting the customer decision journey as a strategic marketing framework', *Journal of Marketing Analytics*, 10, pp. 106–113. Available at: https://doi.org/10.1057/s41270-020-00098-0.

Wijaya, B.S. (2012) 'The development of hierarchy of effects model in advertising', *International Research Journal of Business Studies*, 5(1), pp. 73–85. Available at: https://doi.org/10.21632/irjbs.5.1.73-85.

Zimmermann, R. and Auinger, A. (2023) 'Developing a conversion rate optimization framework for digital retailers—case study', *Journal of Marketing Analytics*, 11, pp. 233–243. Available at: https://doi.org/10.1057/s41270-022-00161-y.

# A Survey Covariates and Analysis Plan

## A.1 Survey covariates

Table 6 lists survey constructs and item groups used for adjustment and profiling.

Table 6: Survey covariates used for adjustment and profiling.

| **Construct** | **Variables (codebook labels)** |
|---|---|
| News frequency | NEWS1a–NEWS1f (newspaper, broadcast, cable, social, radio, political websites/blogs) |
| Media consumption/distrust | CMEDIA1–CMEDIA4 (exist outside, criticize, oppose, challenge mainstream news) |
| Digital media self-efficacy | MDSE1–MDSE3 (spot fake news, distinguish fake vs. legitimate, identify inaccurate content) |
| Political/news interest | APOL, newsint |
| Demographics | birthyr, gender, race, educ, faminc_new, employ, marstat, ownhome, urbanicity2, inputstate, child18, immigrant, smoke100 |
| Weights | weight |

*Note:* NEWS = news-frequency scale; CMEDIA = media consumption/distrust scale; MDSE = digital media self-efficacy scale.

## A.2 Analysis plan mapping

Table 7 maps each hypothesis to outcomes, model classes, and primary reported outputs. Table 8 reports standardized mean-difference diagnostics.

Table 7: Hypotheses mapped to outcomes, models, and outputs.

| **Hyp.** | **Outcome** | **Model** | **Primary outputs** |
|---|---|---|---|
| H1 | Funnel compression (count + time) | Fractional response models | Coefficient table, compression distribution |
| H2a/H2b | Time to intent / conversion-proximate | Survival (Cox), KM curves | Hazard ratios, timing plot |
| H3 | Discovery mix (pre-intent shares) | Fractional response | Discovery mix bar chart |
| H4 | Transition structure | Transition deltas + bootstrap CIs | Transition table and annotated text summary |
| H5 | Breadth (unique sources) | Poisson | Breadth model estimate |
| H6 | Survey alignment | Linear profiling (WLS) | SMD table, profile summary |
| H7 | Moderation | Interaction models | Interaction coefficient and summary text |

*Note:* KM = Kaplan–Meier; CI = confidence interval; WLS = weighted least squares.

Table 8: Covariate balance diagnostics (standardized mean differences).

| Covariate | Mean (LLM) | Mean (Non) | SMD |
|---|---|---|---|
| MDSE (mean) | 4.937 | 4.796 | 0.141 |
| CMEDIA (mean) | 4.160 | 4.152 | 0.007 |
| NEWS (mean) | 4.059 | 3.915 | 0.151 |
| Birth year | 1,973.757 | 1,972.199 | 0.099 |
| Total events | 9,144.368 | 3,049.619 | 0.431 |
| Active days | 25.676 | 21.965 | 0.372 |
| Gender: male | 0.546 | 0.478 | 0.137 |
| Race: Black | 0.146 | 0.139 | 0.019 |
| Race: Hispanic | 0.081 | 0.067 | 0.053 |
| Race: Middle Eastern | 0.000 | 0.002 | -0.071 |
| Race: Native American | 0.011 | 0.020 | -0.074 |
| Race: Other | 0.022 | 0.012 | 0.071 |
| Race: Two or more races | 0.054 | 0.032 | 0.107 |
| Race: White | 0.649 | 0.694 | -0.097 |
| Education: 4-year | 0.330 | 0.221 | 0.244 |
| Education: High school graduate | 0.195 | 0.289 | -0.220 |
| Education: No HS | 0.038 | 0.047 | -0.047 |
| Education: Post-grad | 0.178 | 0.127 | 0.143 |
| Education: Some college | 0.178 | 0.216 | -0.095 |

Notes: SMD = standardized mean difference (unweighted). Weighted SMDs are not shown because weighted means were not exported in the current analysis artifact.

# B Appendix B: Implementation details for labeling pipeline

This appendix consolidates low-level implementation details referenced in the Methods section. The labeling system uses strict no-lookahead context construction, stable `event_hash` identifiers, and deterministic domain/app hinting. Timeline context is trimmed to a fixed token budget for batch inference to keep throughput stable across hardware. Entity extraction includes a high-precision price pass that restricts to rows with explicit currency/price tokens and retains a price only when the output matches a token within a fixed tolerance (±2%, ±5%, ±10%, ±15% in verification). A token-only fallback is not used in the high-agreement training set; unmatched or ambiguous multi-price records remain missing. Additional infrastructure parameters (batch sizes, GPU memory settings, and endpoint configuration) are maintained in internal runbooks to preserve reproducibility without exposing implementation-specific credentials or vendor identifiers.

# C Appendix C: LLM exposure definitions

The primary exposure definition used in the paper follows a broader AI/LLM registry that underlies the user-level exposure counts. It includes canonical LLM domains/apps as well as adjacent conversational surfaces and wrappers that can be identified in observed domain, URL, bundle, or text fields. The stricter robustness definition retains only canonical LLM domains and app bundles that map directly to the in-journey `llm` channel class (e.g., `chat.openai.com`, `claude.ai`, `gemini.google.com`, `perplexity.ai`) and their mobile app bundles. Table 5 compares headline results under the primary broad registry and the stricter canonical definition.

# D Appendix D: Model specification registry and reproducibility parameters

Table 9 documents the estimator-to-outcome mapping, core specification, and operational settings used in the manuscript's reported models. This appendix is intended to make model-choice rationale, parameters, and weighting/inference settings auditable in one place.

# E Appendix E: Sensitivity analyses

Table 10 reports the full gap/cap grid summarized in the main-text robustness discussion. Table 11 reports paired within-user contrasts between non-augmented ($G1$) and augmented ($G2$) journeys for the 91 exposure-positive users who contribute both states.

**Execution parameters and software stack.** Main-paper artifacts were generated with:

```
python run_all.py -gap-hours 24 -cap-hours 168
-bootstrap 500 -min-km-events-per-group 10
```

The environment used Python 3.12.3 and core packages: pandas 2.3.3, numpy 2.2.6, statsmodels 0.14.6, lifelines 0.30.1, matplotlib 3.10.8, and seaborn 0.13.2.

Table 9: Specification registry for main hypotheses.

| Hyp. | Outcome(s) | Model | Specification and parameter settings |
|---|---|---|---|
| H1 | Compression (count; time) | Fractional logit (GLM Binomial, logit link) | Unadjusted: `outcome` $\sim$ `llm_exposed` (exposure-state contrast using $E$). Adjusted: adds baseline activity (total events, active days), survey covariates (MDSE, CMEDIA, NEWS), and available demographic indicators. Decomposition table reports $G0/G1/G2$ counts for $A_{pre}$ separation. Caseid-clustered SE. No user-specified starting values. |
| H2 | Time-to-intent; time-to-conversion-proximate | Cox proportional hazards | Durations in minutes from first consideration; right censoring retained. Negative observed lags recast to censored; durations $\leq 0$ clipped to $10^{-6}$. Caseid-clustered robust variance. Penalizer fallback ladder: 0.00, 0.01, 0.10, 1.00 if convergence fails. |
| H3 | Pre-intent channel shares (LLM, classic search, retailer/marketplace, direct brand, review/UGC, other) | Fractional logit (GLM Binomial, logit link) | One model per share outcome. Same unadjusted/adjusted covariate pattern as H1 (exposure-state contrast). $A_{pre}$ decomposition is reported separately to avoid journey-versus-user conflation. Caseid-clustered SE. No user-specified starting values. |
| H4 | Transition deltas ($C \to I$, $C \to V$, $C \to C$) | Caseid bootstrap | Differences computed as (LLM-exposed minus non-exposed) transition probabilities from consideration. Main run uses 500 bootstrap draws for 95% percentile CIs. |
| H5 | Breadth (unique pre-intent sources) | Poisson GLM (log link) | Unadjusted and adjusted specifications parallel H1. Caseid-clustered SE. IRR reported for adjusted model. |
| H6 | Survey alignment (MDSE, CMEDIA, NEWS) | Weighted least squares | Case-level models with provided survey `weight`. Unadjusted: `outcome` $\sim$ `llm_exposed`. Adjusted: adds baseline activity and available demographics. |
| H7 | Moderation (LLM $\times$ MDSE on compression) | Fractional logit interaction model | Model includes `llm_exposed`, `mdse_mean`, interaction term, and baseline activity controls. Caseid-clustered SE for interaction coefficient. |

*Note:* Main journey segmentation parameters are inactivity gap = 24 hours and maximum journey cap = 168 hours. The summary sensitivity table appears in Table 5; the full segmentation grid is reported in Table 10.

Table 10: Segmentation sensitivity grid (unadjusted).

| Segmentation | H1 compression (AME) | H2 intent (HR) | H4 $C \to I$ diff. |
|---|---|---|---|
| Gap 10m; cap 6h | 0.002 | 1.116 | 0.052 |
| Gap 10m; cap 24h | 0.002 | 1.117 | 0.052 |
| Gap 10m; cap 72h | 0.002 | 1.117 | 0.052 |
| Gap 30m; cap 6h | 0.001 | 1.127 | 0.053 |
| Gap 30m; cap 24h | 0.001 | 1.114 | 0.053 |
| Gap 30m; cap 72h | 0.001 | 1.114 | 0.053 |
| Gap 60m; cap 6h | 0.000 | 1.185 | 0.053 |
| Gap 60m; cap 24h | 0.000 | 1.147 | 0.053 |
| Gap 60m; cap 72h | 0.000 | 1.149 | 0.053 |

Notes: Each row re-segments journeys with a different inactivity gap and maximum journey cap, then re-estimates unadjusted H1 and H2 models and recomputes the $C \to I$ transition difference.

Table 11: Within-user augmentation contrasts among exposure-positive users with both $G1$ and $G2$ journeys.

| Outcome | Paired difference | 95% CI | Users |
|---|---|---|---|
| Compression (count) | -0.006 | [-0.015, 0.000] | 91 |
| Compression (time) | -0.155 | [-0.236, -0.085] | 91 |
| Pre-intent LLM share | 0.030 | [0.022, 0.038] | 91 |
| Pre-intent review/UGC share | 0.013 | [-0.013, 0.040] | 91 |
| Breadth (unique pre-intent sources) | 0.044 | [0.012, 0.083] | 91 |

Notes: Paired differences are computed as user-level mean($G2$) − mean($G1$), with bootstrap confidence intervals over caseids. Positive values indicate larger metrics in augmented journeys.