



amazon ads

Delivering ad relevance without third-party cookies:

third-party cookies

Advanced AI-powered
contextual techniques

Table of contents

<i>Introduction</i>	03
<i>Amazon Ads approach to contextual targeting</i>	04
<i>In depth:</i>	
<i>Model architecture and use of LLMs</i>	05
<i>Understanding open web and mobile supply</i>	09
<i>Understanding demand via product-based classification and beyond</i>	11
<i>Evaluating relevance and performance</i>	14
<i>The customer impact</i>	18
<i>A look into the future</i>	20

With the deprecation of ad identifiers, advertisers are turning toward alternative solutions, including well established ones such as contextual targeting. This renewed interest means that these products are having a bit of a renaissance and innovation to include the newest advances, including predictive modeling and generative AI.

Contextual targeting has always allowed advertisers to place ads aligned with relevant content consumers are viewing in real time. Amazon Ads has redefined what contextual means and moved beyond a simple *'if keyword is present on page, then serve ad'* heuristic by leveraging Amazon's unique shopping insights and AI. Amazon DSP analyzes the context in a nuanced and scalable manner, accounting for complex relationships between words, images, and video content with the intent of being aligned with consumers' shopping journey. AWS-powered AI's ability to parse through vast amounts of unstructured data, recognize semantic themes, and understand audiences current intent, elevates contextual targeting beyond traditional keyword matching. As a result, **advertisers are now able to target categories and contexts they previously couldn't, or could do so using behavioral signals only.**

This innovation aligns with broader industry trends, as **60% of marketers now use AI in advertising**, with nearly half focusing on contextual applications. With the AI advertising market growing 35% annually, the demand for identifier-independent strategies continues to rise. Investments in AI-driven contextual targeting are yielding measurable results, including a **25% increase in consumer engagement**, highlighting the value of Amazon Ads advanced approach.

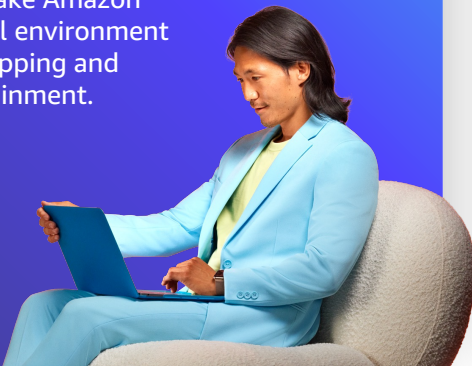
This whitepaper delves into how AI is transforming contextual targeting, and it showcases Amazon DSP innovative solutions to deliver value while minimizing the necessity of third-party cookies.

Amazon Ads *approach* *to contextual targeting*

Cutting through the clutter of messages is a challenge for every campaign. Common techniques such as loud, dynamic creative, targeting low ad-density media, and increasing frequency caps are often employed to make ads impactful. However, reaching the right audiences based on past behavior has become increasingly difficult due to the reduction in available ad identifiers.

To address these challenges, Amazon DSP has redefined what contextual means in the industry and developed a **suite of Contextual Targeting products** that leverages advertiser text and language to identify the optimal moment to serve ads while audiences are actively shopping. Our approach capitalizes on a deep understanding of shopper behaviors, products, and categories to deliver the right message at the right moment—without relying on third-party cookies. Instead, we utilize **advanced machine learning and large language models (LLMs)** to analyze the context of ad placements, ensuring relevance to shopper interests and alignment with advertisers' strategies. By moving beyond traditional keyword targeting, Amazon DSP ensures ads are placed efficiently and effectively where they'll have the greatest impact.

- For ad placements on Amazon-owned channels like the **Amazon Store**, **Twitch** or **Prime Video**, we leverage our robust product categorization and AI-powered recommendation systems that make Amazon an ideal environment for shopping and entertainment.



- When it comes to ad placements across the entire internet including third-party publishers, our challenge is understanding diverse contexts spanning a wider array of topics while maintaining the same high bar for relevance. **Generative AI models**, accessible via **AWS Bedrock**, and embedding-based approaches are uniquely suited to tackle this challenge at any scale. These advanced models empower advertisers to harness the transformative potential of modern AI to deliver contextually relevant ads effectively across the entire open web.

Read on to understand contextual targeting on third-party supply within Amazon Ads, covering key elements such as **modeling architecture and LLM selection**, **datasets for understanding supply**, **product targeting and taxonomy choices for contextual category targeting**, and **relevance evaluation based on Amazon's principles using human judgments and LLMs**. We will delve into the complexities faced by the Amazon Ads team and the innovative solutions being developed to overcome these challenges, providing a comprehensive view of Amazon Ads unique approach to contextual targeting.

In-depth: model architecture and use of LLMs

Large Language Models (LLMs) have transformed productivity by enabling machines to interpret and reason about language, supporting complex tasks like text summarization, classification, and comprehension. Since the 2017 paper "[Attention is All You Need](#)" introduced the **transformer architecture**, rapid advances in LLMs have driven both commercial and open-source innovation.

Evaluation methods for LLMs now focus on real-world applications, including ad placements, where understanding content context is essential for relevance. These models identify patterns and nuances in language, allowing for accurate ad targeting based on contextual understanding, which directly enhances ad placement accuracy and effectiveness.



In-depth: model architecture and use of LLMs

Here are some **foundational concepts** to understand how Amazon Ads leverages LLMs for contextual targeting:

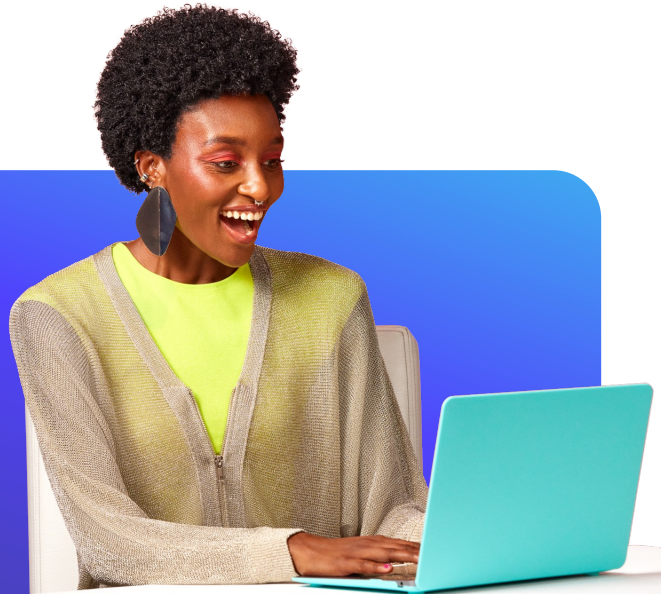


Transformers are the architecture that makes LLMs possible, and a type of neural network that use an "attention mechanism" to **understand the relationships between words in a sentence**. For example, a transformer can distinguish between different meanings of the word "bank," whether it refers to a riverbank or a financial institution, by considering the surrounding words. It represents these relationships numerically, which enables the model to handle more abstract tasks like summarizing the key points of a document. Finally, transformers can be split into configurations such as **encoder-only, and encoder-decoder**.

Generative LLMs are AI models that use an **encoder-decoder transformer** architecture: the model takes an input (like a question or prompt), encodes it into numbers (representations), and then generates a response based on that encoding. These models are great for interactive tasks like chatbots but they can be large and high latency, which makes them impractical for tasks that need to be completed quickly or cost-effectively. That's where **simpler transformer "encoder" architectures** can be helpful, as they can pre-compute, cache information and use that to match supply and demand quickly. Encoders convert large sets of text into numerical representations called **"embeddings"**, which are like summaries of the text. These embeddings can then be searched and matched to relevant queries efficiently using algorithms like **Approximate Nearest Neighbors (ANN)**. An ANN is a technique used to find items that are most similar (or "nearest") to a given query, within a large dataset, in a fast and efficient manner.

In-depth: model architecture and use of LLMs

In programmatic advertising, where ads are matched with content in real-time, we process billions of requests every minute with strict time limits. **At Amazon Ads, we use both encoder-decoder and encoder LLMs to understand content and link it to advertisers' desired targeting settings via a two-step approach:**



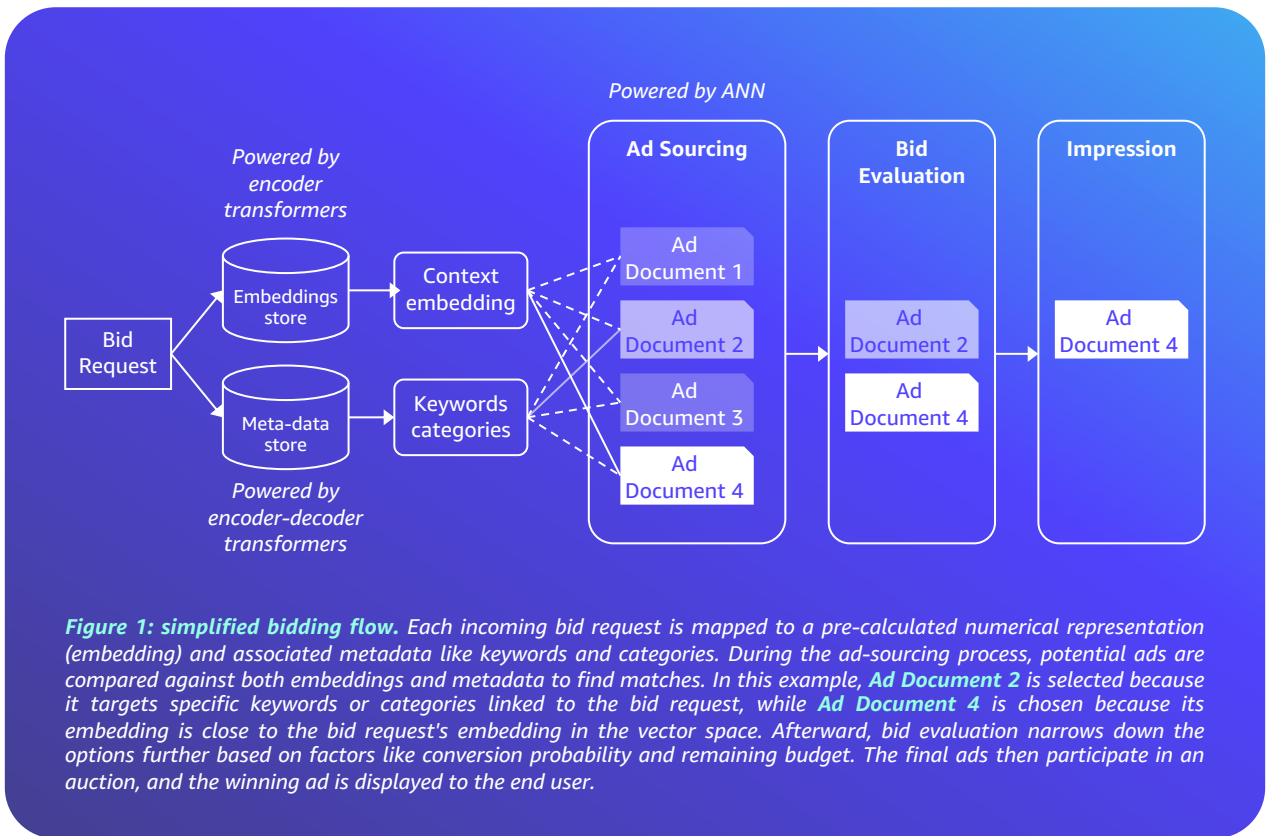
First, we run **offline batch processes that categorize content based on keywords** using encoder-decoder LLMs and match them with ads in real time using indexes.

Second, we use **embeddings** (numerical representations of content) to encode context and quickly find relevant ads in real time using algorithms like **Approximate Nearest Neighbors (ANN)**.

LLMs LLMs
LLMs LLMs
LLMs LLMs
LLMs LLMs
LLMs LLMs

In-depth: model architecture and use of LLMs

The first method ensures that we can achieve high precision and coverage amongst a set of pre-defined categories, the second enhances our recall and addressability, especially in cases where relevance needs to be evaluated for topics that are not part of a pre-defined taxonomy.



In summary, large language models (LLMs) and transformer architectures have significantly improved language interpretation and reasoning, making them valuable for tasks like summarization, classification, and comprehension. At Amazon Ads, we leverage these technologies by combining language models like encoders with fast algorithms such as Approximate Nearest Neighbors (ANN) to

process billions of bid requests efficiently, allowing us to retrieve relevant ads in real time. This system ensures both accurate ad placement and the flexibility to handle more abstract or evolving topics, maximizing relevance while maintaining the strict latency demands of real-time ad auctions.

In-depth: Understanding open web and mobile supply

Beyond Amazon Ads rich first-party inventory, Amazon DSP also unlocks valuable bid opportunities across the open web and beyond. While some publishers provide direct contextual signals, scaling this effort requires understanding and classifying content across a wide range of sources. To achieve this, we extract meaningful contextual signals from open web and mobile app inventory, while respecting publisher preferences on how their content is used.

- For both desktop and mobile web, **Amazon DSP processes over 8.5TB of unstructured webpage data daily, extracting valuable text and metadata.** Importantly, this process operates transparently, identifying it-self as the [Amazon AdBot](#) and respecting all instructions in a site's robots.txt file, including frequency and other access restrictions. This allows publishers to maintain control over how and when their content is processed.

- We focus on webpages where Amazon DSP can serve ads and use **advanced models to avoid Made For Advertising (MFA) pages**, which historically waste 15% of open web ad spend, according to the Association of National Advertisers. By crawling frequently changing pages in near real-time and refreshing more static pages every 48 hours, we ensure our signals are both relevant and up-to-date.



In-depth: Understanding open web and mobile supply

Using these signals, our LLMs assess whether a webpage or app is contextually aligned with an advertiser's targeting requirements, enabling precise ad placement. The entire process is illustrated in Fig. 2.

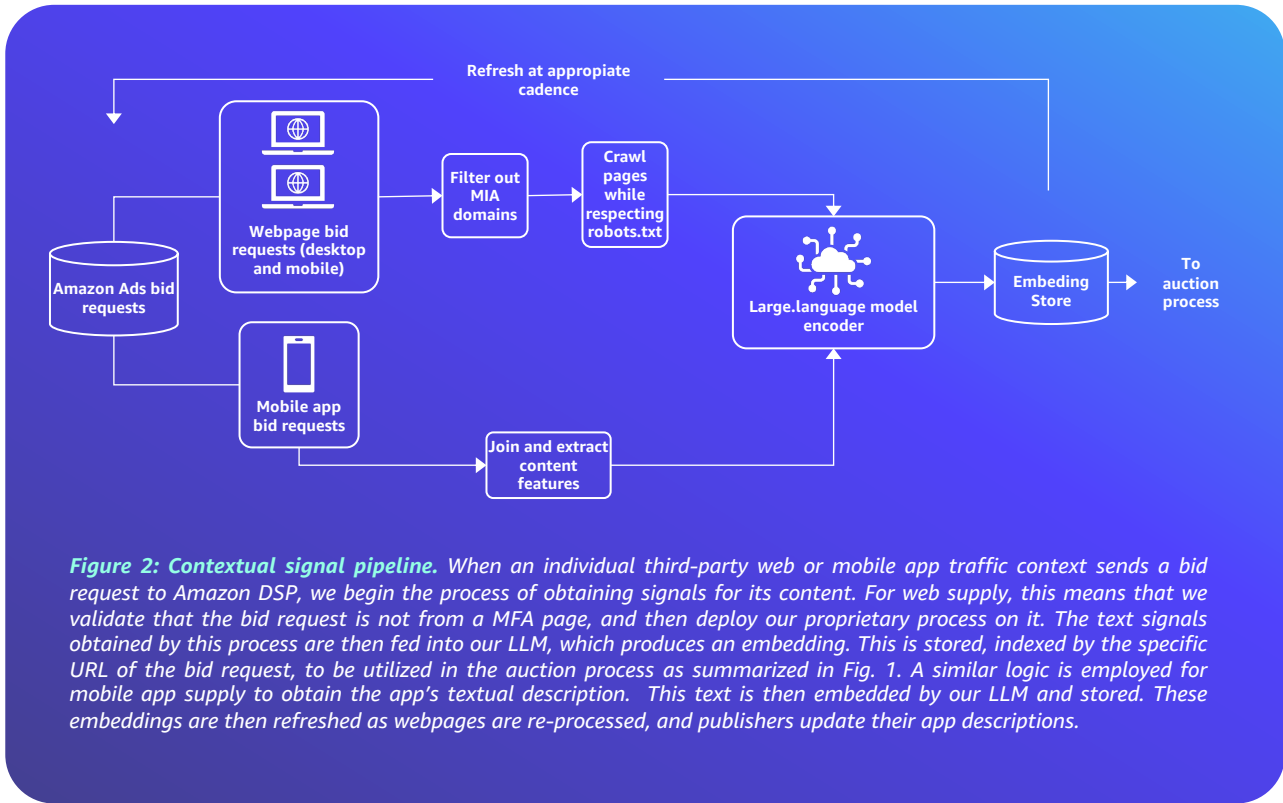


Figure 2: Contextual signal pipeline. When an individual third-party web or mobile app traffic context sends a bid request to Amazon DSP, we begin the process of obtaining signals for its content. For web supply, this means that we validate that the bid request is not from a MFA page, and then deploy our proprietary process on it. The text signals obtained by this process are then fed into our LLM, which produces an embedding. This is stored, indexed by the specific URL of the bid request, to be utilized in the auction process as summarized in Fig. 1. A similar logic is employed for mobile app supply to obtain the app's textual description. This text is then embedded by our LLM and stored. These embeddings are then refreshed as webpages are re-processed, and publishers update their app descriptions.

In conclusion

Amazon DSP extends its reach beyond first-party inventory by extracting contextual signals from the open web and mobile apps. It processes 8.5TB of data daily, including app insights, ensuring precise ad placements. Our approach respects publisher controls, avoids inefficient placements, and supports multilingual capabilities to meet advertiser targeting requirements.

In-depth: Understanding demand via product-based classification and beyond

Advertisers have multiple options to articulate their contextual targeting strategies within Amazon DSP, ranging from pre-built taxonomies leveraging Amazon’s retail expertise to more direct keyword-based targeting. These approaches allow advertisers to effectively address objectives at various stages of the sales funnel across Amazon owned & operated and third-party sites.

Related Product Targeting

We leverage our shopping insights from hundreds of millions of Amazon customers to find **complementary and substitutable products**. Related products includes well-known concepts such as “Frequently Bought Together” items, “Customers Who Bought This Also Bought”, “Customers who Viewed This Also Viewed” and “What do customers buy after viewing this item?”.

Contextual Targeting – Retail Category

This solution provides access to the hierarchical **Amazon browse tree** product category taxonomy. This taxonomy, constructed with Amazon’s extensive retail expertise, categorizes products with increasing granularity as one moves deeper within the hierarchy. A simplified and truncated example of one of the sub-trees is shown in Fig. 3. Our contextual model respects the hierarchical structure of this taxonomy, offering advertisers a strategic balance between specificity and reach. Relevance is treated as transitive to parent categories, allowing broader categories to provide

greater reach while more specific lower-level categories help advertisers focus on targeted contexts. On the open web, our LLMs encode this information to generate ad-line embeddings, enabling ad sourcing beyond just the limited volume of pages where explicit categories are provided.



In-depth: Understanding demand via product-based classification and beyond

Contextual Targeting - Retail Product

This solution allows advertisers to target individual products directly from Amazon's retail catalog. Similar to Retail Categories, on the open web, this product information is embedded using our LLMs for **vector-based Approximate Nearest Neighbor (ANN)** searches, while targeting remains deterministic for Amazon's first-party inventory.

With Amazon's product catalog comprising hundreds of millions of items, indexing and searching through embeddings is challenging. To make this scalable, we employ **scalar quantization** to reduce embedding size while maintaining retrieval accuracy. This approach empowers advertisers to target customers who already show strong interest in specific products - enabling targeted promotions for their own products or competitor products, emphasizing their unique differentiators.

To enhance this targeting mode, we also offer an **optional related product targeting feature** that uses Amazon's **retail co-view, co-purchase, and search data** to identify and target products similar to those being targeted, thereby expanding the range of relevant contexts.

Keyword Targeting

This solution helps advertisers seeking greater flexibility or aiming to target highly specific contexts not covered by our taxonomies. Advertisers can input free-form keywords and choose between exact matches, broad matches, or both.

- In **exact matching**, ads are served only on open web pages containing the keyword in their textual content or when provided by the publisher, as well as on Amazon's first-party pages (e.g., product detail pages or search pages explicitly associated with the keyword).
- **Broad matching** is powered by our LLMs, which leverage embeddings and **ANN algorithms** to identify related webpage embeddings, similar to our product-based taxonomy offerings. This combination provides advertisers with the power to extend their reach intelligently while ensuring that contextual relevance is maintained.



In-depth: Understanding demand via product-based classification and beyond

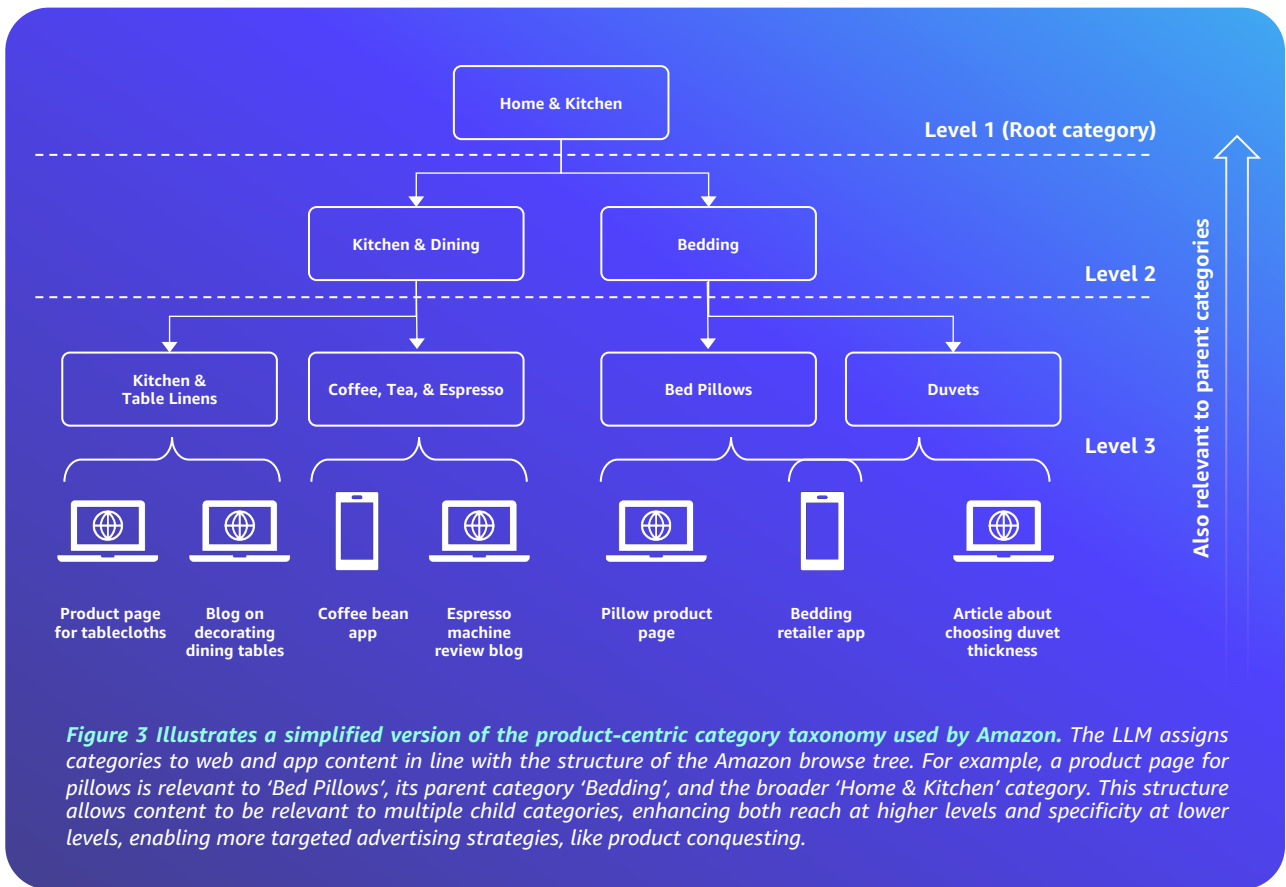


Figure 3 illustrates a simplified version of the product-centric category taxonomy used by Amazon. The LLM assigns categories to web and app content in line with the structure of the Amazon browse tree. For example, a product page for pillows is relevant to 'Bed Pillows', its parent category 'Bedding', and the broader 'Home & Kitchen' category. This structure allows content to be relevant to multiple child categories, enhancing both reach at higher levels and specificity at lower levels, enabling more targeted advertising strategies, like product conquering.

In conclusion

Amazon Ads offers a range of sophisticated contextual targeting solutions that leverage our understanding of shopper behavior and advanced AI models. By utilizing hierarchical product taxonomies, retail product targeting, and flexible keyword options, we provide advertisers with powerful tools to connect with their audiences at the right moment. Our approach maintains high relevance, enabling advertisers to effectively engage customers across the sales funnel and drive meaningful results.

In-depth: Evaluating relevance and performance

Contextual targeting affects the end-to-end ads delivery system, requiring a balance between accurate context matching and maximizing ad reach. We rigorously assess our machine learning models to ensure they both meet advertisers' targeting criteria and unlock new ad opportunities at scale. Here's an overview of our evaluation process:

Defining informative metrics

To evaluate the effectiveness of our models, we use a comprehensive set of metrics covering both **offline** and **online** evaluations. The offline evaluation serves as an initial step to benchmark the model's predictions directly against ground truth training data for relevance, while the online evaluation tests how the models perform in a live ad auction environment. During the offline stage, we focus on the following key metrics:

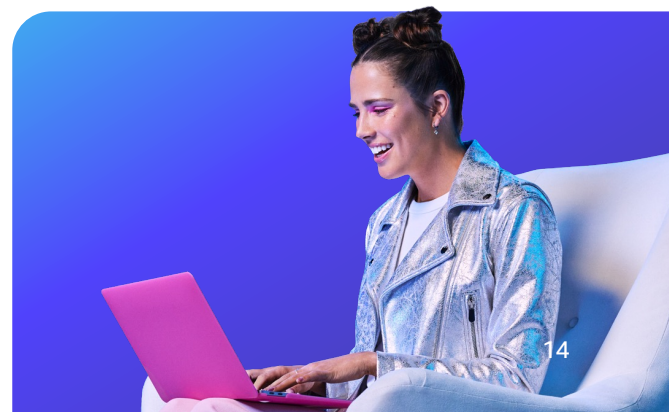
1. Precision@k:

One of our primary goals is to ensure that the models adhere to advertiser instructions and accurately match relevant contexts. To assess this, we evaluate how well our models associate targeted labels (such as retail categories, products, and keywords) with individual webpages or apps. We extract a large sample of representative contexts, and for each context, we analyze the fraction of the top-k associated labels that are semantically relevant to the content. This fraction is known as Precision@k, and we measure it for progressively larger values of k. For example, a **Precision@25** score of 0.8 for retail categories on a webpage means that 80% of the top 25 categories identified by the model are semantically relevant to the page's content. Similarly, a **Precision@10** score of 0.9 indicates that 90% of the top 10 categories are relevant.

This metric helps us gauge how well our models align ad placements with the targeted content. For a deeper understanding of how we measure semantic relevance and evaluate it at scale, refer to *Validating our accuracy*.

2. Ad-line to bid request match rates:

To ensure that our models generate sufficient ad supply for advertisers, we also monitor how the predicted labels (categories, products, and keywords) impact the number of eligible placements for each ad-line. Specifically, we measure the proportion of bid requests within a given day that match a particular ad-line based on the targeted labels and the similarity of their embeddings to the bid request's embedding (whether it's a webpage or app). This metric allows us to verify that our models are unlocking enough relevant ad opportunities, ensuring broad reach for advertisers.



In-depth: Evaluating relevance and performance

Live Campaign Metrics: Once deployed, we monitor the following key performance indicators (KPIs) for advertisers:

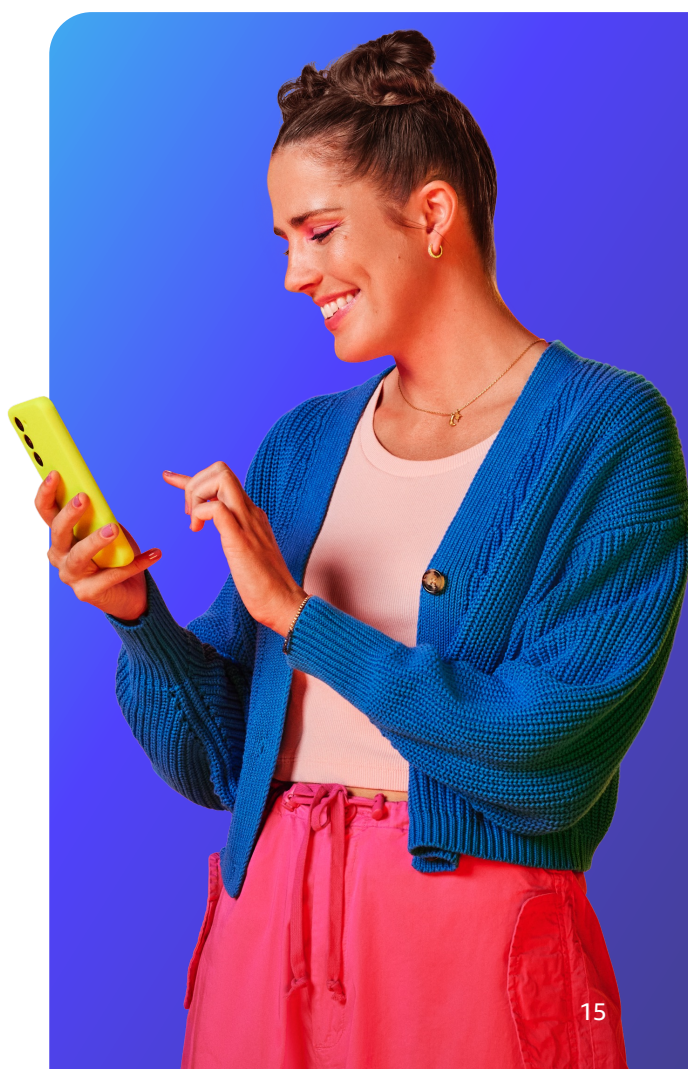
1. Performance KPIs:

These include industry-standard metrics like cost per action (CPA), cost per thousand impressions (CPM), and return on ad spend (ROAS). Additionally, we track engagement and conversion rates such as click-through rate (CTR) and detail-page-view rate (DPVR), ensuring our models deliver meaningful engagement and value.

2. Delivery KPIs:

We also measure delivery-related metrics, including the number of impressions and total spend, ensuring that campaign budgets are fully utilized and maximizing ad delivery. For a deeper understanding of how we measure semantic relevance and evaluate it at scale, refer to *The Customer Impact*.

This comprehensive approach ensures our models are both accurate and effective, delivering on advertiser goals while maximizing reach.



Validating our accuracy

To scale the content annotation process, we use a hybrid approach that combines expert human annotators with an advanced automated system powered by conversational agents. We employ an innovative technique called **tenet-based chain-of-thought (CoT) prompting** to guide AI's reasoning - a novel in-context learning prompt-engineering technique. This approach enhances the model's ability to make accurate decisions by breaking down complex reasoning tasks into smaller steps and aligning them with **Amazon's core tenets mechanisms** - a core

principle of how Amazon operates, codifying organizational beliefs to accelerates decision-making - to systematize LLM decision-making processes by providing a rule-based CoT reasoning framework.

This framework enables the model to understand content and context more effectively, producing accurate annotations at scale. A simplified, high-level example of this is shown in **Fig. 4**.

Tenet-based CoT instruction prompt template

<context>

General context relevant to the task (e.g. Webpage content to be annotated).

</context>

<variables>

Tenet definition and description of variables (e.g. Different inputs in context) required to complete the task. Example:

Tenets:

(1) Tenet 1: A rule or principle to be applied.

(2) Tenet 2: Another rule or principle.

</variables>

<instructions>

General task description, explaining what needs to be evaluated and how to use the tenets. Example Task:

Determine whether the webpage content meets the criteria set by the tenets. For each tenet, provide a rationale for your decision.

</instructions>

<format>

Comments: Discuss your rationale on solving the task by reasoning based on the tenets.

Evaluation: Final output.

</format>

Figure 4: High-level illustration of a Tenet-based CoT prompt. When using LLMs to perform large scale annotation, we adapt the following template to guide its decision making using tenets. In practice, the contextual features of the bid request, and predicted retail categories are supplied within the 'context' section. Specific tenets are provided in the 'variables' section, and detailed instructions in relation to how these are to be applied are stated in 'instructions'. This approach augments traditional in-context learning techniques, such as providing demonstrations or personas, and the additional tenet-based CoT components of the prompt are highlighted in purple.

Validating our *accuracy*

By leveraging this automated pipeline, we can process and validate model predictions across large-scale data sets, supporting all languages on Amazon DSP, at a rate that is **faster** and **more cost-effective** than relying solely on human annotators. The annotation accuracy achieved by the automated system is well within the variance of human expertise, and it is **25% more accurate** than crowd-sourced annotations. Additionally, human annotators continue to play

a key role, receiving extensive training to handle complex edge cases and provide oversight in situations where ambiguity arises.

This hybrid system allows us to achieve both the **scalability of automated annotations** and the **precision and flexibility of human evaluation**, ensuring that our contextual targeting remains both accurate and efficient.



The customer *impact*

Now that we've explored the complex models and embeddings of our Contextual Targeting solution, it's time to evaluate its real-world impact on advertising campaigns. While the technical advancements are many, what truly matters are the tangible benefits for advertisers in practical scenarios.

For starters, we find that contextual targeting delivers a higher proportion of ads on quality, content-rich webpages compared to targeting based on ad ids. In fact, the proportion of **contextually targeted impressions** served on meaningful pages (e.g., in-depth articles) is **1.3x higher** than those served through ad ids that reflect historical behaviors. And when we compare the conversion rates, we see that **bid requests without ad ids are twice as likely to convert** on content-rich pages compared to less meaningful ones.

This makes intuitive sense: for audiences based on traditional ad ids, ads are tailored based on past interactions, however, for **targeting strategies that do not depend on ad ids**, the context of the page becomes a critical factor in driving conversions, as the **content itself provides the most relevant signal for ad placement**.

In a real-world A/B test against traditional contextual solutions, our AI-powered contextual targeting consistently out-performs in key areas:

24% decrease in engagement costs.

23% reduction in CPM for brand campaigns.

15% reduction in CPM for performance campaigns.

10% increase in return on investment (ROI).



The customer *impact*

Beyond these broad metrics, here are specific advertiser success stories:



PEPSICO

Maximizing Prime Day Success with Contextual Keyword Strategy

Goal:

PepsiCo sought to engage value-driven consumers on third-party sites during Prime Day 2024 to drive purchases on Amazon.com.

Results:

Achieved a **3x higher ROAS** compared to behavioral-based audiences, achieved **62% lower CPA**, and expanded unique reach with a 60% reduction in CPM.



Cost Savings and Expanded Reach with Contextual

Goal:


Smiles wanted to increase consideration for their new credit card that maximizes mileage earning by engaging suitable audiences across various formats and devices.

Results:

Implementing a contextual strategy led to a **47%** reduction in CPC on desktop and a **27%** reduction on mobile, effectively driving consideration for the credit card offering.


Direct feedback from advertisers who have successfully leveraged our contextual targeting solution include:



This innovative product has created new opportunities to engage previously non-addressable audiences precisely when their interests align with real-time content consumption. We've observed a significant uplift in our core KPIs, demonstrating that achieving performance without ad IDs is a tangible reality today. We have integrated it into our core targeting strategy, and see it as a robust, long-term approach that warrants continued investment. 

Groupe SEB, Spain



The ability to efficiently and effectively reach incremental, qualified shoppers at the moment they consume content that complements our brand or product is invaluable. This targeting capability unlocks a new method for brands to generate incremental demand and conversions without relying on ad IDs. 

Flywheel Digital, US

A lookout into the future



Contextual targeting will **remain a key strategy for advertisers** to connect with audiences in relevant ways, and at Amazon Ads, it will continue to be a priority investment area.

As we look forward into 2025 and beyond, we plan to build upon and extend the comprehensive set of contextual targeting mechanisms described in this whitepaper by **building models that leverage LLMs to understand targeting holistically at an ad line level**. This will directly connect contexts to the shopper personas advertisers describe with complex targeting

criteria, which potentially include **both contextual and behavioral strategies**. We're also exploring the potential to incorporate **AI-powered context understanding into our closed measurement loop**. This could improve campaign performance by enabling smarter recommendations and introducing a richer set of signals for our machine learning models. Together, we expect these enhancements to improve outcomes for advertisers of all sizes, across different verticals, on both Amazon inventory and across the rest of the internet.

Authors

Daniele Barchiesi – Applied Science Manager @ Amazon Ads

Anurag Deshpande – Machine Learning Scientist @ Amazon Ads

Guilherme Illunga – Applied Scientist @ Amazon Ads

Elias Kassapis – Applied Scientist @ Amazon Ads

Aditya Singh - Applied Scientist @ Amazon Ads

Ona Prat – Sr Product Marketing Manager – Ad Tech @ Amazon Ads

Steve Pinto – WW GTM Lead – Ad Tech @ Amazon Ads

About Amazon DSP Ad Relevance

Ad Relevance is the innovative approach Amazon Ads uses to understand relevant ad opportunities for all products and services advertised through the Amazon DSP. Ad Relevance is built on Amazon Ads extensive understanding of what creates great shopping experiences and their connection with ad interactions, interests, and cadence of actions along the path to purchase. It uses the latest in AI and machine learning technology to analyze billions of browsing, buying, and streaming signals in conjunction with real-time information about the content being viewed to understand where customers are in their shopping journeys, and serves them relevant ads across devices, channels, and content types without needing third-party cookies.