

COINBench: Moving Beyond Individual Perspectives to Collective Intent Understanding

Anonymous ACL submission

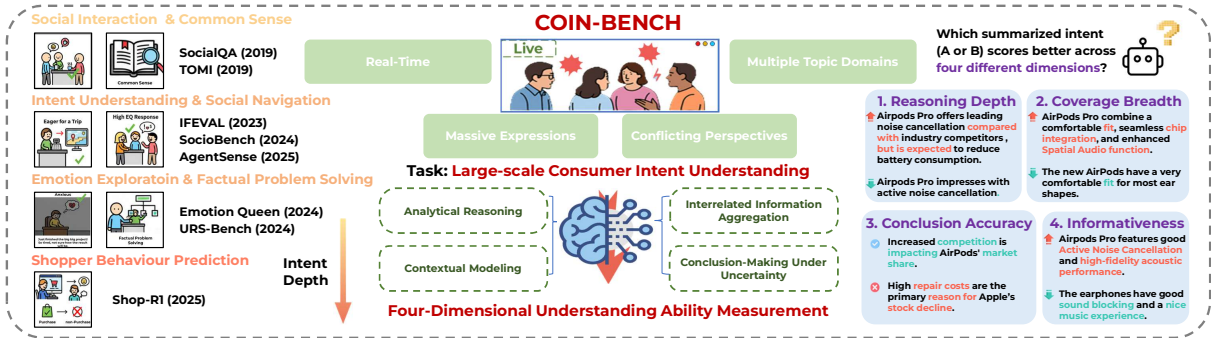


Figure 1: **Motivation for COIN-BENCH.** Existing benchmarks predominantly rely on individual perspectives and synthetic data, which fail to assess model performance in deciphering complex, real-world collective intent. We propose COIN-BENCH, a large-scale, live-updating benchmark that evaluates collective intent understanding across four key dimensions: depth, breadth, accuracy, and informativeness.

Abstract

Understanding human intent is a high-level cognitive challenge for Large Language Models (LLMs), requiring sophisticated reasoning over noisy, conflicting, and non-linear discourse. While LLMs excel at following individual instructions, their ability to distill **Collective Intent**—the process of extracting consensus, resolving contradictions, and inferring latent trends from multi-source public discussions—remains largely unexplored. To bridge this gap, we introduce COIN-BENCH, a dynamic, real-world, live-updating benchmark specifically designed to evaluate LLMs on collective intent understanding within the consumer domain. Unlike traditional benchmarks that focus on transactional outcomes, COIN-BENCH operationalizes intent as a **hierarchical cognitive structure** (from explicit scenarios to deep causal reasoning). We implement a robust evaluation pipeline that synergizes rule-based method with an LLM-as-the-Judge approach. This framework incorporates COIN-TREE for hierarchical cognitive structuring and retrieval-augmented verification (COIN-RAG) to ensure expert-level precision in analyzing raw, collective human discussions. An extensive evaluation of 20 state-of-the-art LLMs across four dimensions—*depth, breadth, informativeness, and correctness*—reveals that while current models handle surface-level aggregation, they struggle with the analytical depth required for complex intent synthesis. COIN-BENCH

establishes a new standard for advancing LLMs from passive instruction-followers to expert-level analytical agents capable of deciphering the collective voice of the real world.

1 Introduction

The emergence of Large Language Models (LLMs) (OpenAI, 2025; Grattafiori et al., 2024; Guo et al., 2025) has catalyzed a fundamental paradigm shift in artificial intelligence, moving beyond mere text generation toward sophisticated reasoning and real-world problem-solving (Team, 2025). Current state-of-the-art models exhibit remarkable proficiency in structured, task-oriented environments, as evidenced by their performance on benchmarks such as SWE-bench (Jimenez et al., 2024a) for software engineering, SPIDER2.0 (Lei et al., 2024) for SQL generation, and GAIA (Mialon et al., 2023) for tool-augmented reasoning. However, these evaluations predominantly focus on the model as a *reasoning agent* interacting with an *environment*, leaving a pivotal dimension of intelligence largely unexplored: understanding collective human intent. This task requires the sophisticated ability to distill coherent consensus from the noisy, fragmented, and often conflicting perspectives inherent in a population.

Real-world public discourse, such as community debates or consumer feedback, is characterized by "swarm-like" intelligence. Unlike linear

instructions, these discussions involve an intricate interplay of divergent goals, emotional nuances, and implicit assumptions. To truly comprehend this environment, an LLM must go beyond parsing individual sentences; it must synthesize multi-source signals, resolve contradictions, and infer latent trends within an evolving discourse. This capability—acting as an expert-level analyst capable of navigating complex public opinion—represents a high-level cognitive frontier that current evaluation frameworks fail to capture.

To rigorously investigate this capability, we identify the **consumer domain** as an ideal microcosm for **Collective Intent**. Consumer discussions are far more than transactional records; they embody grounded yet complex human reasoning, rich in the noise and redundancy inherent to collective intelligence. While prior benchmarks have explored social or emotional dimensions (Sap et al., 2019; Chen et al., 2024a; Mou et al., 2025), they often rely on hand-crafted or semi-synthetic datasets that lack the "raw noise" of real-world discourse. Furthermore, recent consumer-focused benchmarks like Shop-R1 (Zhang et al., 2025) primarily treat intent as a surface-level prediction of shopper actions (e.g., purchase vs. non-purchase). We argue that such transactional outcomes overlook the deeper **Collective Intent**—the multi-layered cognitive structure distilled from a diverse population. There is a pressing need for a benchmark that shifts the focus from simple action prediction to the hierarchical synthesis of collective human voices.

To bridge this gap, we introduce COIN-BENCH (Collective Intent Benchmark), a large-scale, dynamic, and live-updating framework designed to evaluate LLMs on their synthesis of real-world collective consumer intent. COIN-BENCH encompasses over **200k opinions** across 9 major domains and 54 sub-categories, spanning 1,400+ products. Crucially, we depart from passive evaluation (e.g., multiple-choice questions) in favor of an **Active Probing Paradigm**. We task LLMs with acting as "meta-analysts" to reconstruct chaotic public discourse into structured questionnaires. This paradigm is underpinned by three core cognitive pillars:

- **Consensus Extraction:** Designing valid questionnaires requires models to implicitly perform denoising and clustering, identifying salient topics amidst the noise.
- **COIN-TREE for Hierarchical Resolution:**

Collective intent is inherently stratified. We map model outputs onto a hierarchical cognitive topology to strictly distinguish between surface-level aggregation (L1–L3) and deep causal reasoning or trend prediction (L4–L5).

- **COIN-RAG for Grounded Verification:** To ensure correctness, we employ a retrieval-augmented verification pipeline that penalizes hallucinations by ensuring all inferred consensus is strictly grounded in the raw discussion context.

Evaluating 20 diverse LLMs reveals a critical performance dichotomy: while reasoning-enhanced models (e.g., OpenAI o1/o3) demonstrate superior depth in penetrating complex cognitive layers, a significant gap remains in their ability to achieve broad and expert-level analytical synthesis. Our findings suggest that even the most advanced models struggle with the multifaceted nature of collective intent, highlighting substantial room for future development.

In summary, our contributions are as follows:

- **A Large-Scale, Live Benchmark:** We introduce COIN-BENCH, featuring 200k+ real-world consumer discussions across 1,400+ products. Its dynamic nature ensures real-time updates and mitigates data contamination.
- **Robust Active Evaluation Pipeline:** We propose the Active Probing Paradigm, which transforms traditional passive structural questioning into a process of model-led self-exploration and autonomous Q&A. By integrating the COIN-TREE with the COIN-RAG verification framework, this pipeline ensures high-precision assessment, fundamentally shifting the evaluation focus from passive response generation to the active, structured reconstruction of collective intent.
- **Extensive Empirical Analysis:** We provide a comprehensive evaluation of 20 leading LLMs, revealing fundamental limitations in current models' ability to synthesize deep, multi-source human intent.

2 BENCHMARK CONSTRUCTION

2.1 Data Curation

Constructing a benchmark for collective intent requires a corpus that reflects the noisy, fragmented, and contradictory nature of real-world public discourse. We established an automated curation

pipeline designed to retrieve high-density discussions while filtering out irrelevant noise. The process consists of three integrated stages:

Hybrid Acquisition (Search & Retrieval). To maximize coverage, we employ a hybrid retrieval strategy combining semantic vector search with LLM-augmented API queries. For a given focal product (seed entity), the system first generates context-aware keywords using an LLM to expand the search space. It then retrieves discussion threads from diverse open-source web platforms. To ensure semantic relevance, we utilize vector embeddings to measure similarity between the seed topic and retrieved content, retaining only threads surpassing a strict similarity threshold.

Denosing and Filtering. Raw web data is inherently unstructured. We apply a cascade of quality filters: (1) *Rule-based Filtering* removes low-information content (e.g., extremely short texts < 20 chars); (2) *Semantic Filtering* employs an LLM-based verifier to exclude discussions that drift from the core topic; (3) *Temporal Alignment* prioritizes recent discussions to ensure the intent reflects current market realities. This pipeline ensures the final dataset preserves the complexity of swarm intelligence while maintaining high contextual fidelity.

2.2 Data Statistics

As detailed in Table 1 and Figure 3, COIN-BENCH establishes a robust foundation for evaluating collective intent. The benchmark spans 9 primary domains and 54 sub-categories, covering over 1,400 focal products. Crucially, unlike datasets that focus on sparse queries, we enforce high information density: each product entry aggregates an average of 200 user comments, totaling over 200k individual opinions. This scale is essential for testing an LLM’s ability to synthesize a "many-to-one" consensus from a large volume of diverse signals, detailed information in Appendix A.3.

2.3 Evaluation Dimensions

To rigorously quantify the LLM’s cognitive capability in collective intent synthesis, we categorize evaluation into four orthogonal dimensions:

Depth (Cognitive Penetration). This dimension measures the level of reasoning required to derive the intent. We introduce a five-level taxonomy (COIN-TREE levels):

- **Observational Understanding (L1–L3):** The model identifies explicit information directly

present in the text, such as *Usage Scenarios* (L1), *Product Aspects* (L2), and *User Feelings* (L3).

- **Inferential Reasoning (L4–L5):** The model must synthesize dispersed information to perform advanced reasoning. This includes *Comparative Analysis* against competitors (L4) and *Predictive Tendency* regarding future improvements or market shifts (L5).

A high depth score indicates the model moves beyond surface summarization to expert-level insight.

Breadth (Semantic Coverage). Collective intent is multifaceted. This dimension evaluates the model’s completeness in capturing the full spectrum of the crowd’s voice. A high breadth score signifies that the model has successfully identified the "long tail" of user concerns—covering diverse usage scenarios and edge cases—rather than fixating on a single dominant topic.

Correctness (Grounded Fidelity). In swarm analysis, hallucination is the primary failure mode. This dimension evaluates whether the inferred intent is strictly grounded in the provided discussions. Using our COIN-RAG pipeline, we measure the factual alignment between the model’s generated consensus and the raw crowd evidence, ensuring the output is free from bias and unverified extrapolation.

Informativeness (Expressive Diversity). An expert analyst should articulate insights with nuance. We assess this through *Lexical Richness* (vocabulary range) and *Semantic Distinctiveness* (lack of redundancy). High performance here indicates the model avoids generic boilerplate responses ("The product is good") and instead employs precise, varied language to capture the specific texture of the consumer intent.

3 Methodology

Evaluating **Collective Intent** presents a unique challenge: unlike factual QA, there is no single 'gold standard' sentence that captures the entirety of a chaotic public discussion. Passive evaluation methods (e.g., multiple-choice selection) fail to test whether a model can actively synthesize fragmented views into a coherent whole.

To address this, we operationalize evaluation through an **Active Probing Paradigm**. We task the LLM with acting as a 'Meta-Analyst' requiring it to **reconstruct** the unstructured crowd feedback into a structured **Questionnaire**. This design relies on the cognitive premise that to formulate a

Table 1: Comparison with existing related benchmarks. "Real-world" indicates whether the data is sourced from real-world scenarios rather than synthetic or online existing resources. "Live Update" denotes whether the benchmark can be regularly updated.

Benchmark	Domain	Tasks	Real World	Live Update
DABstep (Egg et al., 2025)	Data Science	450	✓	✗
FutureX (Zeng et al., 2025)	Future Prediction	500/week	✓	✓
GAIA (Mialon et al., 2023)	General QA	466	✓	✗
OSWorld (Xie et al., 2024)	Computer Use	369	✓	✗
OPT-Bench (Li et al., 2025a)	Iterative Optimization	30	✓	✗
Spider2.0 (Lei et al., 2024)	Text-to-SQL	632	✓	✗
SWE-Bench (Jimenez et al., 2024b)	Code	2,294	✓	✗
SociaBench (Chen et al., 2024a)	Social Intent	6,000	✗	✗
URS-bench (Wang et al., 2024)	Intent Understanding	1,846	✗	✗
COIN-BENCH (ours)	Consumer Intent	1,475	✓	✓

valid question-answer pair from noise, the model must successfully execute two high-level reasoning steps:

- **Salience Identification (The Question):** By formulating a relevant question stem, the model demonstrates it can identify meaningful topics amidst noise.
- **Consensus Aggregation (The Answer):** By designing options that reflect the majority perspective, the model demonstrates a precise grasp of intents within diverse opinion clusters.

To rigorously assess the quality of this reconstructed intent, we propose a robust evaluation framework combining topological analysis (COIN-TREE) and evidence-based verification (COIN-RAG), as illustrated in Figure 2.

3.1 COIN-TREE: Hierarchical Cognitive Stratification

Moving beyond shallow keyword matching, COIN-TREE provides a **hierarchical intent graph** derived from real-world discussions to measure deep understanding. It allows to quantify the *Hierarchic* of the model’s understanding—distinguishing between surface-level observation and deep reasoning.

Construction. We operationalize collective intent as a five-level weighted hierarchical tree (COIN-TREE), where the root denotes the focal product. The construction process involves two key stages: (i) **Branch Extraction:** We leverage an LLM to distill raw, unstructured discourse into standardized semantic branches based on a cognitive template (ranging from *Usage Scenarios* to *Future Tendencies*); (ii) **Node Aggregation:** Semantically redun-

dant nodes across disparate branches are consolidated into high-weight nodes. The final weights are computed by integrating the frequency of node occurrences with real-world engagement metrics, such as view counts and upvotes, to reflect the popularity and consensus of specific viewpoints. Further technical details and the formal extraction prompt are provided in Figure 2 and Appendix A.4.

Surface Understanding (L1–L3): Nodes at these levels capture explicit content directly observable in user discussions, including *Usage Scenarios* (L1), specific *Aspects* (L2), and *User Feelings* (L3) of the product.

Deep Reasoning (L4–L5): These levels require the model to synthesize dispersed information. L4 (Comparison) represents comparative reasoning against competing products, while L5 (Tendency) represents predictive insights into future product improvements or market shifts.

Evaluation (Lighten Tree Nodes). We evaluate the Depth and Breadth of the LLM by projecting its generated questionnaire onto the COIN-TREE nodes. Specifically, the content of each generated question is extracted into a semantic branch and matched shallow-deep against the COIN-TREE nodes using a Sentence Transformer. Successful matches ‘lighten’ the corresponding nodes. The lightened sub-tree illustrates how well the model understands this human intent.

Depth Score: Calculated as the ratio of the total weight of lightened nodes at a specific level (L1–L5) to the total weight of all nodes at that level. The overall depth is the average all five levels. A high score at L4/L5 indicates the model penetrated the “deep structure” of collective intent.

Breadth Score: Calculated as the sum of weights of all lightened nodes. This reflects the model’s

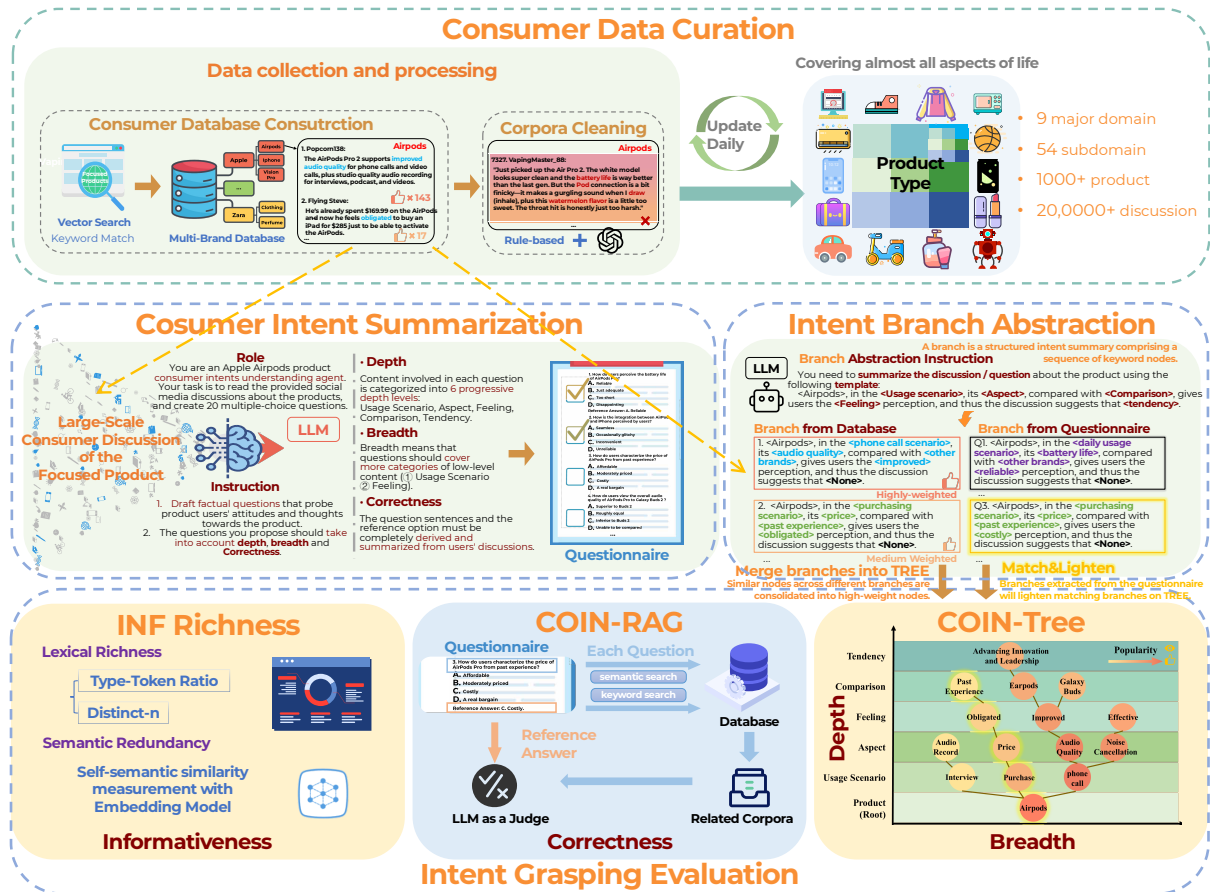


Figure 2: **COIN-BENCH Pipeline: From Raw Discourse to Structured Evaluation.** **Top (Curation):** An automated pipeline that consolidates 200k+ public opinions into a consumer database via hybrid keyword–semantic retrieval and dual-stage rule/LLM denoising. **Middle (Active Synthesis):** The *Active Probing* stage, where LLMs act as meta-analysts to convert fragmented discourse into structured questionnaires. **Bottom (Evaluation):** Performance is assessed along four dimensions: **Depth** and **Breadth** via **COIN-TREE** (a five-level hierarchical intent map), **Correctness** via **COIN-RAG** (grounded corpus verification), and **Informativeness** (lexical and semantic richness).

coverage of the multifaceted nature of the discussion.

3.2 COIN-RAG: Evaluation of Grounded Correctness

In the context of swarm intelligence, ‘correctness’ implies fidelity to the crowd’s consensus rather than external facts. To prevent the evaluation from favoring eloquent but hallucinated insights, we implement COIN-RAG, a retrieval-augmented verification pipeline acting as a “Fact-Checker.”

Construction. COIN-RAG serves to ground the LLM’s inferred intent in the raw evidence. We employ a two-stage retrieval process:

Dual-Embedding: User discussions are vectorized using both TF-IDF (for precise keyword matching) and all-MiniLM-L6-v2 (for semantic nuance). This ensures retrieval captures both specific terminology and implicit sentiments.

Joint Retrieval: Key opinions extracted from the LLM’s generated questionnaire are queried against

the vector database to retrieve the top- k most relevant raw discussion entries.

Evaluation. We employ a reasoning-based verification step. The retrieved raw discussions (Ground Truth Evidence) are compared against the LLM’s generated intents (which are the question stems and given answers in the questionnaires). If the LLM’s synthesized consensus aligns with the majority perspective found in the retrieved evidence, it is deemed correct. This pipeline strictly penalizes hallucinations that drift away from the authentic collective voice.

3.3 Informativeness: Assessing Analytical Density and Diversity

Beyond structural accuracy, expert-level intent synthesis must be both information-dense and analytically diverse. We evaluate the **Informativeness** of generated questionnaires to ensure that models capture the full nuance of collective discourse rather than producing generic or repetitive text.

Lexical Richness To reflect diverse population expression, models must employ precise and domain-relevant vocabulary. We measure lexical variety using the **Type-Token Ratio (TTR)** (Johnson, 1944) and phrase-level diversity using **Distinct- n** (Li et al., 2016). Higher scores indicate effective abstraction of raw crowd language into richer analytical expression.

Semantic Redundancy A high-quality synthesis should maximize semantic coverage rather than concentrate on a narrow theme. We quantify redundancy by computing the average maximum cosine similarity between embeddings of generated question pairs (Chen et al., 2021b). Higher redundancy indicates limited analytical breadth, where the model repeatedly focuses on a single topic while overlooking other latent concerns and divergent perspectives. Detailed calculations are provided in Appendix A.5.

4 Experiment

4.1 Experimental Setup

We evaluated our method on a diverse range of reasoning and general LLMs. This includes proprietary models (GPT family, Claude) accessed via API, and open-source models (Qwen 1.5B–72B, LLaMA, DeepSeek, InternLM) deployed locally via LMDeploy.

4.2 Main Results

Table 2 presents the evaluation of 20 leading LLMs on COINBENCH. Our results reveal a critical performance dichotomy between surface-level information aggregation and deep intent synthesis.

Hierarchy of Intent. We observe a pronounced performance decay as we move from explicit to implicit intent layers within COIN-TREE. While most models maintain reasonable scores in capturing *Usage Scenarios* (L1), their performance collapses when tasked with inferring *Future Tendencies* (L5). This "depth cliff" validates the hierarchical design of our benchmark and highlights a fundamental reasoning bottleneck: models can easily parse what is explicitly stated but struggle to synthesize what is implicitly anticipated by a collective population.

Reasoning vs. General Models. Reasoning models, particularly **GPT-o3**, demonstrate a distinct superiority in the deepest cognitive layers (L5) and **Correctness** (80.35). Unlike general-purpose models that often aggregate noise, **GPT-o3** exhibits

"critical deliberation"—the ability to filter out the vocal minority and anchor its synthesis in true majority consensus. This result suggests that navigating swarm-like discourse requires not just memory, but the ability to perform causal inference and temporal projection to resolve contradictions within the data.

Breadth vs. Precision. An intriguing pattern emerges with large-scale open-source models like **Qwen3-30B-A3B**. While it excels in **Breadth** (53.20) and surface-level recall (L1), it suffers from high **Semantic Redundancy** (68.94) and lower **Correctness**. This indicates a "Verbosity Trap": the model achieves high coverage by exhaustively listing possible viewpoints (brute-force recall) rather than distilling them into a concise, high-entropy consensus. In contrast, **GPT-5** demonstrates superior semantic span, capturing the "long-tail" of niche concerns while maintaining higher analytical density.

Analytical Density and Information Richness. Our **Informativeness** metrics reveal that **GPT-o3** achieves the highest lexical richness (85.52) and the lowest semantic redundancy (52.27). This implies that reasoning models actively restructure raw discussion points into a "distilled" analytical form, whereas general models often "parrot" dominant topics, leading to repetitive and diluted outputs. Furthermore, a significant performance gap remains between proprietary models and open-source models, particularly in **Correctness**, where even the strongest open-source models (e.g., Qwen3-32B) trail behind the GPT-4/5 series by nearly 15-20%.

Scaling Effects. A clear scaling law is observed: models with fewer than 7B parameters (e.g., Qwen2.5-1.5B, DS-7B) fail significantly across all dimensions, particularly in **Correctness** and **Breadth**. This underscores that understanding collective intent is an emergent capability that requires a threshold of world knowledge and reasoning capacity to distinguish signal from noise in real-world public discourse.

4.3 Ablation Study: The Trade-off Between Structure and Noise

The COIN-TREE distills raw discussions into nodes weighted by topic popularity, effectively highlighting mainstream discourse while marginalizing peripheral noise. To investigate the impact of

Table 2: Performance of reasoning LLMs, general LLMs, and open-source LLMs on COIN-BENCH, with the best performance highlighted in **bold**.

Model	Depth						Breadth	Informativeness		Correctness
	L1	L2	L3	L4	L5	Overall		Lexical	Semantic↓	
<i>Proprietary LLMs</i>										
GPT-5	18.29	25.81	6.25	3.49	0.06	10.78	53.48	80.21	62.75	62.65
GPT-4.1	20.97	25.03	5.90	3.37	0	11.01	53.41	79.07	63.82	59.05
GPT-4o	20.99	23.70	4.90	2.58	0.05	10.44	52.95	79.56	62.86	75.75
Claude-3.5-sonnet	19.25	23.25	5.34	2.95	0	10.16	52.83	73.94	61.11	53.35
GPT-o3	16.17	22.43	5.69	3.18	0.07	9.51	52.73	85.52	52.27	80.35
<i>Open-Source LLMs</i>										
Qwen3-30B-A3B	25.01	23.66	5.43	2.51	0.06	11.33	53.20	70.58	68.94	61.60
DS-Distill-Qwen-14B	17.00	25.56	6.12	3.30	0	10.40	53.32	67.47	75.53	58.45
Qwen2.5-32B-Instrcut	19.26	23.63	5.31	2.89	0.01	10.21	52.46	65.72	74.60	54.95
Qwen3-32B	20.77	23.46	5.57	2.82	0	10.52	51.15	65.84	68.16	55.26
Qwen3-8B	15.95	22.43	4.89	2.39	0.01	9.13	50.58	57.51	81.25	50.42
Qwen2.5-72B-Instrcut	18.89	22.27	5.73	3.10	0	10.00	50.52	54.63	77.49	64.11
DS-Distill-Qwen-32B	16.60	24.56	5.90	3.30	0	10.07	50.34	59.30	76.58	53.90
Qwen2.5-14B-Instrcut	13.56	22.23	5.45	2.87	0.02	8.83	48.27	52.39	80.06	60.88
LLama3.2-8B-Instrcut	13.88	19.75	5.62	2.73	0	8.40	47.91	47.87	88.25	52.31
Qwen2.5-7B-Instrcut	11.87	19.73	4.16	1.97	0	7.54	47.43	43.58	85.07	49.24
Internlm3-8B-Instrcut	11.07	20.76	4.87	2.61	0.03	7.87	45.91	49.83	75.51	51.67
LLama3.1-8B-Instrcut	11.23	19.46	5.53	2.91	0	7.83	45.41	42.36	88.00	52.67
Qwen2.5-3B-Instrcut	13.49	18.63	4.22	2.09	0	7.69	42.73	39.32	79.35	35.43
Qwen2.5-1.5B-Instrcut	2.83	4.94	0.99	0.45	0	1.84	14.31	4.56	87.65	36.90
DS-Distill-Qwen-7B	1.80	4.91	1.35	0.55	0	1.72	11.54	3.30	73.25	13.40

Table 3: Comparison of reasoning LLMs, general LLMs, and open-source LLMs using COIN-TREE on COIN-BENCH, with the best performance highlighted in **bold**.

Model	Depth						Breadth	Informativeness		Correctness
	L1	L2	L3	L4	L5	Overall		Lexical	Semantic↓	
GPT-o3 (wo/Tree)	16.17	22.43	5.69	3.18	0.07	9.51	52.73	85.52	52.27	80.35
GPT-o3 (w/Tree)	45.13	36	15	11.74	1.37	21.95	59.16	72.47	71.05	57.10
	(+28.96)	(+13.57)	(+9.31)	(+8.56)	(+1.30)	(+12.44)	(+6.43)	(-13.05)	(-0.22)	(-23.25)
GPT-4o (wo/Tree)	20.99	23.70	4.90	2.58	0.05	10.44	52.95	79.56	62.86	75.60
GPT-4o (w/Tree)	37.38	31.39	12.78	9.60	0.85	18.40	57.00	70.77	72.36	64.15
	(+16.39)	(+7.69)	(+7.88)	(+7.02)	(+0.80)	(+7.96)	(+4.05)	(-8.79)	(-0.50)	(-11.45)
Qwen2.5-7B (wo/Tree)	11.87	19.73	4.16	1.97	0	7.54	47.43	43.58	85.07	42.15
Qwen2.5-7B (w/Tree)	32.39	29.78	11.29	7.96	0.54	16.39	49.28	42.33	79.63	49.24
	(+20.52)	(+10.05)	(+7.13)	(+6.00)	(+0.54)	(+8.85)	(+1.85)	(-1.25)	(-5.44)	(+7.09)

raw swarm noise on model performance, we conducted an ablation study where models were evaluated using the structured COIN-TREE as input instead of raw discussions. The results in Table 3 reveal a fascinating Structure-Noise Trade-off:

Structure as a Scaffold for Small Models: For smaller open-source models (e.g., Qwen2.5-7B), using the COIN-TREE significantly improves *Correctness*. Raw noise often overwhelms these noise-intolerant models. The COIN-TREE acts as a scaffold that clarifies core intents, revealing both the distribution and logical progression with those intents.

Structure as a Bottleneck for SOTA Models. For Proprietary models such as GPT-4o, apply-

ing COIN-TREE leads to declines in *Correctness* and *Informativeness*. Although the tree aids denoising, it is inherently *lossy* for SOTA models: while it highlights core intents, it suppresses the subtle nuances they can naturally recover from raw discourse. This suggests that expert-level intent understanding depends on navigating signal within noise, rather than relying solely on pre-structured summaries.

4.4 Case Study: Quality vs Quantity

The quantitative results hinted at a discrepancy: **Qwen3-30B-A3B** outperforms **GPT-5** at the shallowest level (L1) but lags significantly in deeper reasoning (L2-L4), breadth and informativeness.

We further analyze this via a case study on the 'Google Nest' discussion (detailed in Appendix A.6-A.7).

A fine-grained qualitative analysis reveals a clear divergence in the **analytical resolution** of collective intent. Advanced reasoning models such as **GPT-5** and **GPT-o3** exhibit strong **structural inference**, uncovering high-level latent concerns—e.g., “*Matter protocol interoperability*”—that go beyond surface keywords. Rather than merely aggregating terms, these models perform causal mapping to identify the underlying logic of user sentiment.

In contrast, **Qwen3-30B-A3B** often shows **semantic dilution**, generating verbose but redundant outputs (e.g., repeated variants of “*Audio Fidelity*,” “*Clarity*,” and “*Sound Quality*”) with limited taxonomic precision. While its MoE architecture supports broad recall, it lacks the **inferential rigor** for **discriminative distillation**—i.e., compressing diverse signals into mutually exclusive and collectively exhaustive insights. These findings highlight that understanding collective intent requires not just output breadth but high **analytical density** and structured synthesis. Expert-level models must demonstrate both depth and breadth in reasoning.

5 Related Work

LLM Evaluation The rapid progress of Large Language Models (LLMs) has led to a proliferation of benchmarks for evaluating their generalization and reasoning abilities. Early benchmarks such as MMLU (Hendrycks et al., 2020) and BIG-bench (Srivastava et al., 2022) assessed broad knowledge and reasoning. Later efforts targeted specific domains, including linguistic and commonsense reasoning (Wang et al., 2018, 2019; Talmor et al., 2019; Zellers et al., 2019; Lin et al., 2022), mathematical and programming tasks (Hendrycks et al., 2021; Cobbe et al., 2021; Chen et al., 2021a; Austin et al., 2021), and agent-based evaluations (Chan et al., 2024; Xie et al., 2024; Li et al., 2025b,a). Despite these advances, existing benchmarks rarely examine whether LLMs can understand human intent in dynamic, real-world decision-making—particularly in multi-user, emotionally nuanced, and evolving contexts. To address this, we propose COIN-BENCH, a large-scale benchmark designed to assess LLMs’ ability to comprehend and reason about human intentions in complex real-world scenarios.

LLM Human Intent Evaluation Human intent evaluation has increasingly shifted toward understanding human-centric intent in complex, real-world contexts. Prior benchmarks capture different facets of this problem: SocialIQA (Sap et al., 2019) focuses on social and commonsense intent, TOMI (Le et al., 2019) evaluates Theory of Mind capabilities, and IFEVAL (Zhou et al., 2023) measures instruction-following. Other efforts such as SociaBench (Chen et al., 2024a), AgentSense (Mou et al., 2025), EmotionQueen (Chen et al., 2024b), and URS-bench (Wang et al., 2024) examine intent understanding from social, emotional, or user-response perspectives. However, existing frameworks typically isolate single dimensions of intent, whereas real-world intent is multifaceted and dynamic, combining social, emotional, and practical factors. Consequently, there remains no comprehensive benchmark for evaluating whether LLMs can fully understand human reasoning and mental states. To address this gap, we propose COIN-BENCH, a benchmark for assessing LLMs’ ability to understand complex, dynamic real-world human intent.

6 Conclusion

In this work, we propose COIN-BENCH, a comprehensive benchmark consisting of over 200k product-level discussions across 9 major domains, 54 sub-domains, and over 1,400 products, designed to evaluate the performance of Large Language Models (LLMs) in understanding real-world human intent, particularly within consumer domains. Our evaluation framework measures LLMs’ ability to comprehend intent across four key dimensions: depth, breadth, correctness, and informativeness. We implement a robust evaluation pipeline to mitigate bias and hallucinations. Specifically, we construct COIN-TREE to assess LLMs’ depth and breadth of intent understanding, use COIN-RAG for evaluating correctness, and measure informativeness through lexical diversity and semantic richness. Through extensive experiments on both closed-source and open-source models, we demonstrate that reasoning models outperform general models on average. However, significant gaps remain between closed-source and open-source models, and even the most advanced models struggle with deep and broad intent understanding. Our mission is to advance LLMs toward expert-level reasoning and improve their ability to understand complex real-world intent.

600 Limitations

601 Due to resource constraints, we were unable to
602 include additional state-of-the-art models such as
603 the Gemini 3 series, OpenAI GPT-5.2, and Claude
604 4.5 in our experiments.

605 References

606 Jacob Austin, Augustus Odena, Maxwell Nye, Maarten
607 Bosma, Henry Michalewski, David Dohan, Ellen
608 Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1
609 others. 2021. Program synthesis with large language
610 models. *arXiv preprint arXiv:2108.07732*.

611 Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James
612 Aung, Dane Sherburn, Evan Mays, Giulio Starace,
613 Kevin Liu, Leon Maksin, Tejal Patwardhan, and 1
614 others. 2024. Mle-bench: Evaluating machine learn-
615 ing agents on machine learning engineering. *arXiv
616 preprint arXiv:2410.07095*.

617 Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen
618 Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chen-
619 liang Li, Ji Zhang, and Fei Huang. 2024a. Social-
620 bench: Sociality evaluation of role-playing conversa-
621 tional agents. In *ACL (Findings)*.

622 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan,
623 Henrique de Oliveira Pinto, Jared Kaplan, Harri Ed-
624 wards, Yuri Burda, Nicholas Joseph, Greg Brock-
625 man, and 1 others. 2021a. Evaluating large lan-
626 guage models trained on code. *arXiv preprint
627 arXiv:2107.03374*.

628 Wang Chen, Piji Li, and Irwin King. 2021b. [A training-
629 free and reference-free summarization evaluation
630 metric via centrality-weighted relevance and self-
631 referenced redundancy](#). In *Proceedings of the 59th
632 Annual Meeting of the Association for Computational
633 Linguistics and the 11th International Joint Confer-
634 ence on Natural Language Processing (Volume 1:
635 Long Papers)*, pages 404–414. Association for Com-
636 putational Linguistics.

637 Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu,
638 Yueze Li, Yi Zhao, and Yanghua Xiao. 2024b. [Emo-
639 tionqueen: A benchmark for evaluating empathy of
640 large language models](#). *Preprint*, arXiv:2409.13359.

641 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,
642 Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias
643 Plappert, Jerry Tworek, Jacob Hilton, Reiichiro
644 Nakano, and 1 others. 2021. Training verifiers
645 to solve math word problems. *arXiv preprint
646 arXiv:2110.14168*.

647 Alex Egg, Martin Iglesias Goyanes, Friso Kingma, An-
648 dreu Mora, Leandro von Werra, and Thomas Wolf.
649 2025. Dabstep: Data agent benchmark for multi-step
650 reasoning. *arXiv preprint arXiv:2506.23719*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,
651 Abhinav Pandey, Abhishek Kadian, Ahmad Al-
652 Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,
653 Alex Vaughan, and 1 others. 2024. The llama 3 herd
654 of models. *arXiv preprint arXiv:2407.21783*.
655

656 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao
657 Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shi-
658 rong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025.
659 Deepseek-r1: Incentivizing reasoning capability in
660 llms via reinforcement learning. *arXiv preprint
661 arXiv:2501.12948*.

662 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,
663 Mantas Mazeika, Dawn Song, and Jacob Steinhardt.
664 2020. Measuring massive multitask language under-
665 standing. *arXiv preprint arXiv:2009.03300*.

666 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul
667 Arora, Steven Basart, Eric Tang, Dawn Song, and Ja-
668 cob Steinhardt. 2021. Measuring mathematical prob-
669 lem solving with the math dataset. *arXiv preprint
670 arXiv:2103.03874*.

671 Carlos E Jimenez, John Yang, Alexander Wettig,
672 Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R
673 Narasimhan. 2024a. [SWE-bench: Can language
674 models resolve real-world github issues?](#) In *The
675 Twelfth International Conference on Learning Repre-
676 sentations*.

677 Carlos E Jimenez, John Yang, Alexander Wettig,
678 Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R
679 Narasimhan. 2024b. [Swe-bench: Can language mod-
680 els resolve real-world github issues?](#) In *ICLR*.

681 Wendell Johnson. 1944. The type-token ratio: A mea-
682 sure of lexical diversity. *Language*, 20(3):169–187.

683 Matthew Le, Y-Lan Boureau, and Maximilian Nickel.
684 2019. Revisiting the evaluation of theory of mind
685 through question answering. In *Proceedings of the
686 2019 Conference on Empirical Methods in Natu-
687 ral Language Processing and the 9th International
688 Joint Conference on Natural Language Processing
689 (EMNLP-IJCNLP)*, pages 5872–5877.

690 Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng
691 Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo,
692 Hongcheng Gao, Wenjing Hu, Pengcheng Yin, and 1
693 others. 2024. Spider 2.0: Evaluating language mod-
694 els on real-world enterprise text-to-sql workflows.
695 *arXiv preprint arXiv:2411.07763*.

696 Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao,
697 and Bill Dolan. 2016. A diversity-promoting ob-
698 jective function for neural conversation models. In
699 *Proceedings of NAACL-HLT*.

700 Xiaozhe Li, Jixuan Chen, Xinyu Fang, Shengyuan
701 Ding, Haodong Duan, Qingwen Liu, and Kai Chen.
702 2025a. Opt-bench: Evaluating llm agent on large-
703 scale search spaces optimization problems. *arXiv
704 preprint arXiv:2506.10764*.

- Xiaozhe Li, Xinyu Fang, Shengyuan Ding, Linyang Li, Haodong Duan, Qingwen Liu, and Kai Chen. 2025b. Np-engine: Empowering optimization reasoning in large language models with verifiable synthetic np problems. *arXiv preprint arXiv:2510.16476*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.
- Grégoire Mialon, Clémentine Fourier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.
- Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuan-Jing Huang, and 1 others. 2025. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4975–5001.
- OpenAI. 2025. [Openai o1 system card](#). Technical report, OpenAI.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialliqa: Commonsense reasoning about social interactions. In *Conference on Empirical Methods in Natural Language Processing*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.
- Tongyi DeepResearch Team. 2025. Tongyi-deepresearch. <https://github.com/Alibaba-NLP/DeepResearch>.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 353–355.
- Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. [A user-centric multi-intent benchmark for evaluating large language models](#). *Preprint*, arXiv:2404.13940.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.
- Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, and 12 others. 2025. [Futurex: An advanced live benchmark for llm agents in future prediction](#). *Preprint*, arXiv:2508.11987.
- Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruo Chen Jiao, Junze Liu, Kun Qian, Yuxin Tang, Ran Xue, Houyu Zhang, Qingjun Cui, Yufan Guo, and Dakuo Wang. 2025. [Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning](#). *Preprint*, arXiv:2507.17842.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.



Figure 3: Overview of COIN-BENCH: It includes over 200k product-level discussions across 9 major domains, 54 sub-domains, and more than 1,400 products.

A Appendix

A.1 Use of Large Language Models

Large Language Models are used for grammar check and polishing in this paper.

A.2 Data Statement

The COIN-BENCH dataset was constructed from public available websites, and all privacy-sensitive personal information has been removed during the data curation process. To mitigate the potential for technology misuse, the benchmark will be released under a restrictive license for academic research purposes only, explicitly prohibiting malicious applications.

A.3 Data Statistics

In this section, we provide a detailed statistical breakdown of the COIN-BENCH dataset. To ensure a comprehensive evaluation of collective intent across diverse human activities, we constructed a two-tier hierarchical taxonomy consisting of 9 primary domains and 54 fine-grained sub-categories, as visualized in figure 3.

Domain Distribution and Scale As illustrated in the inner ring of our taxonomy, COIN-BENCH covers the breadth of the modern consumer ecosystem. The dataset is anchored by major sectors including *Home and Daily Necessities* (284 products), *Consumer Electronics* (201 products), *Software* (199

products), and *Vehicles* (166 products). Other significant domains include *Personal Care and Beauty* (145), *Fashion and Accessories* (142), *AI Field* (125), *Sports and Fitness* (111), and *Food and Beverage* (102). This distribution ensures that LLMs are tested not only on general consumer goods but also on high-context, specialized fields such as AI technologies and industrial software.

Granular Sub-categories The outer ring of the sunburst chart highlights the high granularity of our benchmark. Each primary domain is further decomposed into specific sub-categories to capture nuanced usage scenarios and collective intents. For instance:

- **High-Tech & Digital:** Includes niche sectors like *AI Learning Machines*, *NLP/CV tools*, and *Programming/Development Tools* (the largest sub-category with 30 products).
- **Lifestyle & Essentials:** Spans from *Skin Care Products* and *Individual Care* to *Kitchen Appliances* and *Fresh Food*.
- **Mobility:** Covers the transition from traditional *Fuel Vehicles* to *Electric Vehicles* and *New Means of Transportation*.

Data Consistency and Density A key feature of COIN-BENCH is its balanced distributional design. Most sub-categories are populated with approximately 28 to 29 distinct product discussions, ensuring that the benchmark is not skewed toward any single niche. With an average of 200 user comments per product, the total corpus comprises over 200,000 unique human perspectives. This density provides a sufficiently "noisy" and "swarm-like" environment, forcing models to demonstrate advanced synthesis and denoising capabilities rather than simple pattern matching.

A.4 COIN-TREE construction details

First, the LLM (GPT-4o) is utilized to extract branches from each consumer discussion. During this process, the model is prompted to summarize a user's discussion following the template: ⟨Product_series⟩, in the ⟨Usage Scenario⟩, its ⟨Aspect⟩, compared with ⟨Comparison⟩, gives consumers the ⟨Feeling⟩ perception, and the discussion suggests that ⟨Tendency⟩.

Next, the branches are used to construct the tree. All branches are connected to the tree root node, forming an initial tree. Sentence Transformers are

then employed to merge semantically similar nodes layer by layer from the top down within the initial tree. During this process, nodes in the same layer that share the same parent node and are semantically similar are merged into one. The child nodes of each pre-merged node are then designated as the child nodes of the merged node. Additionally, the weight of the merged node is calculated as the sum of the weights of its child nodes. Ultimately, the fully merged tree is referred to as COIN-TREE, which will resemble the structure shown in figure X. Nodes with higher weights will appear in the shallow layers of the tree; this is because the core topics of discussion (e.g., usage scenarios, aspects, feelings) often overlap across discussions from different consumers, and such high-weight nodes represent the aspects of the product that users focus on most.

This process enables the clear presentation of user discussion content in a tree structure while highlighting discussion hotspots. Meanwhile, by updating the discussion data and reconstructing COIN-TREE, we can analyze changes in child nodes under the same parent node between the two trees—thereby identifying users’ immediate concerns and long-term strategic considerations.

After summarization, the key terms in the template are fetched to form a single branch. For the branches derived from one discussion, the initial weight of each node is equal, ranging from 1 to 3, and determined by the discussion’s upvotes and view count. Notably, not every discussion can be summarized to fill all six key terms—more successfully filled key terms correspond to a longer branch path, which in turn reflects a more in-depth consumer intent. Next, all branches are connected to the tree root node, forming an initial tree. Sentence Transformers are then employed to merge semantically similar nodes layer by layer from the top down. During this process, nodes in the same layer that share the same parent node and are semantically similar are merged into one. The child nodes of each pre-merged node are then designated as the child nodes of the merged node. Additionally, the weight of the merged node is calculated as the sum of the weights of its child nodes. Ultimately, the fully merged tree is referred to as COIN-TREE. Nodes with higher weights will appear in the shallow layers of the tree; this is because the core topics of discussion (e.g., usage scenarios, aspects, feelings) often overlap across discussions from different consumers.

Lighten the Tree: To assess the depth and breadth of a LLM’s understanding of consumer intent, content from questionnaires will be extracted into branches. Each branch will undergo semantic matching with the nodes in the COIN-TREE from top to bottom using a Sentence Transformer. Nodes that are successfully matched will be marked as "lightened," and the lightened nodes in the COIN-TREE will form a subtree. and the questionnaire will receive the score corresponding to that node.

Specifically, for each question in the questionnaire, the question stem and its four options will first be concatenated into four opinion statements. The LLM will then extract four branches from these four statements. These four branches will be matched with the nodes in the COIN-TREE from top to bottom—each branch will lighten a path and obtain a score based on the weights of the nodes along that path. The branch with the highest score for a given question will be used to "lighten" the COIN-TREE. Notably, nodes in the COIN-TREE cannot be repeatedly lightened by different questions. After iterating through all questions, the lightened nodes in the COIN-TREE will form a subtree.

A.5 Informativeness

Lexical Richness: The evaluation of lexical richness relies on two key metrics: Type-Token Ratio (TTR) (Johnson, 1944) and Distinct-n (Li et al., 2016). TTR quantifies the ratio of unique tokens to the total number of words in the text. It is defined as:

$$\text{TTR} = \frac{\text{Count}(\text{unique token})}{\text{Count}(\text{tokens})}$$

where a higher TTR indicates greater lexical richness. Distinct-n focuses on the n -gram level, measuring the ratio of unique n -grams to the total number of n -grams. This study focuses on bi -grams, and the Distinct-n is calculated as:

$$\text{Distinct-n} = \frac{\text{Count}(\text{unique bi-gram})}{\text{Count}(\text{bi-grams})}.$$

Semantic Redundancy is evaluated using a self-referential manner (Chen et al., 2021b), where the average maximum semantic similarity is computed between each question and all other questions in the questionnaire, as well as between each question’s options and all other questions’ options. Given a set of questions $Q = \{q_1, q_2, \dots, q_n\}$, the semantic similarity between any two questions q_i and q_j is calculated using cosine similarity:

$$\text{Sim}(q_i, q_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|},$$

where \mathbf{v}_i and \mathbf{v}_j represent the vector embeddings of questions q_i and q_j , respectively. The redundancy score is then computed as the average of the maximum similarity values across all pairs of questions:

$$\text{Redundancy} = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \text{Sim}(q_i, q_j).$$

Notably, a lower redundancy score indicates less repetition in question paradigms and option designs, which reflects the LLM's ability to understand consumers' intentions from multiple perspectives and conduct multi-source causal inference.

A.6 Case Study

A.6.1 Google Nest Smart Speaker questionnaire from GPT-5

1. How do you primarily use your Google Nest speakers at home?

- A. For music playback
- B. For controlling smart devices
- C. For asking questions/time/weather
- D. For security alerts or doorbell announcements

Answer: A. Users reported using Nest speakers most often for music, followed by smart home control and daily reminders such as weather or timers.

2. How satisfied are you with the sound quality of Nest Audio compared to Nest Mini?

- A. Nest Audio is leagues better, especially bass
- B. Mini is enough for casual listening
- C. Nest Audio is adequate but not impressive
- D. No difference noticed

Answer: A. Users consistently said Nest Audio has much better bass and overall quality compared to Mini, making it preferable for music.

3. Have you experienced connection issues with your Nest speakers in recent years?

- A. Yes, frequent disconnections and 'sorry something went wrong'
- B. Yes, occasional hiccups
- C. No, they work reliably
- D. Issues only due to Wi-Fi provider/router

Answer: A. Many users reported worsening connection reliability over time, though some fixed issues by upgrading Wi-Fi or resetting devices.

4. How well does Google Home integrate with your non-Google devices (like Tuya, ZigBee, or Ikea smart products)?

- A. Very smooth integration
- B. Works but often buggy
- C. I cannot integrate them at all
- D. I only use 100% Google products

Answer: B. Several users noted persistent issues integrating Tuya/Lidl ZigBee and Ikea products with Google Home compared to their native apps.

5. Which competing ecosystem do you feel has smoother smart home integration?

- A. Apple HomeKit/HomePod
- B. Google Nest/Home
- C. Amazon Alexa/Echo
- D. None of them are ideal

Answer: A. Some users with Apple devices felt HomePod integration was smoother, while Google or Amazon were preferred for value and compatibility.

6. What's your biggest frustration with Nest smart speakers?

- A. Laggy responses
- B. Misunderstood commands
- C. Device going offline
- D. Limited software updates

Answer: A. Users most commonly complained about slow responses and frequent 'didn't understand' errors, especially in recent years.

7. How often do you use your Nest devices for multi-room audio?

- A. Daily
- B. Occasionally for parties/events

Table 4: Performance of reasoning LLMs, general LLMs, and open-source LLM on Google Nest Smart Discussion.

Model	Depth					Overall	Breadth	Informativeness		Correctness
	L1	L2	L3	L4	L5			Lexical	Semantic↓	
GPT-5	3.33	21.79	6.53	4.19	0.00	7.17	50.84	0.87	0.30	0.75
GPT-o3	2.51	20.02	5.33	2.77	1.91	6.51	50.19	0.75	0.46	0.95
Qwen3-30B-A3B	31.32	16.61	2.77	1.49	0.00	10.44	50.51	0.81	0.39	0.75

- C. Rarely
- D. Never

Answer: A. Users said multi-room audio used to work seamlessly but has become frustratingly unreliable, with groups often not functioning.

8. Would you consider switching away from Google due to reliability issues?

- A. Yes, to Alexa
- B. Yes, to Apple
- C. Yes, to Home Assistant or local-first systems
- D. No, I will stay with Google

Answer: A. Some users said they were ready to move to Alexa or Home Assistant because of lag and reliability problems with Nest.

9. Do you find Nest speakers useful for families with children?

- A. Yes, great for music and educational features
- B. Somewhat, but limited reliability frustrates kids
- C. Not useful at all for kids
- D. Only good as background music

Answer: B. Users mentioned features like 'animal of the day' were fun for kids, but white noise and music playback have become unreliable.

10. How important is price compared to reliability when choosing between Google, Amazon, or Apple smart speakers?

- A. Price is the most important
- B. Balance between price and reliability
- C. Reliability is the most important
- D. Ecosystem integration matters most

Answer: A. Users often noted that Google's affordability got them locked in, but reliability issues make them consider pricier alternatives.

11. Have you faced limitations when casting media from phones to Nest devices?

- A. Yes, apps don't always connect
- B. Sometimes there's lag/delay
- C. No issues at all
- D. I don't use casting

Answer: A. Users reported frequent casting issues, especially with Spotify, YouTube, and video streams not reaching certain Nest devices.

12. If you primarily wanted a device for music, which would you pick?

- A. Google Nest Audio
- B. Amazon Echo 4th gen
- C. Apple HomePod Mini
- D. Traditional Bluetooth speakers

Answer: A. Users debated between Nest Audio, Echo, and HomePod. Many said Nest Audio had good bass but Echo was decent, while some still preferred Sonos or passive Bluetooth sets.

13. How do you feel about Google discontinuing/reducing stock of Nest devices?

- A. Concerned about product support
- B. Neutral, waiting for new models
- C. Considering switching to another brand
- D. Not worried at all

Answer: A. Several users worried Nest Mini and Audio are discontinued, wondering if Google will abandon the smart speaker hardware.

14. What feature would make you more likely to stick with Google Nest speakers?

1131	A. Improved reliability and faster response	C. No, too much risk of discontinued support	1175
1132	B. Better music/audio quality	D. I prefer other brands' smart displays	1176
1133	C. Deeper integration with third-party devices		1177
1134	D. Clear roadmap and updates from Google		1178
1135		<i>Answer: C. Users were hesitant to buy discontinued Nest Hubs/Max, fearing bricking or lack of updates.</i>	1179
1136	<i>Answer: A. Users said better reliability, sound improvements, and smoother ecosystem updates would convince them to remain loyal.</i>		1180
1137			
1138			
1139	15. Do you experience more issues with Google Assistant understanding you in multilingual households?	19. When connecting Nest with services like Spotify or YouTube Music, what's your experience?	1181
1140			1182
1141			1183
1142	A. Yes, it constantly misinterprets	A. Smooth, works well	1184
1143	B. Sometimes, especially switching languages	B. Works but occasionally lags	1185
1144	C. No, it works fine in multiple languages	C. Often breaks or blocks premium-only features	1186
1145	D. I only use one language	D. I don't link music services	1187
1146			1188
1147	<i>Answer: A. Users noted Assistant struggles badly in multilingual homes, often failing basic commands or mixing languages incorrectly.</i>	<i>Answer: C. Several users reported Spotify on Nest sometimes says 'premium only' even with premium, and YouTube Music integration often fails.</i>	1189
1148			1190
1149			1191
1150			1192
1151	16. What's your perspective on Nest speakers' long-term durability?	20. What future direction should Google take with Nest smart speakers?	1193
1152			1194
1153	A. Still working fine years later	A. Bring Gemini AI with better natural understanding	1195
1154	B. Performance has worsened over time	B. Focus on keeping devices reliable	1196
1155	C. Hardware is durable but software declines	C. Produce new affordable hardware	1197
1156	D. They feel like e-waste now	D. Open-source support if retiring devices	1198
1157			1199
1158	<i>Answer: C. Some users praised durability, while many complained hardware outlasted software support, calling devices obsolete early.</i>	<i>Answer: B. Users speculated Google must fix reliability, offer Gemini AI improvements, and either release new hardware or open source old devices.</i>	1200
1159			1201
1160			1202
1161			1203
1162	17. How do you primarily resolve issues with Nest devices?	A.6.2 Google Nest Smart Speaker questionnaire from GPT-o3	1204
1163			1205
1164	A. Factory reset	1. When users retrofit 1980s intercoms with Nest Mini units, which room-specific control do they hope to achieve later on?	1206
1165	B. Router and Wi-Fi upgrades		1207
1166	C. Reinstalling Google Home app	A. Only ceiling fans and lights of that bedroom	1208
1167	D. Contacting Google support	B. Satellite TV channels in the garage	1209
1168	<i>Answer: A. Most users resorted to factory resets or Wi-Fi upgrades; official support was rarely mentioned as helpful.</i>	C. Printer queues in the study	1210
1169		D. Irrigation valves in the backyard	1211
1170			1212
1171	18. Would you invest in another Nest smart display (like Hub/Hub Max) now?	<i>Answer: A. Users describe planning "a speaker in every bedroom with some intricate setup to both only control devices specific to that room (like ceiling fans and lights) as well as shared devices."</i>	1213
1172			1214
1173	A. Yes, I still trust Google ecosystem		1215
1174	B. Maybe, if I find a second-hand deal		1216

1219 2. **What adjective did a long-time owner use**
1220 **to praise Nest Audio’s stereo sound after**
1221 **pairing two units?**

- 1222 A. Incredible
- 1223 B. Tinny
- 1224 C. Muffled
- 1225 D. Overpriced

1226 *Answer: A. A commenter said “I use 2 Nest*
1227 *Audios in a stereo setup, and the audio is in-*
1228 *credible,” reflecting positive feelings about*
1229 *sound quality.*

1230 3. **Which competing smart speaker line did**
1231 **several Redditors say they might switch**
1232 **to because Google devices have become**
1233 **“laggy” and “driving me insane”?**

- 1234 A. Amazon Echo / Alexa
- 1235 B. Sonos Era
- 1236 C. Bose Smart Ultra
- 1237 D. Marshall Uxbridge

1238 *Answer: A. Many posts mention considering*
1239 *Amazon Echo or Alexa displays as an alterna-*
1240 *tive when Nest performance deteriorated.*

1241 4. **In the thread about buying a Nest Hub Max,**
1242 **which security-related use case was high-**
1243 **lighted as a reason to still want the display?**

- 1244 A. Acting as a digital photo frame with cam-
- 1245 **era recording**
- 1246 B. Hosting a VPN server
- 1247 C. Controlling sprinklers via Zigbee
- 1248 D. Calibrating 3D printers

1249 *Answer: A. A buyer said they liked “the cam-*
1250 *era/security recording function and using it*
1251 *as a digital photo frame,” showing the usage*
1252 *scenario.*

1253 5. **Which phrase did frustrated owners repeat-**
1254 **edly hear instead of successful commands,**
1255 **prompting them to call Google Home a**
1256 **“support group”?**

- 1257 A. “Sorry, something went wrong, try again
- 1258 **later.”**
- 1259 B. “Firmware upgrade in progress.”
- 1260 C. “Device is paired in another room.”
- 1261 D. “Low battery, shutting down.”

Answer: A. Multiple users quote the device
replying “Sorry, something went wrong, try
again later,” illustrating a common pain point.

1265 6. **Why did one user say the Pixel Tablet on**
1266 **its dock feels like an “old TV/VCR combo”**
1267 **compared with a real Nest Hub?**

- 1268 A. It can’t be asked to play music on other
- 1269 **Google speakers**
- 1270 B. It lacks Wi-Fi 6E support
- 1271 C. The screen is smaller than 5 inches
- 1272 D. It forces Amazon Prime ads

1273 *Answer: A. They complained that you “can’t*
1274 *tell it to play music on it from another Google*
1275 *speaker,” so the hybrid device does neither*
1276 *role well.*

1277 7. **Which connectivity problem did a border-**
1278 **area listener report when TuneIn stations**
1279 **kept dropping on Nest speakers?**

- 1280 A. Occasional to frequent loss in signal
- 1281 B. Crackling Bluetooth interference only at
- 1282 **night**
- 1283 C. Wrong language playback
- 1284 D. Overheating power adapters

1285 *Answer: A. The post says “I have experienced*
1286 *occasional to frequent loss in signal when lis-*
1287 *tening to stations that utilize TuneIn.”*

1288 8. **When discussing Matter devices going of-**
1289 **line, which brand of mesh router system**
1290 **was singled out for Thread settings confu-**
1291 **sion?**

- 1292 A. Eero 6E
- 1293 B. TP-Link Deco
- 1294 C. UniFi Dream Router
- 1295 D. Netgear Orbi

1296 *Answer: A. A user wrote “I have an Eero*
1297 *6e mesh router system. . . The Threads fea-*
1298 *ture is toggled on,” yet their Matter gear still*
1299 *dropped.*

1300 9. **How did a Nest Mini owner describe the**
1301 **music delay when the speaker was added**
1302 **to a stereo link in Google Home?**

- 1303 A. The delay is HUGE.
- 1304 B. It syncs perfectly.
- 1305 C. Only milliseconds of lag.

D. Delay happens once a month.

Answer: A. The post states, "If I play any music... the delay is HUGE," emphasizing a negative feeling about latency.

10. **Which future-oriented speculation did shoppers raise after noticing no Nest Audio stock in multiple country stores?**

- A. A new generation might be announced at the Pixel event
- B. Google is switching to Apple HomeKit
- C. All smart speakers will become subscription-based
- D. Wi-Fi will be removed from Nest

Answer: A. They asked, "Are people expecting a new generation to be announced at the Pixel event in a couple weeks?"—a tendency toward anticipating new hardware.

11. **Which cloud storage dilemma did dual-ecosystem users discuss while already owning many Nest Hubs and iCloud devices?**

- A. Paying for both 200 GB iCloud and 200 GB Google One plans
- B. Choosing between Dropbox and Box free tiers
- C. Losing access to Microsoft OneDrive photos
- D. Migrating from Amazon S3 Glacier Vaults

Answer: A. The repeated post describes both iCloud and Google One hitting the 200 GB limit and not wanting to upgrade both.

12. **What network feature on apartment Wi-Fi prevented an elderly resident's Nest Mini from completing setup?**

- A. AP Isolation turned on
- B. Hidden SSID broadcast
- C. WPA3-Enterprise only
- D. Dual NAT tunneling

Answer: A. The care home enables "AP Isolation," so the speaker throws the message "Please check your Wi-Fi network settings."

13. **Which sound-related improvement motivated users to prefer Nest Audio over their old Google Home Minis?**

- A. 'Bass is the most noticeable improvement' at high volume
- B. Built-in CD player support
- C. Dolby Atmos rear channels
- D. Quad-mic noise cancelling

Answer: A. One review says, "Bass is the most noticeable improvement, high volume performance is better," highlighting the aspect of audio quality.

14. **How much did Canadian bargain hunters report paying at Lowe's or Home Depot for clearance Nest Audio units?**

- A. \$39.97
- B. \$129.99
- C. \$199.00
- D. \$15.00

Answer: A. Posts note "Nest Audio for sale for \$39.97... is it worth getting," reflecting pricing sentiment.

15. **Which workaround did some owners adopt because the Nest Hub could no longer resume music on the intended room speaker?**

- A. Using the broadcast command instead of TTS
- B. Switching to Zigbee bulbs
- C. Turning on microphone sensitivity
- D. Downgrading firmware via USB

Answer: A. One poster said they had to "re-sort to using broadcast commands which are clunky" when TTS stopped working.

16. **What phrase did a Nest thermostat user shout after eco mode kept activating despite settings being disabled?**

- A. "Jeezus Google."
- B. "Bravo Assistant!"
- C. "Mission accomplished!"
- D. "Danke Alexa."

Answer: A. The frustrated quote is "Jeezus Google," showing irritation with unwanted eco behaviour.

17. **When debating cloud versus local AI, which low-power device did a homeowner consider dedicating as an "always-on screen" for NotebookLM?**

- A. An old MacBook Pro
- B. A Raspberry Pi Zero
- C. A Lenovo Tab M8
- D. A Pixel 2 phone

Answer: A. They planned “to dedicate an old MacBook Pro for the AI assistant” but were open to tablets.

18. **Which free radio service did a listener compare to TuneIn, noting that Audacy retained signal ‘much better’ on Nest speakers?**

- A. Audacy
- B. Pandora
- C. SiriusXM
- D. iHeartWeather

Answer: A. The poster said “The other services like Audacy work much better in terms of signal retention,” offering a comparison.

19. **What is one reason a user said HomePods were a ‘let down’ compared with Google Nest speakers?**

- A. ‘You have to AirPlay from your phone’ for many tasks
- B. They lack any bass at all
- C. They cannot set alarms
- D. HomePods don’t support Wi-Fi

Answer: A. The criticism: “on HomePod minis you have to air play from your phone... google are just way ahead,” indicating a comparison and feeling.

20. **Faced with a 2025 Nest price hike e-mail, how many days of event video history does the ‘Nest Aware Plus’ plan promise to keep?**

- A. 60 days
- B. 30 days
- C. 7 days
- D. 120 days

Answer: A. The notice states the plan keeps “60 days of event video history... up to 10 days of 24/7 video history,” summarising the aspect of subscription value.

A.6.3 Google Nest Smart Speaker questionnaire from Qwen3-30B-A3B

1. **In your smart home setup, how do you integrate Google Nest Mini speakers into an older home intercom system?**

- A. By replacing the entire intercom system with new wireless speakers
- B. By using a 24VDC power supply and buck converters to repurpose old wiring
- C. By connecting the speakers directly to a home Wi-Fi network without rewiring
- D. By mounting the speakers to the wall using adhesive pads only

Answer: B. By using a 24VDC power supply and buck converters to repurpose old wiring

2. **What is the primary reason for using Google Nest Audio speakers in a stereo setup for music playback?**

- A. They offer better sound quality and deeper bass than other smart speakers
- B. They are the only Google devices compatible with Apple Music
- C. They support Bluetooth audio output without delay
- D. They are designed specifically for use with Sonos systems

Answer: A. They offer better sound quality and deeper bass than other smart speakers

3. **When using Google Nest Mini as a temporary speaker during a move, what feature is most important for portability?**

- A. Built-in battery and rechargeable power source
- B. Ability to connect via USB-C to a TV
- C. Support for HDMI input from a computer
- D. Integration with Apple HomeKit for remote access

Answer: A. Built-in battery and rechargeable power source

4. **Why might a user choose Google Nest Mini over Apple HomePod Mini for a smart home setup?**

- A. Because Apple HomePod Mini requires a HomePod Hub for full functionality

- 1481 B. Because Google Nest Mini supports
- 1482 more third-party integrations and voice
- 1483 commands
- 1484 C. Because HomePod Mini lacks music
- 1485 playback capabilities
- 1486 D. Because Google Mini is compatible with
- 1487 all smart home ecosystems

1488 *Answer: B. Because Google Nest Mini sup-*

1489 *ports more third-party integrations and voice*

1490 *commands*

1491 **5. What has caused frustration for users who**

1492 **rely on Google Nest speakers for music**

1493 **playback and voice commands?**

- 1494 A. Frequent unresponsiveness and inconsis-
- 1495 tent command recognition
- 1496 B. Inability to support Spotify Premium ac-
- 1497 counts
- 1498 C. Limited compatibility with Bluetooth de-
- 1499 vices
- 1500 D. High cost of replacing outdated models

1501 *Answer: A. Frequent unresponsiveness and*

1502 *inconsistent command recognition*

1503 **6. How do users perceive the reliability of**

1504 **Google Nest devices after the company's**

1505 **shift to Gemini AI?**

- 1506 A. They believe the new AI will improve
- 1507 performance and voice recognition
- 1508 B. They are concerned about service degra-
- 1509 dation and potential obsolescence
- 1510 C. They expect all devices to be replaced
- 1511 within a year
- 1512 D. They trust that Google will maintain full
- 1513 support indefinitely

1514 *Answer: B. They are concerned about service*

1515 *degradation and potential obsolescence*

1516 **7. What do users suggest as the best way to**

1517 **maintain a smart home ecosystem without**

1518 **relying on Google's services?**

- 1519 A. Using Home Assistant with a local server
- 1520 for full control
- 1521 B. Switching to Apple HomeKit for better
- 1522 integration
- 1523 C. Purchasing only Amazon Echo devices
- 1524 for compatibility
- 1525 D. Replacing all devices with non-smart al-
- 1526 ternatives

Answer: A. Using Home Assistant with a local

server for full control

1527 **8. What is a common workaround when a**

1528 **Google Nest speaker fails to connect to Wi-**

1529 **Fi during setup?**

- 1530 A. Changing the account region to the
- 1531 United States in the app settings
- 1532 B. Using a wired Ethernet connection in-
- 1533 stead of Wi-Fi
- 1534 C. Factory resetting the device and re-
- 1535 registering it
- 1536 D. Replacing the speaker with a newer
- 1537 model

1538 *Answer: A. Changing the account region to*

1539 *the United States in the app settings*

1540 **9. Why do some users prefer Google Nest Au-**

1541 **dio over other smart speakers for music**

1542 **listening?**

- 1543 A. It has superior stereo sound quality and
- 1544 bass response
- 1545 B. It is the most affordable smart speaker
- 1546 with no subscription fees
- 1547 C. It supports more music streaming ser-
- 1548 vices than any other model
- 1549 D. It integrates with Google Photos for vi-
- 1550 sual album displays

1551 *Answer: A. It has superior stereo sound qual-*

1552 *ity and bass response*

1553 **10. What is the main concern for users who are**

1554 **invested in both Apple and Google ecosys-**

1555 **tems?**

- 1556 A. Managing duplicate cloud storage be-
- 1557 tween iCloud and Google One
- 1558 B. Choosing between Apple HomePod and
- 1559 Google Nest for smart home control
- 1560 C. Upgrading to new smartphones to main-
- 1561 tain compatibility
- 1562 D. Finding compatible smart bulbs for both
- 1563 platforms

1564 *Answer: A. Managing duplicate cloud storage*

1565 *between iCloud and Google One*

1566 **11. How do users feel about the future of**

1567 **Google Nest hardware following recent ser-**

1568 **vice issues?**

- 1571 A. They believe Google will soon release a
- 1572 new generation of devices
- 1573 B. They are skeptical about long-term sup-
- 1574 port and product continuity
- 1575 C. They expect Google to fully open-source
- 1576 the software for user maintenance
- 1577 D. They are confident that all devices will
- 1578 continue to function indefinitely

1579 *Answer: B. They are skeptical about long-term*
 1580 *support and product continuity*

12. **What challenge arises when trying to control Google Nest devices from a non-Google ecosystem device?**

- 1584 A. Voice commands are not recognized due
- 1585 to language mismatch
- 1586 B. The speaker cannot detect voice input
- 1587 from Apple devices
- 1588 C. Integration with Home Assistant be-
- 1589 comes unreliable or impossible
- 1590 D. The device refuses to connect to Wi-Fi
- 1591 networks outside the Google network

1592 *Answer: C. Integration with Home Assistant*
 1593 *becomes unreliable or impossible*

13. **What is a key limitation when using Google Nest Mini speakers for doorbell alerts?**

- 1596 A. The alert only plays on one speaker, not
- 1597 all connected devices
- 1598 B. The doorbell chime only works when the
- 1599 speaker is unplugged
- 1600 C. The speaker cannot broadcast alerts dur-
- 1601 ing music playback
- 1602 D. Alerts are delayed by 10–15 seconds due
- 1603 to app routing

1604 *Answer: A. The alert only plays on one*
 1605 *speaker, not all connected devices*

14. **Why might a user consider replacing a Google Nest Hub Max despite its popularity?**

- 1606 A. It is no longer available for purchase in
- 1607 many regions
- 1608 B. It has a built-in camera that cannot be
- 1609 disabled
- 1610 C. It requires a paid subscription for basic
- 1611 functionality
- 1612 D. It only works with Apple devices and not
- 1613 Android
- 1614
- 1615
- 1616

Answer: A. It is no longer available for pur-
 1617 *chase in many regions* 1618

15. **How do users describe the current performance of Google Nest Hub devices in 2025?**

- 1621 A. They are slow, glitchy, and frequently
- 1622 misinterpret commands
- 1623 B. They have improved dramatically with
- 1624 the new Gemini update
- 1625 C. They function as well as they did five
- 1626 years ago
- 1627 D. They are only useful for playing music
- 1628 and showing photos

1629 *Answer: A. They are slow, glitchy, and fre-*
 1630 *quently misinterpret commands*

16. **What is a major reason users are hesitant to expand their Google Nest ecosystem?**

- 1631 A. Fear of being locked into a system that
- 1632 may no longer be supported
- 1633 B. High cost of additional smart home ac-
- 1634 cessories
- 1635 C. Incompatibility with older Wi-Fi routers
- 1636 D. Lack of support for non-English lan-
- 1637 guages

1638 *Answer: A. Fear of being locked into a system*
 1639 *that may no longer be supported* 1640

17. **What is the most common solution users suggest for improving audio performance across multiple Nest speakers?**

- 1641 A. Upgrading to a higher-end model like the
- 1642 Google Home Max
- 1643 B. Using a dedicated audio amplifier for bet-
- 1644 ter sound quality
- 1645 C. Replacing all speakers with Sonos de-
- 1646 vices for consistent playback
- 1647 D. Using a local server to manage audio
- 1648 streams and reduce latency

1649 *Answer: D. Using a local server to manage*
 1650 *audio streams and reduce latency* 1651

18. **What is a major drawback of using Google Nest Audio as a TV speaker?**

- 1652 A. It has a significant audio delay when con-
- 1653 nected via Bluetooth
- 1654 B. It lacks a 3.5mm input port for direct
- 1655 audio connection
- 1656
- 1657
- 1658
- 1659
- 1660

- 1661 C. It cannot play video content from stream-
1662 ing apps
1663 D. It only works with Google TV devices

1664 *Answer: A. It has a significant audio delay*
1665 *when connected via Bluetooth*

1666 **19. What do users suggest as an alternative to**
1667 **Google Nest speakers for a private, local**
1668 **smart home system?**

- 1669 A. Using a Raspberry Pi with a local LLM
1670 for voice control
1671 B. Replacing all devices with Amazon Echo
1672 Dot models
1673 C. Switching to Apple HomePods for better
1674 privacy
1675 D. Using only non-smart speakers with man-
1676 ual controls

1677 *Answer: A. Using a Raspberry Pi with a local*
1678 *LLM for voice control*

1679 **20. How do users perceive the value of older**
1680 **Google Nest devices like the Nest Mini and**
1681 **Nest Audio?**

- 1682 A. They are still functional and affordable,
1683 especially when bought secondhand
1684 B. They are outdated and no longer sup-
1685 ported by Google
1686 C. They are only useful for basic tasks like
1687 playing alarms
1688 D. They are incompatible with modern Wi-
1689 Fi networks

1690 *Answer: A. They are still functional and af-*
1691 *fordable, especially when bought secondhand*

1692 **A.7 Lighted Tree of the questionnaires on**
1693 **Google Nest Smart Speaker**

