SynthAVE: Scalable Synthetic Labeling for E-Commerce with LLM-Arena Validation

Andrea Scarinci[1], Virginia Negri[1], Brayan Impata[1], Suleiman Khan[1], Victor Martinez[1], Marcello Federico[1]

[1]Amazon

{andscar, vrgngr, biimpata, suleimkh, vicmg, marcfede}@amazon.com

Abstract

Fine-tuning large language models (LLMs) for e-commerce attribute extraction requires labeled data representative across thousands of product types, attributes, and multiple languages. This combinatorial scale translates to millions of annotations, rendering human labeling prohibitively costly. While recent work has demonstrated synthetic label generation using LLMs (Negri et al., 2025), deploying such approaches at industrial scale requires integrated quality control mechanisms. We present **SynthAVE**, a large-scale benchmark for attribute-value verification spanning 12,726 products across 229 product types, 792 attributes, and 4 languages (Spanish, French, Italian, German). To validate synthetic labels at scale, we introduce a multi-LLM arena framework where each sample is evaluated by 21 judge configurations (7 model families $\times$ 3 prompts), with final labels determined via majority voting; disagreements with the synthetic label are expert-adjudicated and agreement cases are audited on a stratified sample. The majority vote ensemble agrees with expert labels at Cohen's $\kappa = 0.92$ (95.0% agreement), while individual judges agree with one another only moderately (Fleiss' $\kappa = 0.76$)—by design, since we select for diversity. This demonstrates that diverse models with varying individual judgments aggregate into highly reliable predictions, enabling cost-effective validation at scale while concentrating expert effort on the cases where it changes the label. We estimate the resulting label quality at 97.9%.

1 Introduction

Fine-tuning and evaluating large language models (LLMs) for e-commerce applications demands massive volumes of high-quality labeled data. Specifically, predicting and evaluating product attribute values from unstructured catalog text (e.g., product titles, descriptions, bullet points) requires labeled examples that maintain representativeness across three key dimensions: product categories (e.g., electronics, clothing, furniture), attributes (e.g., color, material, dimensions), and languages. At catalog scale—thousands of product categories, thousands of attributes, and multiple languages—achieving reliable model performance requires hundreds of labeled examples per (product category $\times$ attribute $\times$ language) triplet. This combinatorial requirement translates to millions of annotations, making human labeling prohibitively costly.

Prior work demonstrates the feasibility of using LLMs to generate synthetic labels for e-commerce attribute value extraction (Negri et al., 2025). However, deploying such approaches at industrial scale introduces a critical challenge: how to validate label quality efficiently across heterogeneous attributes and languages without exhaustive manual review. A scalable solution must integrate synthetic label generation with automated quality control mechanisms.

This paper presents SynthAVE (Synthetic data for Attribute Value Extraction), a large-scale benchmark for attribute-value verification augmenting prior generation methodology (Negri et al., 2025) with scalable quality assurance. We introduce a multi-LLM auditing framework where each generated label is evaluated by 21 judge configurations (7 model families $\times$ 3 prompts), with final labels determined through majority voting. To mitigate systematic biases, we employ diverse models from different providers alongside varied prompting strategies. Human review is concentrated on the cases where the arena contradicts the synthetic label; agreement cases are auto-accepted and audited on a stratified sample.

Empirical evaluation demonstrates that diverse models with different biases aggregate into reliable predictions: the ensemble achieves 95.0% agreement with expert labels (Cohen's $\kappa = 0.92$) while enabling cost-effective validation at scale.

Our main contributions are:

EMNLP 2026 Industry Track, pages 497–518
October 24-29, 2026 ©2026 Association for Computational Linguistics

- **SynthAVE**: A benchmark of 12,726 products spanning 229 product categories, 792 attributes, and 4 languages for multilingual attribute-value verification.[1]

- **Multi-LLM Validation Framework**: An auditing approach employing majority voting across diverse judge configurations, achieving 95.0% agreement with expert evaluation while enabling cost-effective validation at scale, with an analysis of panel composition.

2 Related Work

Information extraction (IE) from unstructured web text is a foundational capability for large-scale e-commerce systems and knowledge base construction. Common applications include product attribute extraction, entity population, catalog enrichment, and fact acquisition from heterogeneous online sources (Zhang et al., 2024; Martinez-Rodriguez et al., 2020; Dagdelen et al., 2024). Recent advances in large language models (LLMs) have substantially lowered the barrier to deploying flexible extraction systems across domains and schemas (Zhu et al., 2024). However, evaluating the quality of extracted information remains a central challenge, primarily due to the scarcity of high-quality human-annotated data and the scale at which modern IE systems operate.

Classical evaluation methodologies for IE depend on manually annotated ground truth datasets and report performance using metrics such as precision, recall, and F_1 score (Xu et al., 2024; Zhu et al., 2024). While effective in controlled research settings, this paradigm does not scale to real-world web and e-commerce applications. Annotation is expensive, time-consuming, and often requires domain expertise, particularly when extraction targets complex or evolving schemas (Deng et al., 2024; Hsu and Roberts, 2025). As a result, many production pipelines operate with limited or no gold-standard evaluation data, motivating the development of evaluation methods that function under weak or missing supervision.

Existing benchmarks for product attribute extraction illustrate this tension between scale and quality. The Multi-source Attribute Value Extraction (MAVE) dataset (Yang et al., 2022) provides 2.2 million attribute-value annotations across 1,257 attributes from Amazon product profiles—achieving impressive scale through automated pipelines. However, MAVE's quality assurance relies primarily on classifier-based filtering rather than comprehensive human validation of attribute-value annotations. Additionally, MAVE is limited to English and addresses extraction rather than verification tasks. These limitations motivate both the exploration of synthetic evaluation methods and the development of more rigorous validation frameworks.

In response to these constraints, recent work has explored evaluation strategies that reduce or eliminate dependence on human-labeled ground truth. Seitl et al. (2024) propose an automatic IE evaluation framework based on synthetic ground truth generation, in which artificially constructed structured facts are injected into natural documents and systems are evaluated on their ability to recover them. Despite their scalability, synthetic evaluation methods introduce limitations related to realism—artificially injected facts may not fully capture the ambiguity, inconsistency, and noise present in naturally occurring e-commerce text (Negri et al., 2025).

One response to the limits of a single LLM judge is to use a panel of them. Zheng et al. (2023) introduced the arena paradigm, and Verga et al. (2024) showed that several smaller, diverse models can match one large judge. Panel size alone, however, does not guarantee proportionally more evidence: Kohli (2026) find that when judges tend to make the same mistakes, a nine-model panel can be worth little more than two genuinely independent votes. We therefore report inter-judge agreement as a measure of how far our judges' errors overlap, rather than as evidence that their votes are independent.

Another line of work investigates the use of LLMs to generate annotated datasets for downstream evaluation. By leveraging LLMs as annotators, it becomes possible to rapidly construct large quantities of labeled data where human annotation is infeasible (Mohta et al., 2023; Tan et al., 2024). In e-commerce contexts, this supports approximate evaluation of extraction pipelines across diverse product categories (Nadaş et al., 2025). However, automatically generated annotations may reflect model-specific biases or systematic errors, necessitating careful validation and selective human oversight (De et al., 2025). Pipeline design choices—decomposing extraction into multiple stages, structured output formats, and normalization rules—

[1]The public release will additionally include English-language annotations produced using the same pipeline, extending coverage to 5 languages.

further improve performance (Jaradeh et al., 2023; Sabeh et al., 2024; Satyadharma et al., 2025; Zhang et al., 2025), but complicate evaluation when standardized benchmarks are unavailable.

3 Task and Dataset

Attribute-Value Verification Task. The attribute-value verification task requires determining whether a given attribute value can be verified from unstructured product text. Given a product's unstructured text (title, bullet points, description) and an attribute-value pair, the task is to assign one of three labels: CORRECT, INCORRECT, or UNKNOWN.

We illustrate with a concrete example. Given the product title "`Portable Laptop Stand, Adjustable Height, Aluminum Construction, Silver Finish, Compatible with 10-17 inch Devices`", the label depends on the attribute-value pair: (`color`, `Silver`) is CORRECT since the text states "Silver Finish"; (`color`, `Black`) is INCORRECT as it contradicts the text; and (`weight`, `1.2kg`) is UNKNOWN since no weight information is present.

Training models to perform this three-way classification reliably requires labeled examples of all cases across diverse product categories, attributes, and languages. Our synthetic data generation pipeline produces such examples at scale, and our validation framework ensures their quality. The benchmark evaluates *verification* of supplied pairs, not their discovery, so upstream extraction recall lies outside its scope.

Synthetic Data Generation. We constructed the dataset in two stages. First, we sampled real products from a large e-commerce catalog spanning 4 languages (Spanish, French, Italian, German), ensuring categories exist in all languages with minimum 30 products per category. Second, for each sampled product, we generated synthetic variants using the controlled generation methodology of Negri et al. (2025). This process creates a strict one-to-one correspondence: each synthetic product instance is paired with exactly one target attribute and one value slot. The generation pipeline produces all three label types: CORRECT (value explicitly or implicitly supported by text), INCORRECT (text contradicts the value), and UNKNOWN (attribute cannot be determined from available text).

The label distribution in Table 1 is set by the generation protocol: attribute values are deliberately manipulated so that all three label types occur in usable proportions. Validation shifts this mix only mildly (Appendix D).

Table 1: Dataset statistics by language

	ES	FR	IT	DE
TOTAL PRODUCTS	3,159	3,501	3,007	3,059
% OF DATASET	24.8	27.5	23.6	24.0
UNIQUE ATTRIBUTES	636	675	609	673
Label Distribution (%)				
CORRECT	48.6	48.8	49.1	48.7
INCORRECT	15.3	15.8	14.5	14.3
UNKNOWN	36.1	35.4	36.4	36.9

The synthetic dataset comprises 12,726 products across 229 product categories, covering 792 unique attributes and 2,607 distinct product-attribute combinations. The dataset spans major product domains including electronics, apparel, home & furniture, and sporting goods. Each category contains a minimum of 30 products, with larger categories such as NOTEBOOK_COMPUTER reaching 209 products. Categories are evaluated on domain-specific attributes (e.g., `connectivity_technology` for electronics, `closure.type` for apparel). Full category and attribute distributions are provided in Appendix D.

Validation Challenge. The synthetic pipeline reaches only 92.6% accuracy in preliminary analysis, and metrics computed against noisy ground truth yield misleading conclusions. Yet exhaustive human review of 12,726 products across 4 languages is prohibitively expensive. Section 4 presents our solution.

4 Multi-LLM Validation Framework

Having defined the attribute-value verification task and dataset construction process, we now address the core challenge: validating synthetic labels at scale without exhaustive human review. Building upon established generation methods (Negri et al., 2025), we present a multi-LLM auditing framework that aggregates diverse model judgments into reliable quality estimates.

Our framework aggregates many LLM judgments into an estimate that approximates expert consensus. The judges are *not* statistically independent—their errors are correlated, as we quantify in Section 5—so the design goal is to maximise the diversity of failure modes rather than to assume independent votes. We pursue diversity along two dimensions, and require each judge to

Table 2: LLM models evaluated in our arena.

MODEL FAMILY	MODEL NAME
ANTHROPIC	CLAUDE 3.5 SONNET V2
AMAZON	NOVA PRO V1
OPENAI	GPT OSS 120B
MISTRAL	MISTRAL LARGE 3
DEEPSEEK	DEEPSEEK R1
QWEN	QWEN3 VL 235B
GOOGLE	GEMMA 3 27B

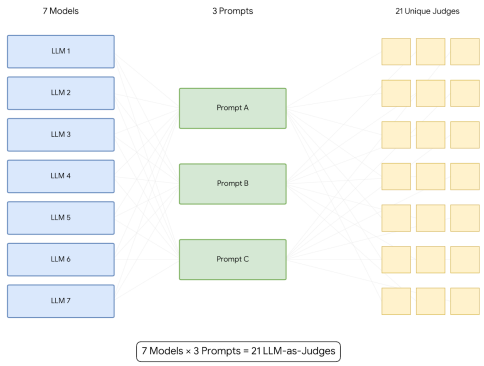

Figure 1: Framework for multi-LLM judge configurations. Seven model families combined with three prompt variations yield 21 judge configurations.

meet qualification criteria analogous to those applied to human auditors (Appendix G.1).

Judge Selection and Configuration. We diversify judges along two axes.

Model family diversity. We selected 7 LLM families from different providers (Table 2): Claude (Anthropic, 2024), Nova (Amazon Artificial General Intelligence, 2024), GPT (OpenAI et al., 2024), Mistral (Mistral AI, 2025), DeepSeek (Guo et al., 2025), Qwen (Bai et al., 2025), and Gemma (Gemma Team et al., 2025). This ensures diversity across model scales and training objectives (general-purpose, reasoning-focused, multilingual), reducing the influence of any single family's biases on the aggregate.

Prompt diversity. For each model, we developed three substantially different prompt versions (detailed in Appendix F). This prevents model outputs from reflecting prompt over-fitting or artificial consistency due to identical phrasing.

This design yields $7 \times 3 = 21$ judge configurations. Each configuration evaluates all 12,726 products across 4 languages (Spanish, French, Italian, German), producing 267,246 total judgments.

Evaluation Task. Judges perform the attribute consistency verification task defined in Section 3, classifying each sample as CORRECT, INCORRECT, or UNKNOWN with brief justification.

Aggregation through Majority Voting. We aggregate judge outputs via majority voting across all 21 configurations, resolving the rare ties as described in Appendix G. The agreement level is itself informative: unanimous votes mark the highest-confidence predictions, while split decisions flag samples warranting inspection. The signature we seek is moderate inter-judge agreement (Fleiss' κ) combined with high ensemble–expert agreement (Cohen's κ), since this indicates judges with differing failure modes cancelling errors rather than duplicating one another. Fleiss' κ is therefore reported throughout as a diversity diagnostic, not as evidence of statistical independence.

Ground Truth Establishment. To evaluate our framework, we established ground truth labels through a disagreement-based annotation strategy that leverages the LLM arena to minimize human effort. Each synthetic sample has an original label assigned by the generation algorithm. We compared this label against the LLM arena majority vote and applied the following protocol:

- **Agreement cases** (11,310; 88.9%): the majority vote matched the original synthetic label, and the sample was accepted *without* human review.
- **Disagreement cases** (1,416; 11.1%): the majority vote contradicted the original label, and the sample was adjudicated by a domain expert who assigned the final label.

Only the disagreement cases are therefore directly adjudicated; 10,910 samples (86%) are never inspected by a human, so we describe cases as *expert-adjudicated*, *automatically accepted*, or *independently audited* rather than uniformly human-validated.

To validate the agreement assumption, an expert annotator independently triaged a stratified random sample of 400 agreement cases (100 per language, comprising 100 unanimous and 300 mixed-agreement samples). The annotator overturned the majority vote in 12/400 cases (3.0%, 95% CI [1.3%, 4.7%]); six further cases were marked *unsure* and counted as non-overturns, giving a conservative upper bound of 4.5%. Unanimous cases—where all 21 judges agreed—were never overturned (0/100), and all 12 overturns occurred in mixed-agreement samples (4.0% overturn rate). Among majority

Table 3: LLM Arena agreement with expert evaluation by language

LANG.	PRODUCTS	ACC.	COHEN κ	FLEISS κ
IT	3,007	96.4%	0.94	0.755
DE	3,059	95.3%	0.92	0.757
ES	3,159	94.7%	0.91	0.757
FR	3,501	94.0%	0.90	0.759
ALL	**12,726**	**95.0%**	**0.92**	**0.757**

labels, UNKNOWN showed the highest overturn rate (5.0%) compared to CORRECT (2.2%) and INCORRECT (0.0%), consistent with UNKNOWN being the "hedge" vote that occasionally masks determinable cases. Agreement between the generation algorithm and a diverse judge panel is therefore a reliable, though not infallible, signal of label correctness (full analysis in Appendix A).

5 Results

We first establish that our validation framework approaches expert-level reliability, then present the final dataset features after cleaning.

Overall Arena Performance. Table 3 summarizes agreement between LLM Arena majority voting and expert evaluation. The arena achieves 95.0% overall agreement with experts, with Cohen's $\kappa = 0.92$ indicating almost perfect inter-rater reliability (Landis and Koch, 1977). Performance is consistent across languages (94.0–96.4%). Individual judges agree with one another only substantially (Fleiss' κ = 0.757), yet the majority vote ensemble reaches almost perfect agreement with experts—the intended outcome of selecting judges for diversity rather than redundancy. This figure is not fully independent of the arena, since agreement cases are accepted because the arena concurred with the synthetic label; we bound the resulting optimism below.

Data Cleaning Performance. Table 4 summarizes the Arena's effectiveness as a data cleaning mechanism. The original synthetic pipeline achieves 92.6% accuracy (946 errors across 12,726 samples). The arena corrects 786 of these errors (83.1% error correction rate), improving accuracy to 95.0%. When framed as binary bad-label detection, the Arena achieves 98.0% precision and 95.2% recall (F1 = 96.6%)—when it flags a label as incorrect, it is almost always right.

Table 4: LLM Arena data cleaning performance

METRIC	VALUE
Baseline	
ORIGINAL SYNTHETIC ACCURACY	92.6%
ORIGINAL ERRORS	946
Arena Correction	
ARENA ACCURACY	95.0%
ERRORS CORRECTED	786 / 946 (83.1%)
NET IMPROVEMENT	+2.4%
Bad Label Detection	
PRECISION	98.0%
RECALL	95.2%
F1 SCORE	96.6%

Table 5: Per-class performance metrics (averaged across languages)

LABEL	PRECISION	RECALL	F1-SCORE
CORRECT	96.2%	98.5%	97.3%
INCORRECT	91.5%	87.0%	89.2%
UNKNOWN	95.0%	94.1%	94.5%

Per-Class Performance. Table 5 presents per-label metrics. The arena performs best on CORRECT (F1 = 97.3%), followed by UNKNOWN (F1 = 94.5%). INCORRECT labels prove most challenging (F1 = 89.2%), primarily due to lower recall—the arena tends to classify borderline INCORRECT cases as UNKNOWN, a conservative behavior preferable in production where false negatives are less costly than false positives.

Performance by Original Label Type. Stratifying by original synthetic label reveals where the arena adds most value (Table 6). Originally CORRECT labels already achieve 97.3% accuracy, with modest arena improvement (+1.2%). Originally INCORRECT labels show the largest gap: synthetic accuracy is only 79.3%, but the arena recovers to 88.7% (+9.4%), demonstrating particular effectiveness at correcting the most error-prone category.

Agreement Level and Accuracy. Accuracy correlates strongly with agreement level: unanimous consensus (36% of products) yields 100% accuracy, very high agreement (>85%) yields 98.8%, while low-agreement cases (<50%, 1.4% of products) achieve only 65.1%, suggesting these warrant human review. Full stratification is provided in Appendix C.

Label Quality. The 92.6% baseline treats agreement cases as correct by construction; correcting for their triage-measured residual error (2.37%)

Table 6: Arena accuracy (%) by original synthetic label type

ORIGINAL	COUNT	ORIG ACC	ARENA ACC	Δ
CORRECT	5,980	97.3	98.5	+1.2
INCORRECT	2,240	79.3	88.7	+9.4
UNKNOWN	4,506	92.8	93.6	+0.8

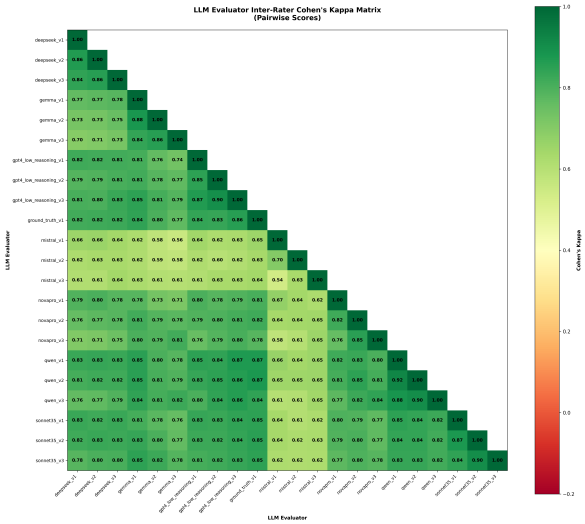

Figure 2: Pairwise Cohen's κ between judge configurations (German). Within-family agreement is high (diagonal blocks), while cross-family diversity drives ensemble accuracy. Patterns hold across all languages (Appendix E).

puts the synthetic labels at 90.5% unconditionally, and the released labels at 97.9% after adjudication (95% CI [96.7%, 99.0%])—a silver rather than gold standard (Northcutt et al., 2021). See Appendix B.

Additional Analyses. No individual judge configuration outperformed the ensemble, and accuracy varied more across model families than across prompts: at three judges, three families with one prompt each (91.8%) beats one family with three prompts (89.9%). The 21 configurations are thus not 21 independent signals—mean pairwise error correlation is 0.36, implying ∼2.6 effective votes (Kohli, 2026)—matching the clustering in Figure 2. The production-critical error is a false CORRECT, asserting an unsupported value: 235 of 630 errors (37.3%), the rest being largely conservative abstentions (INCORRECT → UNKNOWN, 27.5%). Italian shows the highest fix rate (90.2%) and French the lowest (76.1%). Full details are in Appendix C.

Cost Analysis. Total validation cost was $290.50 for 12,726 products ($22.83 per 1,000 products, 267,246 API calls).[2] Detailed breakdowns are in Appendix H.

The SynthAVE Dataset. The resulting dataset comprises 12,726 products spanning 229 product categories, 792 attributes, and 4 languages (47.7% CORRECT, 15.8% INCORRECT, 36.5% UNKNOWN), with estimated label accuracy of 97.9%. Labels are either expert-adjudicated (the 1,416 disagreement cases) or accepted on arena–pipeline consensus and validated by stratified audit (Appendix A). Unlike MAVE (Yang et al., 2022), SynthAVE provides explicit UNKNOWN labels, multilingual coverage, and a documented validation protocol with a quantified residual error rate. The dataset and per-judge predictions will be publicly released.

6 Conclusion

We presented SynthAVE, a large-scale benchmark and validation framework for synthetic e-commerce data. Building on prior work in synthetic data generation (Negri et al., 2025), we introduced an LLM-arena framework that enables scalable quality assurance through multi-model evaluation.

Our evaluation yields three key findings. First, the LLM arena achieves 95% agreement with expert evaluation (Cohen's $\kappa = 0.92$) across 12,726 products, 229 product categories, 792 attributes, and 4 languages. Second, diversity rather than redundancy drives accuracy: judges agree with one another only substantially (Fleiss' $\kappa = 0.76$, a diversity diagnostic rather than a shortfall), yet the ensemble outperforms any single model, with model families mattering more than prompts. Third, expert triage of 400 agreement cases shows the majority panel is overturned only 3% of the time (0% for unanimous cases), putting the estimated quality of the released labels at 97.9%—a silver standard. The arena corrects 83.1% of synthetic labeling errors at $0.02 per product, reserving human review for panel–pipeline disagreements.

These results support adopting multi-LLM evaluation protocols that incorporate prompt and model diversity, moving beyond single-model approaches. The SynthAVE dataset will be publicly released to support future research in scalable data quality assurance for attribute-value verification.

[2]Pricing based on commercial API rates as of June 2025.

Limitations

Several limitations warrant discussion. First, our evaluation covers four European languages with shared Latin script; generalization to languages with different scripts (e.g., Chinese, Arabic), morphological structures, or lower LLM proficiency requires additional validation. The framework was also evaluated only on e-commerce attribute verification—a task with relatively objective ground truth—and not on an external benchmark. MAVE (Yang et al., 2022) is the closest candidate but is English-only, targets extraction rather than verification, and lacks explicit UNKNOWN labels, so transfer beyond SynthAVE remains open.

Second, the framework depends on the availability of diverse, high-capability LLMs, and specific judge configurations may require updating as model families evolve, retire, or change pricing. That some families substantially underperform others (accuracy ranging from 76% to 93%) raises the question of minimum capability thresholds for judge inclusion. We also aggregate by simple majority vote, a deliberately transparent baseline; given the clustered reliability structure documented in Appendix C.4, reliability-weighted aggregation that down-weights redundant within-family judges is a natural extension we did not explore. Since the panel supplies roughly 2.6 effective votes rather than 21, gains from simply adding judges from existing families should be expected to be small.

Third, our ground truth construction relies on a disagreement-based annotation strategy in which human review was triggered only when the LLM ensemble disagreed with the synthetic label. Consequently 88.9% of the benchmark was accepted without direct adjudication, and for those cases the reference labels are not independent of the arena's own judgments. Expert triage of 400 stratified agreement cases bounds the residual error at 3.0% (0% for unanimous cases), giving an estimated 97.9% accuracy for the released labels (Appendix B)– *silver* rather than *gold*-standard reliability, so comparisons between systems whose scores differ by less than roughly two points should be treated as unresolved (Northcutt et al., 2021). Adjudication of the disagreement cases rested on a single expert. A second annotator from a different team working in a similar catalog area independently re-labelled all 400 audit rows (Appendix A), reproducing the first pass at Cohen's $\kappa = 0.98$ and a slightly more conservative overturn rate against the arena majority. Two caveats attach to this figure: the annotators share the general catalog domain and are expected to agree more than fully external experts, and re-annotating the full 1,416 disagreement cases would be required to bound annotator error on that stratum specifically.

Finally, while the framework is designed for offline validation rather than real-time inference, the latency of running 21 judge configurations may still be prohibitive for rapid iteration cycles. Future work could explore distilling ensemble judgments into smaller, specialized models for faster evaluation.

Ethics Statement

SynthAVE is constructed from synthetic product data generated using the methodology of (Negri et al., 2025), with seed products sampled from commercial e-commerce catalogs. All brand names, model identifiers, and potentially identifying product information were anonymized during generation to prevent privacy concerns and mitigate LLM biases from prior knowledge of specific products—no real customer data or personally identifiable information is present in the released dataset. Human validation was performed by domain experts specialized in e-commerce catalog systems, who established ground truth labels through careful review of disagreement cases between the synthetic generation pipeline and LLM arena predictions. SynthAVE is designed as a test set optimized for evaluation coverage: it spans 2,598 product type-attribute combinations with a minimum of 30 samples per combination, prioritizing breadth over the volume typically required for training. We encourage researchers to leverage this benchmark for downstream applications including validating catalog quality assurance systems, improving attribute coverage in product catalogs, developing robust multilingual extraction methods, and advancing LLM-as-judge evaluation methodologies. We hope SynthAVE contributes to more accurate and reliable e-commerce information systems that benefit both businesses and consumers.

References

Amazon Artificial General Intelligence. 2024. The Amazon Nova family of models: Technical report and model card. Technical report, Amazon.

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. https://www.anthropic.com/news/claude-3-family. Accessed: 2024-03-14.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Sayan De, Debarshi Kumar Sanyal, and Imon Mukherjee. 2025. Fine-tuned encoder models with data augmentation beat ChatGPT in agricultural named entity recognition and relation extraction. *Expert Systems with Applications*, 277:127126.

Shumin Deng, Yubo Ma, Ningyu Zhang, Yixin Cao, and Bryan Hooi. 2024. Information extraction in low-resource scenarios: Survey and perspective. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 33–49.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Enshuo Hsu and Kirk Roberts. 2025. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific Reports*, 15(1):8241.

Mohamad Yaser Jaradeh, Kuldeep Singh, Markus Stocker, Andreas Both, and Sören Auer. 2023. Information extraction pipelines for knowledge graphs. *Knowledge and Information Systems*, 65(5):1989–2016.

G. Kohli. 2026. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *Preprint*, arXiv:2605.29800.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2020. Information extraction meets the semantic web: A survey. *Semantic Web*, 11(2):255–335.

Mistral AI. 2025. Introducing Mistral 3: Mistral Large 3 and the Mistral 3 model family. https://mistral.ai/news/mistral-3. Accessed: 2026-02-04.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Mihai Nadaş, Laura Dioşan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *Preprint*, arXiv:2503.14023.

Virginia Negri, Víctor Martínez Gómez, Sergio A Balanya, and Subburam Rajaram. 2025. Attribute-aware controlled product generation with LLMs for e-commerce. *Preprint*, arXiv:2601.04200.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *Journal of Artificial Intelligence Research*, 71:565–619.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Kassem Sabeh, Mouna Kacimi, Johann Gamper, Robert Litschko, and Barbara Plank. 2024. Exploring large language models for product attribute value identification. *Preprint*, arXiv:2409.12695.

Soham Satyadharma, Fatemeh Sheikholeslami, Swati Kaul, Aziz Umit Batur, and Suleiman A. Khan. 2025. Auto prompting without training labels: An LLM cascade for product quality assessment in e-commerce catalogs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 937–953, Suzhou, China. Association for Computational Linguistics.

Filip Seitl, Tomáš Kovářík, Soheyla Mirshahi, Jan Kryštůfek, Rastislav Dujava, Matúš Ondreička, Herbert Ullrich, and Petr Gronat. 2024. Assessing the quality of information extraction. *Preprint*, arXiv:2404.04068.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024.

Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Dey, and 1 others. 2025. LiveBench: A challenging, contamination-limited LLM benchmark. In *Proceedings of the International Conference on Learning Representations*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. MAVE: A product dataset for multi-source attribute value extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 1256–1265, New York, NY, USA. Association for Computing Machinery.

Jiayi Yuan, Hongyi Li, Xingyu Ding, Wenting Xie, Yu-Jhe Li, Wenxuan Zhao, Kehan Wan, Junda Shi, Xuan Hu, and Ziniu Liu. 2025. Understanding and mitigating numerical sources of nondeterminism in LLM inference. *Preprint*, arXiv:2506.09501.

Bryan Zhang, Suleiman Khan, and Stephan Walter. 2025. Leveraging product catalog patterns for multilingual E-commerce product attribute prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 267–275, Suzhou (China). Association for Computational Linguistics.

Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, and Jiawei Han. 2024. Automated mining of structured knowledge from text in the era of large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6644–6654.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

A Human Triage of Agreement Cases

To validate the assumption that agreement between the synthetic generation pipeline and the LLM arena majority vote implies label correctness, an expert annotator independently triaged a stratified random sample of 400 agreement cases. The sample was balanced across locales (100 per language) and stratified by agreement level (100 unanimous, 300 mixed-agreement).

For each sample, the annotator reviewed the product text, attribute-value pair, and majority vote label, then recorded whether they *agree* with the majority, *disagree* (i.e., would overturn it), or are *unsure*. The audit returned 382 agreements, 12 disagreements, and 6 unsure judgements.

Handling of Unsure Cases. Only clear disagreements are counted as overturns. Because the treatment of the six unsure cases is a modelling choice rather than an observation, we report all three defensible conventions: **3.0%** (12/400), counting them as non-overturns, which is the figure used in the main text; **3.0%** (12/394), excluding them from numerator and denominator alike; and **4.5%** (18/400), the conservative upper bound obtained by counting every unsure case as an overturn. The conventions agree to within 1.5 percentage points, so no conclusion in this paper depends on the choice. Substituting the upper bound would raise the estimated residual error of the released labels (Appendix B) from 2.1% to 3.0%, leaving the silver-standard characterisation intact.

Overall Results. The annotator disagreed with the majority vote in 12 out of 400 cases (3.0%), with a 95% confidence interval of [1.3%, 4.7%] (margin of error ±1.7%).

Table 7: Expert triage overturn rates by slice. CI values are percentages.

SLICE	RATE	95% CI	n
OVERALL	3.0%	[1.3, 4.7]	400
By sample kind			
MIXED	4.0%	[1.8, 6.2]	300
UNANIMOUS	0.0%	—	100
By majority label			
UNKNOWN	5.0%	[1.6, 8.3]	161
CORRECT	2.2%	[0.1, 4.4]	181
INCORRECT	0.0%	—	58

Breakdown by Locale. Table 8 shows the expert triage overturn rates by locale.

Key Findings.

- The majority panel achieves a 97% accuracy rate (12/400 overturns) with a tight confidence

Table 8: Expert triage overturn rates by locale

LOCALE	OVERTURNS	RATE	n
IT	5	5.0%	100
DE	3	3.0%	100
ES	3	3.0%	100
FR	1	1.0%	100

interval of [1.3%, 4.7%].

- Unanimous cases have a 0% overturn rate (0/100): when all 21 judges agree, the annotator found nothing to correct in our sample.
- UNKNOWN is the least reliable majority label at 5.0% overturn rate, compared to CORRECT at 2.2% and INCORRECT at 0.0%. This is consistent with UNKNOWN being the conservative vote that can mask determinable cases.
- All 12 overturns came from mixed-agreement samples (4.0% rate); no unanimous sample was overturned.
- Locale differences range from 1.0% (FR) to 5.0% (IT), but with n=100 per locale the margins of error (2–4%) overlap substantially. Distinguishing locale-level effects would require ∼1,000+ samples per locale.

On the 0% Unanimous Overturn Rate. One caveat deserves emphasis. For agreement cases the reference label *is* the original synthetic label by construction, so a unanimous panel that matches it cannot disagree with the reference by definition. What makes the 0% figure informative is the audit itself: an independent annotator, working from the product text rather than from the panel's output, found nothing to correct in 100 unanimous samples. The number should therefore be read as evidence that unanimity is a strong confidence signal, not as a measured classification accuracy of 100%.

Second Expert Evaluation. To calibrate the reliability of the reference labels themselves, a second annotator independently re-labelled all 400 audit samples. Both annotators are researchers on this work with multiple years of experience on multilingual product-catalog attribute problems, drawn from two different teams operating in adjacent areas of catalog attribute research. The second annotator worked from a blinded interface in which the first annotator's label, the arena majority vote and the original triage decision were hidden. The annotators do not share day-to-day working conventions, but they do share the general catalog domain and the attribute taxonomy the benchmark is built on, so this study should be read as agreement between two experienced practitioners of the domain rather than as agreement with a fully external annotator.

Table 9: Second-expert audit: agreement with the first annotator

SLICE	RAW AGREEMENT	COHEN κ	DISAGR.
DE	98.0%	0.968	2
ES	98.0%	0.968	2
FR	99.0%	0.984	1
IT	100.0%	1.000	0
ALL (400)	**98.8%**	**0.980**	**5**

Across the 400 audit rows the two annotators agree on 395, giving Cohen's $\kappa = 0.98$ (95% CI $[0.96, 1.00]$), which sits in the almost-perfect band of Landis and Koch (1977). Applying the second annotator's judgements against the arena majority yields an overturn rate of 5/400 = 1.2% overall, with 0/100 on unanimous samples and 5/300 = 1.7% on mixed-agreement samples—consistent with, and slightly more conservative than, the 4.0% mixed-sample overturn rate reported for the first pass. The first-pass 3.0% figure is therefore reproducible, and both annotators would classify the audit set at 98–99% accuracy.

The five human–human disagreements all occurred in mixed-agreement samples; four out of the five involve the arena's INCORRECT or UNKNOWN calls, where the first annotator sided with the arena and the second annotator would overturn it. Three cases (`washer_type` and `thread_type` in ES, and `guitar_bridge_system` in DE) flip an arena INCORRECT/UNKNOWN call to CORRECT under a broader reading of the value; one DE `max_printspeed_color` case flips INCORRECT to UNKNOWN; and one FR `section_width` case flips a CORRECT call to UNKNOWN because the value is not explicitly stated in the text. Rather than silently override either annotator, we retain the first pass's labels as the released reference and treat the five residual disagreements as documented uncertainty; the point estimate of ∼97% agreement-case accuracy therefore sits between the two annotators' independent overturn rates (1.2% and 3.0%), and the confidence interval in Appendix B accommodates the difference.

Implications. The 3% ± 1.7% overturn rate supports the claim that the majority panel correctly labels ∼97% of agreement cases. Combined with the 0% overturn rate for unanimous cases, this validates the disagreement-based annotation strategy: agreement between the generation pipeline and a diverse 21-judge panel is a reliable signal of label correctness, with residual error concentrated in inherently ambiguous mixed-agreement cases. Appendix B converts these rates into an unconditional estimate of the quality of the released labels.

B Estimating Label Quality

The 92.6% figure reported for the synthetic pipeline in Section 5 is *conditional*: it counts only errors surfaced in disagreement cases and treats agreement cases as correct by construction. Because the expert triage of Appendix A measures a nonzero error rate inside the agreement stratum, we can replace this with an unconditional, stratified estimate.

B.1 Stratified Estimator

Partition the benchmark into the agreement stratum A and the disagreement stratum D, with $P(A) = 88.9\%$ and $P(D) = 11.1\%$. The overall error rate is

$$\varepsilon = P(A)\,\varepsilon_A + P(D)\,\varepsilon_D.$$

Within A, we split further by consensus level. Unanimous samples make up 40.8% of the stratum and show an observed overturn rate of 0% (0/100 audited); mixed-agreement samples make up the remaining 59.2% and show 4.0% (12/300 audited). Hence

$$\varepsilon_A = 0.408 \times 0.0 + 0.592 \times 0.040 = 2.37\%.$$

B.2 Before Expert Review

Before adjudication, ε_D is the rate at which the synthetic label is wrong among cases the arena contests, estimated at 66.8% from the fixes applied to this stratum. This gives

$$\varepsilon = 0.889 \times 2.37\% + 0.111 \times 66.8\% = 9.5\%,$$

i.e. an unconditional synthetic-label accuracy of **90.5%**, roughly two points below the conditional 92.6% figure. The gap is entirely attributable to residual error in cases that the arena and the generator agreed upon and that were therefore never inspected.

B.3 After Expert Review

After expert adjudication, the disagreement stratum is human-labelled and contributes no residual error under the assumption that the adjudicating expert is correct, so $\varepsilon_D \approx 0$ and

$$\varepsilon = 0.889 \times 2.37\% + 0.111 \times 0\% = 2.1\%,$$

giving an estimated accuracy of **97.9%** for the released labels. Propagating the binomial interval on the 12/300 mixed-sample overturn rate through the same weighting yields a 95% confidence interval of [96.7%, 99.0%].

Two caveats apply. First, $\varepsilon_D \approx 0$ assumes the adjudicating expert is authoritative; with a single annotator we cannot bound this from human–human agreement, which is the principal limitation of the protocol. Second, the estimate inherits the sampling error of a 400-case audit, so the interval is wider than the point estimate suggests.

Table 10: Estimated label quality before and after expert review

STRATUM	SHARE	ERR. BEFORE	ERR. AFTER
AGREEMENT	88.9%	2.37%	2.37%
UNANIMOUS	40.8%	0.0%	0.0%
MIXED	59.2%	4.0%	4.0%
DISAGREEMENT	11.1%	66.8%	≈0%
OVERALL ERROR		**9.5%**	**2.1%**
ACCURACY		**90.5%**	**97.9%**

B.4 Gold, Silver and Bronze Standards

It is useful to place this figure on the conventional scale of reference-label quality. *Gold standard* labels are manually verified and adjudicated by trusted domain experts, approaching 100% accuracy at high cost. *Silver standard* labels are produced with high confidence but without exhaustive human verification, typically by consensus among multiple annotators or a highly reliable automated procedure; they retain occasional noise. *Bronze standard* labels come from heuristics, dictionary lookups, or weak models, and carry substantial noise in exchange for speed.

At an estimated 2.1% residual error, SynthAVE sits at the upper end of the silver band: cleaner than the average of the widely used benchmarks audited by Northcutt et al. (2021), but short of the near-perfect agreement expected of a gold standard. This distinction matters for downstream use. Northcutt et al. (2021) show that even small test-set label

error rates can reorder model rankings, so benchmark consumers should treat differences between systems that fall within a couple of points as unresolved. Closing the remaining gap would require adjudicating the mixed-agreement portion of the agreement stratum, which our triage identifies as the sole location of residual error; the unanimous portion required no corrections in 100 audited samples.

C Detailed Results

C.1 Agreement Level and Accuracy

Table 11 shows arena accuracy stratified by the proportion of judges agreeing on the majority vote.

Table 11: LLM Arena accuracy by model agreement level

Agreement Level	Products	Accuracy
Low (<50%)	172	65.1%
Medium (50–70%)	1,198	79.5%
High (70–85%)	1,398	91.1%
Very High (>85%)	9,958	98.8%
Unanimous (100%)	4,612	**100.0%**

C.2 Data Cleaning Performance by Language

Table 12 presents detailed cleaning performance by language, including fix rates and regression rates.

Table 12: Data cleaning performance by language

Lang.	Orig Acc	Arena Acc	Fix	Regr.
IT	91.9%	96.4%	90.2%	3.0%
DE	93.8%	95.3%	93.2%	4.6%
ES	92.5%	94.7%	75.8%	3.8%
FR	92.1%	94.0%	76.1%	4.5%

C.3 Error Pattern Distribution

Table 13 breaks down the arena's 630 errors by confusion pattern. The dominant pattern is INCORRECT → UNKNOWN (27.5%), indicating conservative abstention. Confusion involving UNKNOWN accounts for 80.5% of all errors, while direct CORRECT ↔ INCORRECT misclassification is rare (19.5%).

C.4 Panel Size and Composition

Table 14 reports majority-vote accuracy as a function of panel composition, averaging over every subset of a given size so that the numbers reflect a typical rather than a hand-picked panel. This sweep is recomputed from the per-judge prediction files released with the dataset; it reproduces the full-arena accuracy of Table 3 to within 0.2 percentage points, and all comparisons below are between configurations evaluated on the same basis.

Table 13: Error pattern distribution (Human → Arena)

Error Pattern	Count	% of Errors
INCORRECT → UNKNOWN	173	27.5%
UNKNOWN → CORRECT	147	23.3%
UNKNOWN → INCORRECT	128	20.3%
INCORRECT → CORRECT	88	14.0%
CORRECT → UNKNOWN	59	9.4%
CORRECT → INCORRECT	35	5.6%
Total Errors	**630**	**100%**

Table 14: Accuracy by panel composition. Family subsets are averaged over all $\binom{7}{k}$ combinations.

Composition	Judges	Mean	Min	Max
Family diversity (k families × 3 prompts)				
1 family	3	89.9	81.0	92.9
2 families	6	92.7	90.4	93.8
3 families	9	93.7	91.9	94.5
4 families	12	94.2	93.4	94.7
5 families	15	94.6	94.1	95.0
6 families	18	94.9	94.7	95.0
7 families	21	95.2	—	—
Prompt diversity (1 family, p prompts)				
1 prompt	1	88.0	76.3	92.1
2 prompts	2	88.8	76.8	92.5
3 prompts	3	89.9	81.0	92.9
Family diversity at fixed prompt				
3 families	3	91.8	—	93.4
7 families	7	93.2	—	—

Three conclusions follow. First, no single configuration reaches ensemble accuracy: the best individual judge attains 92.1%, roughly three points below the full panel. Second, family diversity dominates prompt diversity at equal cost. Holding the judge budget at three, three families with one prompt each reaches 91.8% while one family with three prompts reaches 89.9%; adding prompt variants within a family buys under two points in total, whereas adding families buys more than five. Third, returns diminish but do not vanish: a nine-judge, three-family panel reaches 93.7% on average and 94.5% at best, within roughly one point of the full arena at 9% of the cost (Appendix H). For deployments where a one-point accuracy loss is acceptable, this is the configuration we would

recommend; pruning prompts is safer than pruning families.

C.5 Judge Dependence and Effective Votes

Because the panel aggregates correlated judges, the nominal count of 21 overstates the independent evidence available. Following Kohli (2026), we estimate the effective number of votes from the mean pairwise correlation $\bar{\rho}$ of judge error indicators as $n_{\text{eff}} = n/(1 + (n-1)\bar{\rho})$.

We measure $\bar{\rho} = 0.36$ across all 210 judge pairs, decomposing into 0.53 for within-family pairs and 0.34 for cross-family pairs. This yields $n_{\text{eff}} \approx 2.6$ effective votes from 21 configurations. The interpretation is not that the panel is ineffective—it outperforms every individual judge—but that its reliability should be attributed to the diversity of a handful of genuinely distinct error profiles rather than to the nominal panel size. The within- versus cross-family split also explains the composition results above: within-family judges are largely redundant, so the marginal value of a new prompt is small relative to that of a new family.

C.6 Error Distribution by Category, Attribute and Language

The dominant production risk is a false CORRECT, in which the panel asserts a value the product text does not support. Table 15 reports the categories and attributes where the arena is least reliable.

Table 15: Highest and lowest error rates by product category and attribute (slices with $n \geq 60$)

SLICE	n	ERROR
Categories, highest error		
WALL_ART	62	14.5%
AIMING_SCOPE_SIGHT	66	10.6%
CLOCK	66	10.6%
BINOCULAR	69	10.1%
TABLE	89	9.0%
TELEVISION	98	8.2%
Categories, lowest error		
KICK_SCOOTER	105	1.0%
SWIMWEAR	98	1.0%
MICROPHONE	60	0.0%
Attributes, highest error		
HANDLE.MATERIAL	106	7.5%
CONTAINER.TYPE	100	6.0%
NUMBER_OF_POCKETS	71	5.6%
POWER_SOURCE_TYPE	76	5.3%
DISPLAY.TYPE	150	4.7%

Errors concentrate where the attribute is a fine-grained physical property that product copy tends to describe loosely. `handle.material` and `container.type` are the clearest instances: descriptions mention a material or container in passing without specifying the one being asked about, and judges over-commit to CORRECT on the basis of a nearby but different mention. Decorative and optical categories (WALL_ART, AIMING_SCOPE_SIGHT, BINOCULAR) show the same pattern for framing, mounting and coating attributes.

Representative false-CORRECT cases illustrate the mechanism. For a wall clock with attribute `alarm_clock` and generated value `FALSE`, the text neither confirms nor denies an alarm function; the correct label is UNKNOWN, but the panel returned CORRECT. For a safety-glasses product with `frame.type` = "flexible fit", the description praises comfort without characterising the frame; again UNKNOWN, again asserted CORRECT. For a utility cart with `caster.type` = "universal front wheels", the text describes wheels without specifying caster type, and the panel confirmed a value the ground truth marks INCORRECT.

False CORRECT rates also vary by language, from 1.4% in German to 2.5% in Italian, tracking the overall per-language ordering in Table 12. Practitioners deploying this pipeline should apply tighter thresholds, or route to human review, on attribute families of this kind rather than applying a uniform confidence policy across the catalog.

D Dataset Statistics

This appendix provides comprehensive statistics for the SynthAVE test set we release with this paper.

D.1 Overall Scale

- **Total products**: 12,726 unique products
- **Product categories**: 229 distinct types
- **Unique attributes**: 792 across all languages
- **Product-attribute combinations**: 2,607 unique pairs
- **Average entries per combination**: 4.9
- **Languages**: 4 (Spanish, French, Italian, German)
- **Products per language**: 3,007–3,501
- **Judge configurations**: 21 (7 models $\times$ 3 prompts)

D.2 Label Distribution

Table 16: Overall label distribution (majority vote)

LABEL	COUNT	%
CORRECT	6,214	48.8
UNKNOWN	4,602	36.2
INCORRECT	1,910	15.0
TOTAL	12,726	100.0

Table 16 presents the overall label distribution based on majority vote across all judges.

D.3 Product Category Distribution

Table 17 presents the top 20 product categories by volume.

Table 17: Top 20 product categories by count (of 229 total)

CATEGORY	COUNT	%
NOTEBOOK_COMPUTER	209	1.6
CELLULAR_PHONE	190	1.5
HANDBAG	180	1.4
SHOES	140	1.1
PRINTER	138	1.1
WATCH	133	1.0
CAMERA_DIGITAL	127	1.0
CHAIR	123	1.0
SECURITY_CAMERA	122	1.0
WEARABLE_COMPUTER	121	1.0
DRESS	118	0.9
LIGHT_FIXTURE	115	0.9
CABINET	111	0.9
KICK_SCOOTER	110	0.9
BACKPACK	110	0.9
VIDEO_PROJECTOR	108	0.8
BRA	107	0.8
PERSONAL_COMPUTER	107	0.8
HEADPHONES	107	0.8
MONITOR	106	0.8
...209 additional categories		

D.4 Attribute Distribution

Table 18 presents the top 20 attributes by frequency.

D.5 Top Product-Attribute Combinations

Table 19 presents the most frequent product-attribute combinations in the dataset.

D.6 Per-Language Statistics

Table 20 provides detailed per-language statistics including unique product types, attributes, and combinations.

Table 18: Top 20 attributes by count (of 792 total)

ATTRIBUTE	COUNT	%
CLOSURE.TYPE	296	2.3
FRAME.MATERIAL	290	2.3
BRAND	237	1.9
OUTER.MATERIAL	185	1.5
MATERIAL	170	1.3
DISPLAY.TYPE	163	1.3
COLOR	132	1.0
WAIST.SIZE	121	1.0
BATTERY.AVERAGE_LIFE	115	0.9
NECK.NECK_STYLE	115	0.9
HANDLE.MATERIAL	111	0.9
CONTAINER.TYPE	108	0.8
TOP.MATERIAL	107	0.8
CHEST.SIZE	106	0.8
SLEEVE.TYPE	100	0.8
DISPLAY.SIZE	88	0.7
SPECIAL_FEATURE	87	0.7
SIZE	86	0.7
WEIGHT_CAPACITY.MAXIMUM	84	0.7
METAL_TYPE	81	0.6
...772 additional attributes		

D.7 Ground Truth Distribution

Reference labels come from expert adjudication for the disagreement cases and from the accepted synthetic label for the agreement cases (Section 4). Table 21 shows the resulting distribution.

D.8 Effect of Validation on the Label Mix

Table 22 compares the original synthetic labels with the final reference labels. Validation is a mild correction rather than a redistribution: CORRECT gains 93 samples (+0.7 points), INCORRECT loses 232 (-1.8), and UNKNOWN gains 139 (+1.1). This is expected, because the generation protocol fixes a target label for each sample in advance, so validation corrects execution errors rather than re-balancing the design. The largest movement is out of INCORRECT, consistent with the per-class results in Appendix C, where INCORRECT is both the least accurate original label and the class the arena most often revises towards UNKNOWN.

Practical utility. Because the label mix is set by design, SynthAVE is intended as a diagnostic test set: per-class scores should be re-weighted by the label prior of the target application before being read as expected error rates. What transfers directly is the relative ordering of systems and their behaviour on the hard slices identified in Appendix C.

Table 19: Top 20 product-attribute combinations (of 2,607 total)

Product Type	Attribute	N
RADIO	DISPLAY.TYPE	27
THERMOMETER	OUTER.MATERIAL	25
ELEC._CIGARETTE	BATTERY.POWER	25
BLANKET	BLANKET_FORM	24
SUITCASE	HANDLE.TYPE	24
ENVELOPE	CLOSURE.TYPE	23
CLOTHES_RACK	FRAME.MATERIAL	23
AIMING_SCOPE	HAS_NIGHT_VISION	23
WALKING_STICK	BASE.MATERIAL	22
SEASONING	CONTAINER.TYPE	22
HANDBAG	MATERIAL	22
POWER_BANK	BATTERY.CELL_COMP	22
INCONT._PROT.	INCONT._PROT._TYPE	21
OUTBUILDING	FRAME.MATERIAL	21
WATCH_BAND	WATCH_BAND_DESIGN	21
PICTURE_FRAME	FRAME.MATERIAL	21
SELF_BAL._VEH.	WHEEL.SIZE	20
WALLET	WALLET_COMP._TYPE	20
WATCH	DIAL.COLOR	20
SUNGLASSES	POLARIZATION_TYPE	20

Table 20: Detailed statistics by language

	ES	FR	IT	DE
Total entries	3,159	3,501	3,007	3,059
Product types	227	229	225	229
Attributes	636	675	609	673
Combinations	1,443	1,505	1,267	1,483
Majority Vote Counts				
CORRECT	1,536	1,710	1,477	1,491
INCORRECT	482	553	437	438
UNKNOWN	1,141	1,238	1,093	1,130

E Inter-Judge Agreement Heatmaps by Language

This appendix presents pairwise agreement heatmaps for all four languages. Figure 2 in the main text shows German (Cohen's κ). Below we provide the complete set for all languages and both agreement metrics.

E.1 Cohen's Kappa Heatmaps

Figures 3, 4, 5

E.2 Raw Agreement Heatmaps

Figures 6, 7, 8, 9

F Evaluation Prompt Templates

This appendix provides the complete prompt templates used for our LLM Arena evaluation across all three prompt versions. Each version maintains the same evaluation task (attribute consistency verification) while employing different structural approaches and reasoning strategies to ensure IID generation requirements.

Table 21: Ground truth label distribution (human-verified)

Ground Truth	Count	%
CORRECT	6,073	47.7
UNKNOWN	4,645	36.5
INCORRECT	2,008	15.8
Total	12,726	100.0

Table 22: Original synthetic vs. final reference label distribution

Label	Original	%	Final	%
CORRECT	5,980	47.0	6,073	47.7
INCORRECT	2,240	17.6	2,008	15.8
UNKNOWN	4,506	35.4	4,645	36.5
Total	12,726	100.0	12,726	100.0

F.1 Prompt Version 1: Direct Instruction

F.2 Prompt Version 2: XML-Structured with Examples

F.3 Prompt Version 3: Role-Based with Systematic Guidelines

G Evaluation Settings

G.1 Judge Qualification Criteria

Each LLM judge must meet qualification criteria that parallel traditional standards for human auditors. *General competency* establishes baseline capability: models must meet minimum performance thresholds on established benchmarks, such as Global Average $\geq 70\%$ on LiveBench (White et al., 2025), paralleling education or certification requirements for human auditors. *Task-specific competency* ensures domain relevance: models must demonstrate sufficient understanding of attribute verification through internal benchmark evaluation, analogous to subject matter expertise. *Self-consistency* addresses output reliability: given known variability in LLM outputs due to floating-point precision issues (Yuan et al., 2025), judges must achieve minimum consistency when generating responses for identical prompts. *Instruction adherence* ensures usable outputs: models must reliably follow specified output formats, parallel-

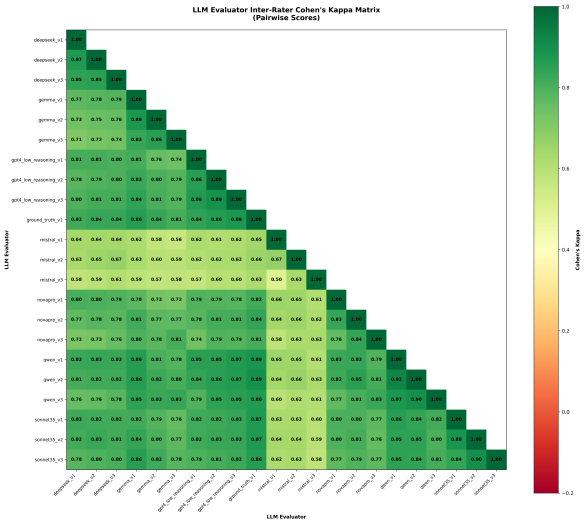

Figure 3: Pairwise Cohen's κ between judge configurations for Italian. Clustering patterns are consistent with other languages, with the same model families forming high-agreement and low-agreement groups.

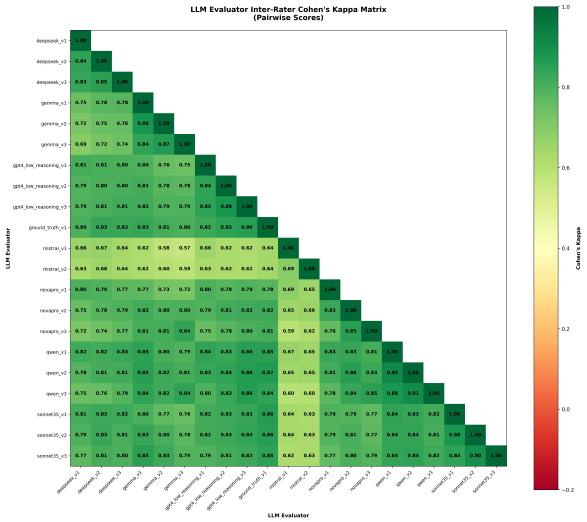

Figure 4: Pairwise Cohen's κ between judge configurations for French. Despite French showing the lowest overall arena accuracy (94.0%), inter-judge agreement patterns remain consistent with other languages.

ing structured reporting requirements for human evaluators.

We stress that these criteria govern the *competence* of individual judges, not their statistical independence. Qualified judges may still fail in correlated ways, and Appendix C quantifies the extent to which they do.

G.2 Input Format

Figure 13 shows the input format provided to each judge.

G.3 Model Configuration

All models use a low temperature (0.1) to ensure consistent, reproducible outputs. We leave default settings for Top-p and Top-k parameters.

For OpenAI's GPT OSS reasoning model, we set `reasoning_effort="low"` to balance computational cost with performance. This parameter controls the depth of the model's internal chain-of-thought reasoning before producing a final answer. Even at the "low" setting, the model demonstrates strong performance on our classification task while significantly reducing inference costs compared to higher reasoning effort levels.

Model-Specific Formats. Each model family requires different API message formats:

- **Claude**: Uses Anthropic's message format with `anthropic_version` specification
- **Nova**: Requires nested content structure with explicit text fields
- **OpenAI**: Standard chat completion format with `reasoning_effort` parameter
- **Mistral**: Uses instruction-tuned format with `[INST]` tags
- **DeepSeek/Qwen3/Gemma**: Standard chat message format

G.4 Response Parsing

Models return responses in different formats requiring specialized parsing:

- **Standard format** (Claude, Nova, Mistral, DeepSeek, Qwen3, Gemma): First line contains the classification label (CORRECT, INCORRECT, or UNKNOWN), followed by reasoning on subsequent lines.
- **OpenAI reasoning format**: Response contains `<reasoning>...</reasoning>` tags encapsulating the model's chain-of-thought, followed by the final classification label. The parser extracts both components separately.

Invalid or unparseable responses default to UNKNOWN with the full response content logged for debugging.

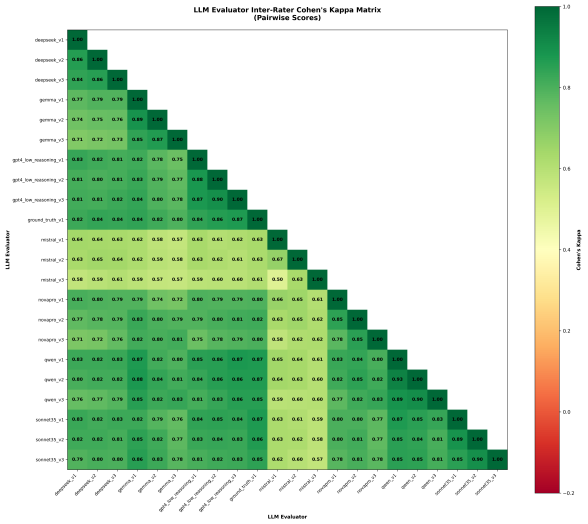

Figure 5: Pairwise Cohen's κ between judge configurations for Spanish.

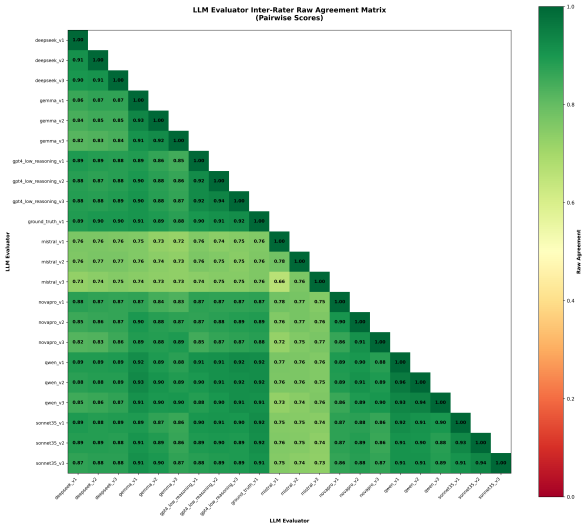

Figure 6: Pairwise raw agreement between judge configurations for Spanish. Agreement ranges from 75% to 96% (within-family pairs).

G.5 Evaluation Pipeline

For each product and judge configuration:

1. **Input formatting**: Product data formatted according to model-specific API requirements
2. **API invocation**: Model called via AWS Bedrock with configured parameters
3. **Response parsing**: Extract classification and reasoning from model output, handling format variations
4. **Validation**: Verify response contains valid classification (CORRECT, INCORRECT, or UNKNOWN); invalid responses default to UNKNOWN
5. **Error handling**: API failures and malformed responses are caught and logged, with graceful degradation to UNKNOWN
6. **Storage**: Save classification, reasoning, model ID, prompt version, and timestamp

All models were accessed through AWS Bedrock's unified API, ensuring consistent infrastructure across different model providers. The pipeline processes 21 judge configurations (7 models $\times$ 3 prompt variants) per product systematically.

G.6 Aggregation Details

Tie resolution. With 21 judges and three labels, a strict tie is possible (for example a 9/9/3 split) but rare, affecting well under 1% of samples. Ties are resolved in favour of the label supported by the highest-accuracy judge among those tied, evaluated on the audited subset. Because tied cases are by construction low-agreement, they fall in the stratum we recommend routing to human review, and the choice of tie-breaking rule has a negligible effect on aggregate accuracy.

Malformed generator output. A small fraction of samples emerged from the generation pipeline without a usable three-way label, either because the generator emitted no label, produced an unparseable value, or attached the attribute to a mismatched product identifier. These samples are counted as disagreement cases and routed to expert adjudication, since no synthetic label exists for the arena to agree with. Practitioners reusing this pipeline should expect a comparable residue and budget adjudication capacity for it.

Judge-level failures. Invalid or unparseable judge responses default to UNKNOWN, as described above. This is a conservative choice that pushes affected judges towards abstention rather than towards a substantive label, and it slightly depresses measured per-judge accuracy relative to a design that discarded such responses.

H Cost Analysis Details

This appendix provides detailed cost breakdowns for the LLM Arena validation framework. All costs are based on AWS Bedrock on-demand pricing as of January 2026.[3]

[3] https://aws.amazon.com/bedrock/pricing/

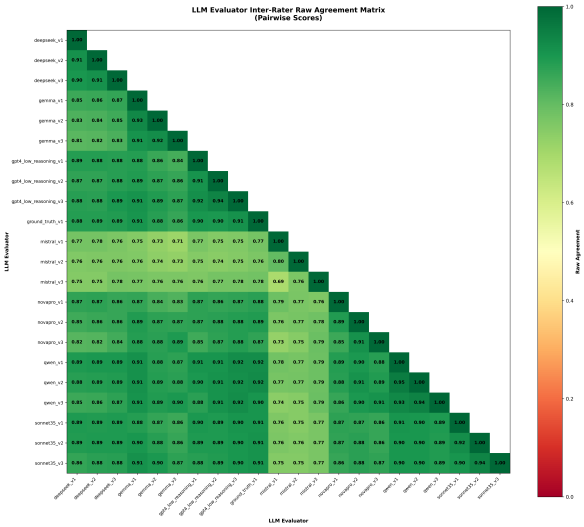

Figure 7: Pairwise raw agreement between judge configurations for German. Agreement ranges from approximately 75% (cross-family pairs with lower-performing models) to 96% (within-family pairs).

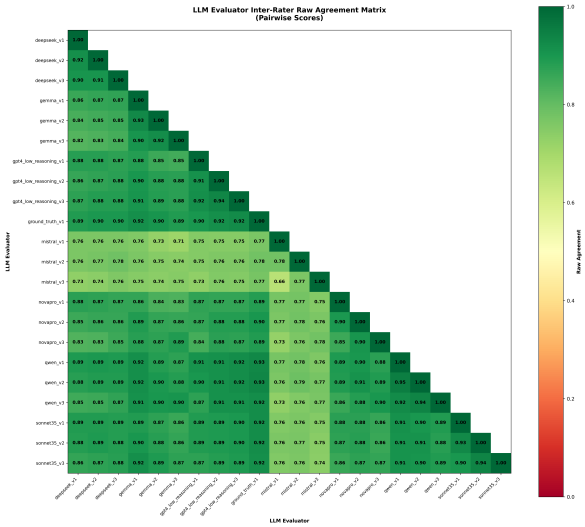

Figure 8: Pairwise raw agreement between judge configurations for Italian. Italian shows the highest overall agreement, consistent with its leading arena accuracy (96.4%).

H.1 Token Statistics

Each API call consists of input tokens (the prompt with product data) and output tokens (the model's classification and reasoning). Input token counts vary based on prompt version and product text length; output tokens are estimated at 50 tokens per response (classification label plus brief reasoning).

Table 23: Average input tokens per prompt by language and version

Lang	Ver	Mean	Med.	Min	Max	Std
DE	V1	727	664	272	2554	308
	V2	1095	1031	643	2925	308
	V3	1229	1167	776	3060	309
ES	V1	688	637	273	2229	285
	V2	1056	1005	639	2596	285
	V3	1191	1140	778	2733	285
FR	V1	716	671	274	2222	291
	V2	1084	1039	638	2587	290
	V3	1219	1174	775	2723	291
IT	V1	714	665	274	2079	301
	V2	1082	1032	644	2446	300
	V3	1217	1167	780	2581	300

Table 23 presents input token statistics. The three prompt versions differ in complexity: V1 provides basic instructions, V2 adds attribute descriptions, and V3 includes worked examples. This is reflected in token counts, with V3 prompts averaging 70% more tokens than V1.

Token counts vary across languages due to differences in product text length and linguistic characteristics. German (DE) has the highest average token counts, while Spanish (ES) has the lowest. The standard deviation of approximately 285–308 tokens reflects the diversity in product description lengths across the dataset.

H.2 Full Arena Costs

Table 24: Full arena costs by language

Lang.	Products	API Calls	$/Prod.	Total
IT	3,007	63,147	$0.023	$68.80
DE	3,059	64,239	$0.023	$70.69
ES	3,159	66,339	$0.022	$70.76
FR	3,501	73,521	$0.023	$80.25
All	**12,726**	**267,246**	**$0.023**	**$290.50**

Table 24 shows the per-language cost breakdown for running the full 21-judge arena (7 models $\times$ 3 prompt versions). Costs scale linearly with dataset size, with French being most expensive due to having the most products.

H.3 Scaling Projections

Table 25 projects arena costs at various scales. At production scale (1M products), the full arena would cost approximately $22,825—substantially less than equivalent human annotation, which would cost $100,000–500,000 at typical rates of $0.10–0.50 per label.

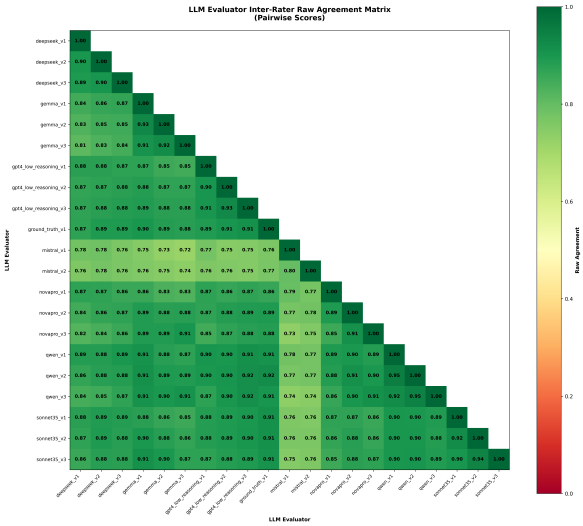

Figure 9: Pairwise raw agreement between judge configurations for French.

Table 25: Projected arena costs at scale

PRODUCTS	API CALLS	COST	$/PROD.
1K	21K	$23	$0.023
10K	210K	$228	$0.023
100K	2.1M	$2,283	$0.023
1M	21M	$22,825	$0.023

H.4 Cost-Efficient Configurations

For budget-constrained deployments, reduced ensemble configurations offer significant savings with potential accuracy trade-offs.

Table 26: Alternative ensemble configurations (per 1,000 products)

CONFIGURATION	JUDGES	$/1K	SAVINGS
FULL ARENA (7×3)	21	22.83	–
EXCL. HIGHEST-COST	18	11.57	49%
TOP-3 FAMILIES (3×3)	9	2.08	91%
SINGLE FAMILY (3 PROMPTS)	3	0.54–11.26	51–97%
SINGLE CONFIG	1	0.14–4.39	81–99%

Table 26 presents alternative ensemble configurations. Cost varies substantially across model families, with open-weight models typically 10–30× cheaper than commercial APIs. Using a single model family with all three prompt versions represents 3–49% of full arena cost depending on family choice.

The trade-off between cost and accuracy presents optimization opportunities. Unanimous agreement cases (36% of products) could use a smaller initial ensemble, with full arena deployment reserved for uncertain cases. Such adaptive strategies could reduce average cost by 30–50% while maintaining high accuracy on confident predictions. However, ablation studies validating accuracy of reduced configurations are required before deployment.

```
You are an expert product attribute evaluator. Your task is to determine whether a generated attribute label is correct,
incorrect, or unknown based on the provided product information.

PRODUCT INFORMATION:
ID: {id}  |  Category: {category}  |  Title: {title}
Product Description: {description}
Bullet Points: {bullet_points}
ATTRIBUTE: {attribute}  |  GENERATED LABEL: {generated_label}

TASK: Evaluate whether the generated label "{generated_label}" is accurate for the attribute "{attribute}" based on the product
information provided. Note that some product information (description or bullet points) may not be available.

RESPONSE OPTIONS:
- CORRECT: The generated label accurately reflects the product's attribute based on the available information
- INCORRECT: The generated label is clearly wrong based on the product information
- UNKNOWN: Cannot determine if the label is correct or incorrect from the available product information
  (including cases where insufficient information is provided)

RESPONSE FORMAT: You must respond with exactly one of: "CORRECT", "INCORRECT", or "UNKNOWN".
After your choice, in a new line, provide a brief explanation (1-2 sentences) of your reasoning.

Example response format:
CORRECT
The product description explicitly mentions this attribute value.
```

Figure 10: Prompt Version 1: Direct instruction format with minimal structure.

```
You are an e-Commerce data analyst specializing in product attribute verification and quality assurance.
Assess whether a generated attribute label correctly represents the specified product attribute based on available listing information.

<examples>
<example>
ID: B08N5WRWNW | Category: LUGGAGE
Title: <brandname> <makename> Hardside Expandable Luggage with Spinner Wheels, Brushed Anthracite, Carry-On 20-Inch
Description: Durable polycarbonate shell construction with scratch-resistant texture
Bullet Points: * 100% Polycarbonate shell * Four multidirectional spinner wheels * TSA-approved combination lock
Attribute Name: material_type | Generated Label: Polycarbonate
Your Response:
CORRECT
The description and bullet points explicitly confirm polycarbonate shell construction, validating the material type label.
</example>
<example>
ID: B07XYZ123A | Category: CLOTHING
Title: Men's Cotton Blend Casual T-Shirt Blue Medium | Description: [Not provided] | Bullet Points: [Not provided]
Attribute Name: sleeve_length | Generated Label: Long Sleeve
Your Response:
UNKNOWN
Title mentions t-shirt but provides no sleeve length information, and additional product details are unavailable.
</example>
<example>
ID: B09ABC456D | Category: ELECTRONICS
Title: Wireless Bluetooth Headphones with Noise Cancellation
Description: Premium over-ear headphones featuring active noise cancellation technology
Bullet Points: * 30-hour battery life * Bluetooth 5.0 connectivity * Foldable design for portability
Attribute Name: connectivity_technology | Generated Label: USB-C
Your Response:
INCORRECT
Product clearly indicates Bluetooth connectivity, directly contradicting the USB-C label assignment.
</example>
</examples>

<inputs>
<id>{id}</id> <category>{category}</category> <item_name>{title}</item_name>
<description>{description}</description> <bullet_points>{bullet_points}</bullet_points>
<attribute_name>{attribute}</attribute_name> <generated_label>{generated_label}</generated_label>
</inputs>

<instructions>
Evaluate the generated attribute label by:
1. Reviewing all available product information for evidence relating to the specified attribute
2. Determining if the generated label matches, conflicts with, or cannot be verified from the product data
3. Handling incomplete information appropriately when description or bullet points are missing
Classify as:
- CORRECT: Label accurately reflects attribute based on available evidence
- INCORRECT: Label contradicts information in product details
- UNKNOWN: Insufficient data to verify label accuracy
Response format: Respond with exactly one of "CORRECT", "INCORRECT", or "UNKNOWN", followed by a brief explanation.
</instructions>
```

Figure 11: Prompt Version 2: XML-structured format with in-context examples.

```
## Role
You are an e-Commerce data analyst specializing in product attribute verification and quality assurance.
Your task is to assess whether a generated attribute label correctly represents the specified product attribute
based on available listing information.

## Task Overview
Evaluate the accuracy of a generated attribute label by analyzing product listing data and determining if the label
is supported, contradicted, or cannot be verified by the available information.

## Evaluation Guidelines
Follow these steps systematically:
1. **Review Available Evidence**: Examine all provided product information (title, description, bullet points)
   for any evidence relating to the specified attribute.
2. **Compare Label Against Evidence**: Determine whether the generated label:
   - Matches the product information (explicitly or implicitly)
   - Contradicts the product information
   - Cannot be verified due to insufficient data
3. **Handle Missing Information**: When description or bullet points are unavailable or incomplete,
   acknowledge the limitation in your assessment.

## Classification Categories
- **CORRECT**: The label accurately reflects the attribute based on available evidence in the product listing
- **INCORRECT**: The label contradicts or misrepresents information found in the product details
- **UNKNOWN**: Insufficient data exists to verify the label's accuracy

## Examples
### Example 1: CORRECT
- ID: B08N5WRWNW | Category: LUGGAGE
- Title: <brandname> <makename> Hardside Expandable Luggage with Spinner Wheels, Brushed Anthracite, Carry-On 20-Inch
- Description: Durable polycarbonate shell construction with scratch-resistant texture
- Bullet Points: * 100% Polycarbonate shell * Four multidirectional spinner wheels * TSA-approved combination lock
- Attribute: material_type | Label: Polycarbonate
Assessment: CORRECT - Description and bullet points explicitly confirm polycarbonate shell construction.

### Example 2: UNKNOWN
- ID: B07XYZ123A | Category: CLOTHING
- Title: Men's Cotton Blend Casual T-Shirt Blue Medium | Description: [Not provided] | Bullet Points: [Not provided]
- Attribute: sleeve_length | Label: Long Sleeve
Assessment: UNKNOWN - Title mentions t-shirt but provides no sleeve length information.

### Example 3: INCORRECT
- ID: B09ABC456D | Category: ELECTRONICS
- Title: Wireless Bluetooth Headphones with Noise Cancellation
- Description: Premium over-ear headphones featuring active noise cancellation technology
- Bullet Points: * 30-hour battery life * Bluetooth 5.0 connectivity * Foldable design for portability
- Attribute: connectivity_technology | Label: USB-C
Assessment: INCORRECT - Product clearly indicates Bluetooth connectivity, contradicting USB-C label.

## Product Information to Evaluate
- **ID**: {id} | **Category**: {category} | **Title**: {title}
- **Description**: {description}
- **Bullet Points**: {bullet_points}
- **Attribute Name**: {attribute} | **Generated Label**: {generated_label}

## Response Format
Provide your assessment as: CORRECT/INCORRECT/UNKNOWN followed by a brief explanation (1-2 sentences).
```

Figure 12: Prompt Version 3: Role-based format with systematic guidelines and markdown structure.

```
{
  "id": "X",
  "category": "FILE_FOLDER",
  "text_fields": {
    "title": "<brand>, Lot de 100 pochettes...",
    "features": [
      "Les pochettes A4 en polyéthylène...",
      "Pochettes transparentes..."
    ],
    "description": "Ces chemises en plastique..."
  },
  "attribute_of_interest": {
    "name": "tab.position",
    "value": "Côté"
  }
}
```

Figure 13: Example input format for LLM judges. Models receive structured product data with target attribute to verify.