# Product Data and Identifiers for AI Shopping Agents: A Measurement of 1.17 Million Shopify Storefronts

Datify and Charles Swayne
October 2026, updated November 2026

## Summary for store owners

AI shopping tools can already find your store. What they learn about each product comes from facts such as price, brand, size, color and barcode, printed in a hidden block on your product page. We checked 1,168,182 Shopify stores for 31 of these product facts.

- **The average store prints 9 of the 31 facts.** Its grade is 28.6 out of 100 (the median store scores 32). The grade is the share of the 31 facts a store prints, so a grade of 32 means about 10 of 31.
- **14% of stores print none of the facts in the format most agents read** (JSON-LD). About 9% of stores use an older format, microdata, instead, so about 5% print no product markup in either format.
- **Under 1% of stores tell agents the color, size or material of a product in a form agents can filter on,** even when shoppers can pick these on the page.
- **Shopify has a category for about two thirds of the products we checked, but for only about 1 in 4 of the oldest ones.** The category is the label (for example, T-Shirts) under which Shopify lists the attributes a product should carry. About two thirds of sampled products have one, against about 4 in 5 of the newest products and about 4 in 10 of the products in stores with more than 5,000 products. Shopify lists about 9 attributes for the average categorized product, and nearly all of the stores that have a categorized product and print Product JSON-LD print no attribute (color, size, material, pattern or audience) and no extra named attribute on any page we checked (Section 6.6).
- **The basics are nearly universal and the details are rare.** Name, brand, price, image and availability appear on about 85% of stores. Category appears on 41.9%, barcode on 16.0%, and size, color and material each on under 1%. Unless a sentence says otherwise, every percentage that describes a fact is the share of stores (not of products) that print the fact on at least one product page we checked; the category results (Section 6.6) say where they are shares of products.
- **Your theme sets a floor, not a ceiling.** The best themes print about 12 of 31 facts, most print 10 or 11, and none prints the details. Filling in the missing facts in your product data and printing them on the page is the work that is left.
- **We did not test whether filling these in changes what AI agents recommend.** This is a measurement of what stores show agents, not of what agents do with it (Section 7.4).

**Terms used in this paper.**

- *Structured data (JSON-LD)*: a block of code on a product page that lists the product's facts in a form AI agents read easily. Shoppers never see it. *Microdata* is an older format for the same facts that agents may also read.
- *GTIN*: the barcode number of a product (a UPC or EAN). Shopping tools use it to tell that two listings are the same product. *SKU* is the store's own stock code. *MPN* is the part number the maker gave the product.
- *Variants and options*: the choices a shopper makes on a product, such as size and color. *Metafields* are extra fields a store adds to a product in Shopify, such as material.
- *Shopify category and category attributes*: Shopify's standard product taxonomy puts a product in a category, for example Apparel and Accessories > Clothing > Clothing Tops > T-Shirts. Each category lists attributes that a product in it should carry (a T-shirt has size, color, material, neckline and sleeve length). Category attributes are also called category metafields.
- *Catalog* has two meanings. A store's catalog is its product list in Shopify. Shopify's global catalog (also called the shared catalog) is a separate list of products from many stores that AI agents can search.
- *Grade*: the share of the 31 facts a store prints in its product page JSON-LD, from 0 to 100.
- *Agent*: an AI shopping tool that reads product pages and platform files to answer a shopper's question.

# Abstract

AI shopping agents can read product facts from product pages, from platform endpoints built for agents, and from platform catalogs. Google documents that it reads schema.org Product markup for shopping results [14], and Shopify offers agents a catalog that spans its stores [13]. We measure how complete this data is across the public Shopify web, using a grade of 31 product facts. From the 133.2 million registered domains in Common Crawl's July to September 2026 web graph we found 1,966,842 addresses that point at Shopify, tried 1,966,840 of them in October 2026, and graded 1,168,182 distinct storefronts on up to five product pages each. Follow-up samples read store product files, stores' own UCP endpoints and Shopify's global catalog. A pass in November 2026 used a large sample of graded stores and sampled products, including their variants.

Agent files and crawler access are nearly uniform: 99.9% of stores serve an agents.md file (its content varies by store) and 99.9% block none of 12 AI crawlers. Product data varies far more. The average store prints 8.9 of the 31 facts (a mean grade of 28.6 out of 100, median 32), 14.1% of stores print no Product JSON-LD (5.0% print no Product markup in any format) and 84.0% print no GTIN on any sampled page. GTINs are scarce in the merchant's own data: 20.4% of sampled stores have a valid barcode on at least one of their first three products, and among products whose vendor is not the store itself, 17.6% have one. In the samples where we checked, neither the store's UCP endpoint nor the global catalog

returned a barcode for any product. About two thirds of the sampled products with a UCP answer carry a Shopify standard category, from about 1 in 4 of the oldest products to about 4 in 5 of the newest and about 4 in 10 of the products in stores with more than 5,000 products, and nearly all of the stores with a categorized product and Product JSON-LD print no category attribute on any page we read, although Shopify defines about 9 per product. About 4 in 10 of the sampled products were found in the global catalog, and about 3 in 4 for US products in stock. Theme explains 28% of the variance in the 31-fact grade, against 1% for country and under 1% for popularity rank, but much of that is markup format: stores on the oldest free themes print microdata instead of JSON-LD. We do not measure whether these gaps change what agents recommend; a small observational pilot found no detectable difference (Section 7.4).

# 1 Introduction

An AI assistant that answers “find me running shoes under $100” has to know which products exist, what they are, what they cost, whether they are in stock, and which listings are the same item sold by different stores. For the last fact the standard answer is the GTIN, the barcode number a brand owner registers with GS1 [12]. Agents can get product facts from three places: the HTML of the product page (often its schema.org Product JSON-LD), platform endpoints built for agents, and platform-wide catalogs.

Shopify hosts a large share of independent online stores and has recently added agent-facing layers to every storefront: an agents.md file, a Universal Commerce Protocol (UCP) profile and endpoint, and a global catalog that agents can search across stores. Vendors, including Datify, which published this study, sell “AI readiness” services to merchants. There is little public measurement of where stores stand. Industry reports on AI shopping measure demand, for example AI referral traffic, from a handful of stores [9]. Academic web measurement has tracked structured data adoption across the whole web [1, 2, 3], but not the product data of one commerce platform across its layers.

This paper asks how complete, on 31 product facts, the machine-readable product data of Shopify stores is, where product identifiers are present, how much of Shopify’s own product category and shared catalog data exists, and what store characteristics are associated with the gaps. It is a measurement of supply (what an agent can read), not of demand or of agent behaviour.

**Contributions.**

1. A crawl of 1,168,182 distinct Shopify storefronts found through the Common Crawl web graph (Section 5).
2. A 31-fact completeness grade for product page JSON-LD, its distribution across stores, and a decomposition of its variance by theme, popularity rank and country (RQ2, RQ4).
3. A measurement of GTIN presence at four points where the identifier can appear (the merchant’s barcode field, the product page, the store’s UCP endpoint and Shopify’s global catalog), anchored on a product-level sample that links the same products across three of them (RQ3).

4. Coverage of Shopify's global catalog on a large sample of products, and a comparison of store products with their catalog entries: fidelity, and what the catalog adds or drops (RQ5).
5. A measurement of Shopify product category coverage and category attribute use on a large sample of products and stores, and how they vary with product age and store size (RQ6).
6. Aggregate tables and the crawler code, so the study can be repeated (Section 10).

## 2 Background: where an agent can read a Shopify product

The merchant's product record in Shopify admin (L1) is the source. Shopify and the store's theme publish it through four other layers. Each is generated from L1, not from the layer before it, so a fact missing in L1 can appear elsewhere only if that layer adds it (for example, a theme that prints a GTIN stored in a metafield rather than in the Barcode field).

| Layer | Who controls it | What we read | Product identifier |
|---|---|---|---|
| L1 Product record in Shopify admin | Merchant | Not readable from outside; inferred from L2 | Variant Barcode field, free text, not validated by Shopify |
| L2 Public product files | Platform, from L1 | /products.json (no barcodes); /products/<handle>.js (barcode, SKU, alt text) | Barcode as entered, in the .js file only |
| L3 Product page HTML (Product JSON-LD or microdata) | Theme template, apps | Up to five product pages per store | gtin, gtin8/12/13/14, if the template prints it |
| L4 Store agent files and UCP endpoint (/agents.md, /llms.txt, /.well-known/ucp, /api/ucp/mcp) | Platform | Public, read only requests: files on every store, and one product per request (up to five products per store) | Not returned (Section 6.3); the product's Shopify category is returned (Section 6.6) |
| L5 Shopify global catalog (catalog.shopify.com) | Platform | Public, read only lookup requests for the variants of the same products (up to five products per store) | Not returned (Section 6.5) |

Schema.org Product markup is the common vocabulary for L3. Search engines have promoted it for rich results for a decade, and it is readable without an API. A GTIN is valid when it has 8, 12, 13 or 14 digits and a correct GS1 check digit [12]; we use that definition throughout. Many products legitimately have no GTIN: handmade, made-to-order and private-label goods, digital products and gift cards. We therefore report GTIN presence, not loss, and Section 6.3 gives presence separately for products sold under a vendor other than the store. The global catalog is documented by Shopify [13]. Shopify's standard product taxonomy gives each product a category and each category a list of attributes. The category is stored on the product record (L1) and returned by the store's UCP endpoint (L4); the attribute values are metafields on the record, which we cannot read from outside.

## 3 Related work

**Structured data on the web.** The Web Data Commons project extracts schema.org data from Common Crawl and has tracked its adoption for over a decade. Meusel, Bizer and Paulheim [1] used successive

crawls to study how the schema.org vocabulary is adopted and how deployed data departs from the standard; Meusel and Paulheim [18] catalogued common errors in deployed schema.org data and heuristics to repair them, which is the setting for our malformed JSON-LD counts. The 2024 corpus [2] reports sites with Product annotations rising from about 581,000 in 2017 to 3.3 million, and the share of them using Product/sku rising from 21% to 60% over five years. The HTTP Archive Web Almanac [3] reports structured data formats and types from the HTTP Archive crawl; its 2024 edition finds Product in JSON-LD on 0.77% of all crawled pages (not only product pages). These studies count which properties appear across the whole web; they do not follow a product from the merchant's data to the page and to a platform's own agent layers, and they do not treat one commerce platform's stores as a population.

**Crawling shopping sites.** Mathur et al. [15] crawled about 11,000 shopping websites to measure dark patterns, the closest precedent for a large crawl of e-commerce storefronts. Our crawl is larger and limited to one platform, which lets us identify each store's theme and read platform endpoints.

**Product identifiers and matching.** Product offers marked up with schema.org are the raw material for product matching research. Primpeli, Peeters and Bizer [19] built a large training set for product matching by clustering offers that share an identifier such as a GTIN or MPN, and WDC Products [4] builds a benchmark from such offers and shows that matchers still struggle with products they have not seen. Identifiers are what make matching across stores reliable, which is why their presence matters for agents that compare sellers.

**AI crawlers and consent.** Longpre et al. [5] audited robots.txt and terms of service for 14,000 web domains used in AI training corpora and found AI-specific restrictions rising sharply in 2023 to 2024. Liu et al. [6] studied how content creators use robots.txt and other tools against AI crawlers. Our robots.txt results (RQ1) describe the opposite setting: commerce sites that, by default, want to be read.

**AI agents as shoppers.** Yao et al. [20] introduced WebShop, a simulated store for training and evaluating language agents that search and buy products. Allouah et al. [7] ran randomized experiments on AI shopping agents and found position bias, concentration on a few products, and large shifts between model versions. Aggarwal et al. [8] showed that content changes can raise a source's visibility in generative engine answers. These study how agents choose; none measures the product data available to them. Ethercycle's 2026 report [9] measures AI referral traffic at ten US stores; it is the demand side of the question we take from the supply side.

**Top lists.** We use the Tranco list [11] to split stores by popularity. Scheitle et al. [16] and Ruth et al. [17] show that top lists are unstable and only loosely track traffic, so we call this a popularity rank and treat its bands as coarse.

**Shopify.** We know of no prior peer-reviewed study of Shopify storefronts at this scale. Public numbers come from vendor blogs and samples of a few thousand stores, usually about speed.

# 4 Research questions

| | Question | Layer | Data |
|---|---|---|---|
| RQ1 | Access: do stores let AI crawlers in and serve the platform's agent files? | L4, robots.txt | Main crawl, 1,168,182 stores |
| RQ2 | Completeness: how many of 31 product facts does the product page's JSON-LD carry? | L3 | Main crawl |
| RQ3 | Identifiers: at which layers between the merchant's barcode field and the global catalog is a GTIN present? | L1 to L5 | Catalog sample (product level), barcode sample, main crawl, UCP pass (category row) |
| RQ4 | Correlates: which store characteristics (theme, popularity rank, country, industry) are associated with completeness? | L3 | Main crawl, industry pass, UCP pass (category) |
| RQ5 | Catalog: how many store products appear in Shopify's global catalog, where, and how faithfully? | L2, L3, L5 | UCP pass (coverage); catalog sample (fidelity) |
| RQ6 | Category: how many products carry a Shopify standard category, which products and stores lack one, and do stores print the attributes Shopify defines for the category? | L1 (read through L4), L3 | UCP pass, a large sample of stores and products, joined to the main crawl's product pages |

All six questions are descriptive. None tests whether completeness changes what an agent recommends; Section 7.4 returns to that.

# 5 Data and methods

## 5.1 Population and sampling frame

The population of interest is Shopify Online Store storefronts reachable on the public web. No public list of them exists, so we built a frame. We took the 133.2 million registered domains in Common Crawl's July to September 2026 web graph [10] and resolved each through public DNS, keeping those that point at Shopify (23.227.38.x or shops.myshopify.com). We added the stores from an earlier 40,746-store run, which came from the Tranco top 1 million list [11] and from myshopify.com addresses in Common Crawl, and 34,511 stores found through app reviews (Section 5.5). The frame held 1,966,842 Shopify-pointing addresses, and 1,966,840 were tried. It misses stores that the web graph and our earlier lists never mention, headless stores on their own hosting, and stores behind another CDN.

## 5.2 Crawl

The main crawl ran October 16 to 21, 2026. For each store that reached a product page it made 13.8 requests on average: the home page, robots.txt, /agents.md, /llms.txt, /.well-known/ucp, /meta.json, /products.json, the sitemap index and the first product sitemap, and up to five product pages. Product pages were drawn at random (seeded by host) from the store's first product sitemap file, or from links on the home page when there was no product sitemap; for large catalogs that span several sitemap files the draw covers only the first file. Requests to one store went one at a time, at least one second apart and longer when robots.txt set a Crawl-delay. The crawler identified itself as DatifyResearch/1.0 and followed each store's robots.txt. It read HTML without running JavaScript, as the large AI crawlers observed by Vercel do [21]. In total the crawl made 20.9 million requests over six days, about 40 per second on average, almost all to Shopify's shared infrastructure; most failed addresses received a single request. Section 9 gives the full crawl rules.

## 5.3 Measures

- **Completeness grade (RQ2).** For each sampled product page we check 31 product facts in Product or ProductGroup JSON-LD. A fact is a product field that an agent can use to find, match or answer a question about a product: the schema.org Product fields that at least 20 stores print, with close spellings counted once (weight, width, height, depth and measurement are one fact, "weight or dimensions") and fields that fit only one kind of product (nutrition, ISBN, drug ingredients) combined into one "category-specific detail" fact. Reviews and star ratings, release dates, related-product links, alternate names, general page fields (keywords, page id, language, video, logo, slogan, awards, links to other profiles) and keys that are not schema.org names are not facts (Appendix F lists them). The 31 facts fall in four groups: the basics (8), finding the exact product (6), category, options and specs (11) and the offer (6); Appendix F lists them with the share of stores that print each. All facts have equal weight. A store gets a fact when it prints it on at least one of its sampled pages (up to five; 4.76 on average, and 91% of scanned stores had all five graded). Its grade is the facts printed divided by 31, times 100, rounded to a whole number, so a store with no Product JSON-LD scores 0. Nine of the facts (name, description, image, price, currency, availability, brand, SKU and GTIN) are properties Google lists as required or recommended for merchant listing results [14]; the other 22 are fields an agent can match, filter or answer with (Appendix F). The grade records whether a fact is printed, not whether it is correct. Product microdata is detected and reported but not graded, so we also report a format-neutral measure: whether a store prints Product markup in any format.
- **GTIN (RQ3).** On a page: a gtin, gtin8, gtin12, gtin13 or gtin14 property holding at least 8 digits. In Shopify (L2): a barcode that passes the GS1 definition in Section 2. The two tests differ because page values are often formatted with spaces or prefixes; Section 8 discusses the asymmetry.
- **Theme (RQ4).** The Theme Store ID a store reports (743,854 main-sample stores match a listed theme by ID), and four theme groups by the theme name left in the page code: current free

themes, Debut, the oldest free themes and all others (lists in Appendix C). Copies and heavily edited themes fall into “all others”.

- **Popularity rank (RQ4).** Tranco rank for stores that came from the Tranco list; Common Crawl host graph rank as a second ranking.
- **Country (RQ4).** The address country in the store’s public /meta.json.
- **Industry (RQ4).** The top level of the Shopify product category of one product, read from the store’s UCP endpoint (the industry pass, Section 5.4).
- **Category set, explicitly Uncategorized, no category field (RQ6).** A product has a Shopify category when the store’s UCP get_product answer carries a Shopify standard category. Two kinds of product lack one and are counted as not categorized: the answer names the category “Uncategorized”, or the answer has no category field. Whether a merchant or Shopify’s own suggestions set the category is not visible. We call a store’s sampled products *all*, *some* or *none* categorized.
- **Category depth and attributes defined (RQ6).** Depth is the number of levels in the category path. The attributes defined for a product are the attributes Shopify’s standard product taxonomy lists for its category, read from a copy of the taxonomy generated on September 20, 2026 (14,606 categories). The count is what Shopify defines, not what a merchant filled in. Attributes *printed* are those the main crawl found on the store’s pages: color, size, material, pattern, age group or gender in the product markup, or any additionalProperty.
- **Product age (RQ6).** The Shopify product id grows with time, so we use it as an order of creation. We cut the answered products into eight id ranges and label each with an approximate year. The order of the ranges is exact; the year labels are estimates from known id ranges and were not checked against creation dates in our data.
- **Store size (RQ6).** The published product count in the store’s public /meta.json, in six bands.
- **Catalog found (RQ5).** The global catalog’s lookup_catalog, called with the variant ids of the product read from the product’s public .js file, returned the product. Available means the store’s UCP answer marked the product available. Products with no readable variant list were not looked up and are left out of rates unless a sentence says otherwise.
- **Six merchant facts (RQ6).** Shopify category, product type, at least one tag, GTIN in the page markup, SKU in the page markup and a description in the page markup. They are facts a merchant controls.

## 5.4 Follow-up samples

| Sample | When | Drawn from | Size | What it reads |
|---|---|---|---|---|
| Barcode sample | Oct 2026 | Random, first 1.0 million stores scanned | A large sample, after exclusions | First three products’ public .js files (barcode), store UCP get_product for one product |

| Sample | When | Drawn from | Size | What it reads |
|---|---|---|---|---|
| Industry pass | to Oct 19 | Among the first 0.83 million stores scanned: 98.4% of Tranco-ranked stores and 8.4% of the rest (mostly .com, .co and .au domains) | A large sample, after exclusions | Store UCP get_product, one product |
| UCP pass | Nov 2 to 3 | A large subset of the graded stores with sampled products | Up to five products per store; nearly all answered. Answered products used for category results | The product's public .js file, the store's UCP get_product answer (sold-out variants included), then the global catalog's lookup_catalog for each variant (50 per call) |
| Catalog sample | Oct 22 | Random, seeded, stores whose products.json answered | A large sample, after exclusions | products.json (first three products, which are the newest), their .js files, the first product page, then lookup_catalog by the first variant id (50 ids per call) |

The UCP pass (RQ5 coverage and RQ6) is the largest follow-up. It keeps no hosts in its published tables. Stores without readable sampled products were excluded, including those with unreadable robots.txt files, blocked requests and a requested crawl delay that was not waited out. The industry pass is kept only for the grade-by-industry table (Appendix D); the UCP pass replaces it for category figures.

All follow-up reads used public, read only requests, for up to five products per store. Shopify did not take part in or endorse the study.

## 5.5 Units and deduplication

The unit is the shop. 1,221,409 addresses were graded. We removed 34,511 stores that entered the frame through app reviews rather than the web graph or Tranco: they were selected because they reviewed SEO, metafield or bulk-editor apps, so they are reported as a separate group (Appendix A) and not mixed into the population estimate. Adding the 16,202 confirmed reviewer stores that are not already in the main sample would leave the mean grade at 28.6 (28.56 to 28.63). We also removed 153 theme demo stores and 18,563 addresses that reach a shop already counted (same myshopify domain or final host), leaving 1,168,182 stores, each counted once. Appendix B lists what happened to every address.

## 5.6 Validation

A separate parser checked the grading code by hand-fetching one product page from each of 30 stores in the October 12 pilot: 29 matched on all nine facts checked then, and in the 30th the checker took the wrong Product block. On October 27 a separate Claude session re-read the 141 product pages the crawler had graded for 30 random stores from the October crawl (raw HTML for all 30 stores, a real browser for

5). Six pages had been removed since the crawl. Of the other 135, 130 matched on Product JSON-LD and the nine facts checked then (all of them among today's 31), and the robots.txt and agents.md results matched for all 30 stores. The 5 that differed were all the pages of one store whose Product block is not valid JSON: the crawler counts such a block as no Product JSON-LD, as a strict parser would, while a lenient parser reads 8 of those 9 facts. On November 2 the 31-fact grade was recomputed by a separate Claude session with its own code from the same stored scans, and it matched the main figures exactly (1,168,182 stores, mean 28.6, median 32, 14.1% at 0, Dawn 31.3, Horizon 32.1, United States 27.6). Four earlier number checks, each run by a separate Claude session (an AI model) that had not written the code and was briefed only with the data and definitions, recomputed the published figures from the raw or per-store data: rates reproduced within 0.1 point, all 428 result chunks matched the counts file, and for the catalog sample 30 of 30 live lookups agreed with the stored result. On November 3 the UCP pass was checked against the main crawl: every store in the UCP pass matched a JSON-LD store record on final host, nearly all of the sampled products matched a graded page on host and path (the rest sit on stores rescanned with another product sample), the share of joined stores that print no Product JSON-LD (14.0%) is close to that of all 1,168,182 stores (14.1%), and the category totals in the pass's own count file equal the analysis tables. On November 3 a separate Claude session that had not read the findings report or the aggregate tables recomputed seven headline figures with its own code from the raw JSON-LD result files and the UCP tables: products categorized, the all, some and none split of stores, the lowest and highest product age ranges, categorized share for stores over 5,000 products, stores with a categorized product that print no attribute, products found in the catalog, and available US products found. All seven matched within 0.1 point. These checks were commissioned by the author and are not independent review in the research sense.

## 5.7 Statistics

Main-crawl figures cover every store in the frame, so we give them without sampling intervals: their main error is coverage of the frame, and each store's grade is itself based on up to five pages. Sample figures carry 95% intervals (Wilson for shares; for the catalog sample's product-level share, a ratio estimator with stores as clusters, design effect 2.5). To compare correlates we report the share of variance in the store grade explained by each grouping alone (eta squared) and by a joint linear model of theme (the 60 most common Theme Store IDs plus "other"), country (the 40 most common plus "other") and Tranco band. The grade is one number on a 0 to 100 scale; because it is the share of 31 facts, a store's grade moves in steps of about 3.2 points. UCP pass figures cover every store of the graded sample, so we give them without sampling intervals. With up to five products per store, a store's categorized share can only be 0, 20, 40, 60, 80 or 100%, so store figures describe the kind of store and not its whole catalog; Section 6.6 tests how far that sample can be trusted. Store figures from the UCP pass are averages over stores of the share of their sampled products, unless a sentence says products. Group tables report countries with at least 2,000 stores, top-ranked country subgroups with at least 200, and industries with at least 300 top-ranked stores.

# 6 Results

## 6.1 RQ1: Access is open and uniform

Agent files and crawler access are nearly uniform, so they do not separate stores.

| Check | Stores (n = 1,168,182) |
|---|---|
| Serve /agents.md (content varies by store) | 99.9% |
| Serve /llms.txt | 99.9% (98.9% identical to agents.md) |
| agents.md links to the store's UCP and MCP endpoints | 99.8% |
| Valid UCP profile at /.well-known/ucp, listing an MCP endpoint | more than 99.9% |
| UCP profile lists payment handlers | 99.7% |
| robots.txt blocks none of the 12 AI crawlers checked | 99.9% |
| robots.txt names any AI crawler at all | 0.8% |
| Home page has a meta description | 74.1% |
| Home page has Open Graph tags | 98.4% |

About 0.1% of stores block at least one of the 12 AI crawlers. In the earlier 40,746-store run the most blocked were training crawlers, CCBot (0.5% of stores) and GPTBot (0.4%). Robots.txt states what a store allows, not what its bot protection does; we did not send requests as each crawler.

## 6.2 RQ2: The average store prints 9 of 31 facts, and one in seven prints none

The mean store grade is 28.6 and the median 32. The average store prints 8.9 of the 31 facts, and the median store about 10. No store prints all 31, and the best reach 84. 53.3% of stores grade between 30 and 39, and only 0.3% reach 60 or more.

The basics are printed by most stores: 85.9% print a name, 85.5% a brand, 85.4% a price and 84.8% an image. The details are rare. 41.9% of stores print a category, 55.0% tell agents a product has variants or options, and under 1% tell agents a color (0.8%), size (0.7%) or material (0.4%). Appendix F gives all 31.

At the bottom, 14.1% of stores print no Product JSON-LD on any sampled page; this is essentially the same group as the 14.1% that score 0. Most of them are not blank: about 9% of all stores print Product microdata instead, so 5.0% print Product markup in neither format. Two quality problems sit on top: 3.6% of stores have malformed JSON-LD on a sampled page, which a strict parser may skip [18], and 7.1% have a page with two or more Product blocks, which can disagree on price or availability.

## 6.3 RQ3: GTINs are scarce at the source, and the agent layers do not add them

**Product level (catalog sample).** The catalog sample links the same products across L2, L3 and L5.

| Point | Measure | Result |
|---|---|---|
| L2 Shopify barcode field (.js file) | Any barcode on any variant / GS1-valid | 17.2% / 11.8% |

| Point | Measure | Result |
|---|---|---|
| L2, by vendor | GS1-valid when the vendor is not the store / is the store | 17.6% / 9.4% |
| L3 Product page JSON-LD | Page prints a GTIN when the product has a GS1-valid barcode | 38.8% (734 of 1,894) |
| L5 Global catalog | Barcode field returned | 0% of products found, including products with valid barcodes |

**Store level.** 84.0% of stores print no GTIN on any sampled page (14.1 points of that are stores with no Product JSON-LD at all), so 16.0% print one on at least one page, and 6.3% print one on every sampled page.

| Point | Measure | Result |
|---|---|---|
| L2 Shopify barcode field | GS1-valid GTIN on at least one of the first three products | 20.4% of stores (95% CI 19.3 to 21.4) |
| L3 Product page JSON-LD | GTIN on at least one / every sampled page | 16.0% / 6.3% of stores |
| L4 Store UCP endpoint (get_product) | Barcode returned | 0% of stores, including stores with valid barcodes |
| L4 Store UCP endpoint | Shopify product category set (products / stores with it on at least one sampled product) | about two thirds / about 4 in 5 |

Three readings follow. First, most products carry no GTIN in Shopify: 11.8% of sampled products have a GS1-valid barcode. Products sold under another vendor’s name, where a GTIN is more often expected, have one 17.6% of the time; this proxy is imperfect, because dropship suppliers and print-on-demand vendors also count as “other vendors”. About 93% of sampled stores that print no GTIN also had no valid barcode on the products we read. Second, when a valid barcode exists, the page prints it 38.8% of the time; 6.2% of sampled stores, or 30% of stores with a valid barcode, had one in Shopify and printed no GTIN on any sampled page. Third, neither L4 nor L5 returned a barcode field in the barcode and catalog samples, though L4 does return the Shopify category. The category figures are different measures and should not be compared directly. In the UCP pass, about two thirds of products have a Shopify category as the store’s UCP endpoint returns it, and about 4 in 5 stores have one on at least one of up to five sampled products, while 41.9% of all stores print a category on a product page (what the theme or an app shows agents on the page). Agents reading the page see a category on fewer stores than have one set. Section 6.6 shows which products lack one.

## 6.4 RQ4: Theme is the closest correlate, mostly through markup format

**Variance.** Theme alone explains 27.6% of the variance in the 31-fact store grade and 33.8% of the variance in whether a store prints Product JSON-LD. Country explains 1.0% and Tranco band 0.6%. In a joint model (R squared 0.28), removing theme costs 0.27, removing country 0.003 and removing rank 0.005. For a format-neutral outcome, printing Product markup in any format, theme explains only 4.7%, and for GTIN presence on a page theme explains 4.6%, country 1.4% and rank 1.1%. Theme therefore largely decides the markup format; none of these characteristics explains much of GTIN presence.

**Theme groups.** Stores on Shopify's current free themes almost always print Product JSON-LD; stores on the oldest free themes almost never do, but they print microdata instead.

| Theme group | Stores | Mean grade (31 facts) | Facts printed (of 31) | No Product JSON-LD | No Product markup in any format | GTIN on every page | Two or more Product blocks | Malformed JSON-LD |
|---|---|---|---|---|---|---|---|---|
| Current free themes (Dawn, Horizon and 23 more) | 307,712 | 31.2 | 9.7 | 1.4% | 1.3% | 4.1% | 4.1% | 1.2% |
| Debut | 43,574 | 24.0 | 7.4 | 18.8% | 7.8% | 0.2% | 2.7% | 12.3% |
| Oldest free themes (Supply, Minimal and 10 more) | 46,312 | 2.3 | 0.7 | 93.6% | 1.1% | 0.3% | 0.6% | 0.5% |
| All other themes | 770,584 | 29.4 | 9.1 | 14.1% | 6.6% | 8.0% | 8.9% | 4.2% |

92.6% of stores on the oldest free themes print Product microdata and no JSON-LD, so they fail a JSON-LD grade rather than print nothing. A clean JSON-LD block is not a complete one: current free themes print the block but their stores average only 9.7 of the 31 facts. Most pages leave out SKU (about 64%) and GTIN (about 92%), which depend on what the merchant enters (6.3), and almost none print the category, options and specs facts. Of the stores with no Product JSON-LD, 66% are on all other themes, 26% on the oldest free themes, 5% on Debut and 3% on current free themes. Appendix A gives every Theme Store theme and theme version.

**Theme and category.** Category is Shopify data returned by the UCP endpoint, not something the theme prints, so a theme cannot cause a category gap through what it prints; theme differences here reflect which stores run which theme. Within one theme family the pattern repeats. Stores on versions 2 to 9 of Dawn categorize about 3 in 5 of their sampled products, stores on versions 15 and up about 4 in 5. The category comes from Shopify's data and not from the theme, so this gap is not a difference in what the theme can do; it is consistent with older stores that were set up once and not revisited. Percentages in the table are rounded to the nearest 5 points.

| Dawn major version | Products categorized (store average) | All sampled products categorized | None categorized | No category field | GTIN on a page |
|---|---|---|---|---|---|
| 2 to 9 | 60% | 40% | 25% | 30% | 5% |
| 10 to 14 | 70% | 50% | 20% | 20% | 10% |
| 15 and up | 80% | 65% | 10% | 10% | 15% |

**Popularity rank.** Stores from the Tranco top 1 million average 34.7 to 35.8 (by band, below) and all other stores 28.4. The gap is about 6 to 7 points on the 100-point scale. Every Tranco-ranked store came

from the list used in our earlier run, so rank and source are confounded.

| Group | Stores | Mean grade (31 facts) | Facts printed (of 31) | No Product JSON-LD | GTIN on every page |
|---|---|---|---|---|---|
| Tranco top 100,000 | 841 | 35.8 | 11.1 | 7.5% | 25.3% |
| Tranco 100,001 to 1 million | 27,655 | 34.7 | 10.8 | 8.6% | 21.4% |
| Not from the Tranco-derived frame | 1,139,686 | 28.4 | 8.8 | 14.2% | 6.0% |

**Country.** Among countries with at least 2,000 stores, mean grades run from 26.1 (Philippines) to 31.7 (Brazil); the United States, with 503,695 stores, averages 27.6. Among top-ranked stores the differences between the highest countries are within their 95% intervals (for example Germany 37.1 plus or minus 0.7, Denmark 37.1 plus or minus 1.5, Spain 36.8 plus or minus 0.9), so we do not rank them; top-ranked US stores average 35.1. Appendix D has the table for the 12 largest countries.

**Industry.** Differences are modest. Among the 12 industries with at least 300 top-ranked stores, mean grades run from 32.3 (media) to 37.3 (sporting goods), with 95% intervals of plus or minus 0.3 to 1.3 points. The industry pass leans toward ranked and .com stores and leaves out stores with no category (more than a third of top-ranked stores), so Appendix D reports it as a sample. Category coverage is similarly modest across industries. Among the larger industries in the industry pass, the average share of a store’s sampled products that are categorized runs from about 75% to about 85%. These stores were chosen because the product we asked about had a category, which lifts all of them above the about two thirds of the whole pass, so the point is the small spread between industries. The stores whose asked product was explicitly Uncategorized are a separate group and average about a third, which again shows that category coverage belongs to the store (Section 6.6).

## 6.5 RQ5: The global catalog holds about 4 in 10 sampled products, covers US products best, and mirrors their text, without identifiers

**Coverage (UCP pass).** The pass looked up every variant of a large sample of products in the global catalog and found about 4 in 10 of the products. Counting only the products the store marked available, about half were found, and for US products in stock about 3 in 4.

Coverage is highest for US stores, and the catalog holds fewer products from stores elsewhere. We cannot see why Shopify’s catalog holds so few non-US products.

A GTIN on the page does not explain the country gap: across stores, a store’s share of products with a GTIN and its share found are uncorrelated (correlation near zero).

**What the catalog adds (UCP pass).** Of the found products matched to a graded page, about 4 in 5 of those with a catalog rating show no aggregateRating on the store’s own page. The review-app share is partial: 14.2% of the stores with a detected review app print aggregateRating on any sampled page, against 1.3% of stores without one (app detection only sees apps that leave a marker on the home page).

**Fidelity (catalog sample).** The October 22 sample looked up each store's three newest products by their first variant and found a little over a third; the November pass looked up every variant of up to five random products and found about 4 in 10. The two differ in sampling, in variants asked and in date (twelve days apart), and we do not separate those effects. The sample stays because it reads each product from the store and the catalog in the same hour, which fidelity needs.

| Measure | Result |
|---|---|
| Sampled products found in the catalog | a little over a third |
| Stores with at least one sampled product found | a little under half |
| Stores with three sampled products: none found / all found | a little over half / about a quarter |
| Same title / availability | nearly all / nearly all |
| Same description (first 1,500 characters) | nearly all |
| Same price / differs by more than one dollar | nearly all / under 5% |
| Barcode, SKU or category field returned | 0% |
| Generated selling points, features and specs | about 7 in 10 |
| Rating shown | about 1 in 5 |
| List price passed on for a real markdown | nearly all |
| Product-level option list missing / no option information at all | about 1 in 7 / under 5% |
| Catalog image alt text equals the product title | about 3 in 4 (a small live check) |

Coverage is close to all or nothing by store, which points to a store-level reason (opt-out, eligibility, country or indexing) that we did not test. The sampled products in the October sample are each store's newest, which may not be indexed yet, so its coverage may be understated (Section 8). Where a product is present, the catalog copies the store's text and price faithfully and adds generated summaries, but it carries no identifier field. Its alt text mostly repeats the title; stores themselves have alt text on about a quarter of images.

## 6.6 RQ6: Category coverage follows product age and store size, and almost no store prints the attributes under it

Of the products with a UCP answer, about two thirds carry a Shopify category, about 1 in 7 carry the explicit value "Uncategorized" and about 1 in 5 have no category field at all. Joined to the page data, about 60% of the answered and graded products are categorized and print Product JSON-LD, about 25% print Product JSON-LD without a category, about 5% have a category and no Product JSON-LD and about 10% have neither.

**Product age.** Old products mostly were not categorized. The share rises with every id range, from about 1 in 4 for the oldest to about 4 in 5 for the newest. The explicit value "Uncategorized" stays between about 1 in 7 and about 1 in 5 in every range, so the age effect is the missing category field: more than half of the oldest products have none, about 1 in 10 of the newest. Percentages in the table are rounded to the nearest 5 points.

| Product id range (oldest to newest) | Approximate years | Categorized | Explicit Uncategorized | No category field | Product JSON-LD on the page | Found in global catalog |
|---|---|---|---|---|---|---|
| under 1e12 | 2017 and before | 25% | 20% | 55% | 40% | 60% |
| 1e12 to 4e12 | 2018 to 2019 | 25% | 20% | 55% | 50% | 55% |
| 4e12 to 6e12 | 2019 to 2020 | 25% | 20% | 55% | 65% | 50% |
| 6e12 to 7e12 | 2021 | 40% | 20% | 45% | 75% | 50% |
| 7e12 to 8e12 | 2022 | 60% | 15% | 25% | 85% | 50% |
| 8e12 to 9e12 | 2023 to 2024 | 70% | 15% | 15% | 90% | 45% |
| 9e12 to 10e12 | 2024 to 2025 | 75% | 15% | 10% | 90% | 45% |
| over 10e12 | 2025 to 2026 | 80% | 15% | 10% | 90% | 35% |

The same holds inside a store. In the stores where some but not all sampled products are categorized, categorized products sit later in the store's creation order (mean rank about 0.55, where 0 is the oldest sampled product and 1 the newest) than products that are not (about 0.45); products with no category field are the oldest (about 0.43). Among stores with three or more sampled products, nearly twice as many have a newest sampled product that is categorized and an oldest that is not as the reverse. A product id is an order of creation, not a date, so the years in the table are estimates.

**Store size.** The biggest catalogs are the least categorized. Stores with more than 5,000 published products have about 4 in 10 of their sampled products categorized, about half of these stores have none categorized, and about half of their products have no category field. They are also the stores most likely to print identifiers: the average share of their sampled pages with a GTIN is about 30% and with a SKU about 75%, against about 10% and about 30% for stores with 10 or fewer products. The correlation between a store's categorized share and the logarithm of its published product count is weakly negative (about -0.1). Large catalogs are mostly old catalogs, which is consistent with the age result, but we did not separate the two. Percentages in the table are rounded to the nearest 5 points.

| Published products | Products categorized (store average) | All sampled products categorized | None categorized | No category field | GTIN on a page | SKU on a page |
|---|---|---|---|---|---|---|
| 1 to 10 | 70% | 60% | 20% | 15% | 10% | 30% |
| 11 to 50 | 70% | 50% | 20% | 15% | 10% | 35% |
| 51 to 200 | 65% | 50% | 20% | 20% | 10% | 45% |
| 201 to 1,000 | 65% | 45% | 20% | 20% | 15% | 50% |
| 1,001 to 5,000 | 55% | 40% | 35% | 35% | 25% | 65% |
| Over 5,000 | 40% | 30% | 50% | 50% | 30% | 75% |

(Stores with an unknown product count are left out of the table.)

**Category belongs to the store.** With up to five sampled products per store, a store's share can only move in steps of 20 points, so we tested whether it describes the store. Among the stores with exactly five answered products, about two thirds have all five or none categorized; if each product were

categorized independently at the same rate (about two thirds), only about 1 in 9 would. Category coverage is therefore a property of the store, and the sample mostly tells which kind of store it is. Across all stores with an answered product, about half have every sampled product categorized, about 3 in 10 some and about 1 in 5 none.

| Sampled products categorized | Share of stores with 5 answered products | Expected if each product were categorized independently (about two thirds) | All stores with an answered product |
|---|---|---|---|
| All five | 50% | 10% | 50% |
| Some, not all | 30% | 90% | 30% |
| None | 20% | under 5% | 20% |

**Attributes defined and printed.** For categorized products with a category path, Shopify's taxonomy lists about 9 attributes on average (median 9; most have 5 or more). The mean runs from 13.7 for Health and Beauty and 13.6 for Electronics through 11.6 for Apparel and Accessories to 3.3 for Business and Industrial, and is 0 for Services, Gift Cards and Product Add-Ons (Appendix G lists the 12 largest categories; the tables archive has all 24). On the page side, among stores that have a categorized product and Product JSON-LD, nearly all print none of color, size, material, pattern, age group or gender and no additionalProperty on any sampled page; under 2% print at least one of the named attributes and under 2% print an additionalProperty (some print both). Across these stores the named attributes are printed by about 1% (color and size) or less (material, age group, gender and pattern). Appendix F gives the same facts as shares of all stores. These figures are a gap between what the taxonomy lists and what pages print. They are not a measure of what merchants filled in: the UCP endpoint returns no metafields, so we cannot tell whether the attribute values are empty in Shopify admin, or filled in and not printed (Section 8).

**Six merchant facts.** Only about 1 in 20 of the answered and graded products have all six merchant facts (category, product type, at least one tag, GTIN, SKU and a description in the markup); the typical product is missing three. Percentages in the table are rounded to the nearest 5 points.

| Facts missing (of six) | Share |
|---|---|
| 0 | 5% |
| 1 | 15% |
| 2 | 20% |
| 3 | 20% |
| 4 | 25% |
| 5 | 10% |
| 6 | 5% |

GTIN is the fact missing most often (about 9 in 10 products), then SKU (a little over half), product type (over 4 in 10), tags (about 4 in 10), the Shopify category (about a third) and the description in the markup (about 1 in 5). At store level, a little under half of the stores have every sampled product categorized and Product JSON-LD on every sampled page, about 1 in 10 stores of those with an

answered product also print a GTIN on at least one page, and fewer than 1 in 20 stores have every sampled product categorized and print GTIN, SKU and description on every sampled page. Facts go missing in clusters (Appendix G): product type and tags are missing together (tags missing for about 3 in 5 products with no product type, against about 4 in 10 overall) but are nearly independent of the Shopify category (no category for about 4 in 10 products with no type and about a third of products with no tags, against about a third overall); a product with no Product JSON-LD has no category more than half of the time, which fits the older-theme and old-store cluster above; we did not separate them.

# 7 Discussion

## 7.1 Variation is in merchant data and themes, not platform files

The agent-facing layers that Shopify controls (L4, L5, robots.txt defaults) are present on nearly every store and do not vary. What varies is what the merchant entered and what the theme prints. For the identifier that lets an agent tell one product from another, presence is low in the merchant's own data: 11.8% of sampled products, and 17.6% of those sold under another vendor's name, have a valid barcode. Advice to merchants that centres on files such as llms.txt or agents.md addresses the layer that is already uniform. The data that is missing is the detail: category, options and specs such as size, color and material, which fewer than half of stores print (Appendix F).

## 7.2 Theme is associated with markup format and completeness

Theme explains far more of the grade than rank, country or industry, but largely through format: the oldest free themes print microdata, and with a format-neutral measure their stores are as likely as any to print some Product markup. This is an association. Stores that never update their theme may also neglect their product data. A theme change could add JSON-LD without new product data; we did not test this.

## 7.3 Platform layers mirror the store; they do not fill gaps

The global catalog reproduces store text and price with high fidelity and adds generated summaries, but it returns no barcode, SKU or category field, and its alt text mostly repeats the title. An agent that compares sellers through the catalog's public responses cannot match products by identifier; product matching by text is hard for products a model has not seen [4]. It is also uneven by geography: about 3 in 4 of the available products of US stores are in it, and fewer from stores elsewhere (Section 6.5). Whether the catalog uses identifiers internally, and why most non-US products are absent, cannot be seen from outside.

## 7.4 What this study does not show

We measure what an agent could read, not what agents do with it. Agent choices are also shaped by position, ratings, sponsorship labels and model version [7], and by how content is written [8]. A store with a perfect grade is not shown to be recommended more often. In October 2026 we ran a small

observational pilot on this question: 196 stores from the earlier run, split into high grade (90 or above) and low grade (below 50) on the earlier nine-fact version of the grade, within two Tranco bands, one product each, asked by exact title and by description to two models with web search (Gemini 3 Flash and GPT-5 mini, through OpenRouter), 1,259 answers in all. We found no detectable difference: stratified differences in how often the store was cited ran from minus 4.5 to plus 10.2 points, and every 95% interval included zero. The pilot could detect only differences of about 12 points or more and was not randomised. Testing the link properly needs controlled experiments on agent outputs, which we leave to future work.

### 7.5 Category coverage is a backfill problem, and the attribute layer is empty on the page

Sections 7.1 and 7.2 found the platform's files uniform and the theme deciding the markup format. The category results add a third pattern. Shopify's category exists for most new products and for few old ones (about 1 in 4 of the oldest products, about 4 in 5 of the newest), and the stores with the biggest catalogs, which are mostly old catalogs, have the least (about 4 in 10 over 5,000 products). The store's vertical changes little (Section 6.4). Inside one store, newer products have a category and older ones do not. A reading that fits the data is that Shopify's category suggestions came after most old products were created and nobody went back to them; we did not test that. Under the category sits a second gap: the taxonomy lists about nine attributes for the average categorized product, and nearly all of the stores that have a categorized product and Product JSON-LD print none of them. The platform fills what it can fill for new products; the back catalog and the attribute values depend on the merchant. Datify is a Shopify app that edits product data (Section 9), so this is the result where our interest is closest to the finding. The tables behind it are published (Appendix G), and the join and the counts were recomputed by a separate session (Section 5.6).

## 8 Limitations and threats to validity

**Construct validity.** The grade checks presence, not correctness: a wrong price or a fake GTIN counts as present. It reads JSON-LD only, so microdata stores score low by design; we report a format-neutral presence measure beside it. Page GTINs are counted at 8 or more digits without a check-digit test, while Shopify barcodes must pass the GS1 test, so page presence is measured more leniently than source presence. A store's grade covers up to five randomly drawn product pages (4.76 on average), the barcode sample up to three products per store, and the catalog sample each store's first three products in products.json order and only each product's first variant for the catalog lookup. "Vendor is not the store" is a rough proxy for products expected to carry a GTIN. The category is what the store's UCP endpoint returns; we cannot tell whether a merchant or Shopify's own suggestions set it. The attributes defined are counted from a copy of Shopify's taxonomy generated on September 20, 2026, and are what the taxonomy lists, not what a merchant filled in: the UCP get_product answer has no metafields, so attribute values in Shopify admin are not observable and "printed" does not mean "filled in". The product id is an

order of creation with approximate years, not a date. A store's category figures rest on up to five products, so a single store's share is coarse; the group figures are stable because the groups are large.

**Coverage (external validity).** The frame is storefronts linked from the public web or in our earlier lists. Stores the web graph never mentions, headless stores and stores behind another CDN are missing. Test and abandoned stores remain in the myshopify.com groups. The industry pass covers the first 0.83 million stores scanned and leans toward ranked and .com stores; the barcode sample covers the first 1.0 million. The catalog sample uses each store's newest products, which may not be in the catalog yet. The UCP pass covers a large subset of the graded stores, not all 1,168,182, and only stores with an answered product enter the store figures. Its sampled products are the pages the main crawl drew from the first product sitemap file (Section 5.2), so for large catalogs the age mix of the sample may differ from the whole catalog.

**Comparisons (internal validity).** Rank is confounded with the list a store came from. Theme groups rely on theme names and IDs, so renamed copies fall into "other". The variance shares are descriptive and do not establish cause.

**Timing.** This is one snapshot from October 2026, with the UCP pass read on November 2 and 3, about two weeks after the pages it is joined to. Store and catalog were read hours apart in the catalog sample. Shopify changes its agent layers often (they are less than a year old), so the L4 and L5 results describe the days we called them. A repeat on a separate sample of stores on October 27 found no change.

**Reproducibility.** The per-store data is private (Section 9), so outside readers can check the aggregates against the published tables but not re-derive them from raw records.

## 9 Ethics and conflict of interest

**Crawl rules.** The crawl read only public pages and files, identified itself honestly (DatifyResearch/1.0), followed each store's robots.txt, made one request at a time per store at least one second apart, honoured Crawl-delay and Retry-After, and never logged in or touched a cart, checkout, account page or form. It did not try to get past blocks. Shopify was not told about the crawl before or during the run, did not take part in the study and does not endorse it. Domains that did not point at Shopify got no request; about 3,400 that did but were not stores got one home-page request. The whole crawl made 20.9 million requests over six days, about 40 per second on average, nearly all to Shopify's shared infrastructure.

**Personal data.** The crawl stored each home page's title and description as written, which can include contact details a store printed itself. No store is named in this paper, and per-store records are not published. To have a store left out of future runs, email apps@jadepuma.com. The study had no institutional ethics review; it was run by a company, not a university.

**Conflict of interest.** The author builds Datify, a Shopify app that edits product data and structured data, and Datify intends to cite figures from this study in its marketing. A study that finds product data gaps serves that business. To limit the effect, the measures, including the 31 grade facts, are defined in the crawler code released with the paper, nine of them follow Google's merchant listing documentation [14],

every headline number was recomputed by separate AI sessions that did not write the code (Section 5.6), the null result of our own pilot is reported (Section 7.4), and Datify's own merchant data is not used. The text was drafted with an AI assistant (Claude) and checked by the author. No external funding.

## 10 Data and code availability

The crawler, aggregation scripts and their tests are published as a zip archive with this paper (datify-ai-readiness-2026-scripts.zip), and aggregate tables (CSV) as a second archive (datify-ai-readiness-2026-tables.zip). Both are linked from the study page at https://datify-app.com/pages/ai-readiness-study. The scripts archive holds the storefront crawler, the analysis scripts and the category method scripts. The tables archive holds the summary tables and the category tables of Appendix G. Per-store records are kept private because they identify stores.

## 11 Conclusion

Across 1,168,182 Shopify storefronts, agent files are present on 99.8% or more of stores, while Product JSON-LD is absent on 14.1%, a page GTIN on 84.0%, and the average store prints 9 of 31 product facts (grade 28.6 of 100). GTINs are scarce in the merchants' own data (11.8% of sampled products have a valid barcode), the page prints a valid barcode about four times in ten, and neither the store's UCP endpoint nor Shopify's global catalog returns one. Theme is the store characteristic most closely tied to page data, largely through markup format. Shopify's own category is set on about two thirds of the sampled products but on about 1 in 4 of the oldest and on about 4 in 10 of the products in stores with more than 5,000 products, and nearly all of the stores with a categorized product and Product JSON-LD print none of the roughly nine attributes Shopify lists for it. Shopify's global catalog holds about 4 in 10 of the sampled products, and about 3 in 4 for US products in stock. Whether any of this changes what agents recommend is open; our small pilot found no detectable effect, and only controlled experiments on agent behaviour can answer it.

## References


1. R. Meusel, C. Bizer, H. Paulheim. A Web-scale Study of the Adoption and Evolution of the schema.org Vocabulary over Time. WIMS 2015.
2. Web Data Commons. WDC JSON-LD, Microdata, RDFa Data Corpus 2024. University of Mannheim, from the October 2024 Common Crawl.
3. HTTP Archive. Web Almanac 2024: Structured Data.
4. R. Peeters, R. C. Der, C. Bizer. WDC Products: A Multi-Dimensional Entity Matching Benchmark. EDBT 2024.
5. S. Longpre et al. Consent in Crisis: The Rapid Decline of the AI Data Commons. NeurIPS 2024, Datasets and Benchmarks Track.

6. E. Liu, E. Luo, S. Shan, G. M. Voelker, B. Y. Zhao, S. Savage. Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. IMC 2025.
7. A. Allouah, O. Besbes, J. D. Figueroa, Y. Kanoria, A. Kumar. What Is Your AI Agent Buying? Evaluation, Biases, Model Dependence, and Emerging Implications for Agentic E-Commerce. arXiv 2508.02630.
8. P. Aggarwal et al. GEO: Generative Engine Optimization. KDD 2024.
9. Ethercycle. State of Ecommerce 2026. Industry report.
10. Common Crawl. Web graphs, July to September 2026 domain graph.
11. V. Le Pochat et al. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. NDSS 2019.
12. GS1. How to calculate a check digit.
13. Shopify. Global catalog, agent documentation.
14. Google Search Central. Merchant listing (Product, Offer) structured data.
15. A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, A. Narayanan. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. CSCW 2019.
16. Q. Scheitle et al. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. IMC 2018.
17. K. Ruth, D. Kumar, B. Wang, L. Valenta, Z. Durumeric. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. IMC 2022.
18. R. Meusel, H. Paulheim. Heuristics for Fixing Common Errors in Deployed schema.org Microdata. ESWC 2015.
19. A. Primpeli, R. Peeters, C. Bizer. The WDC Training Dataset and Gold Standard for Large-Scale Product Matching. WWW 2019 Companion.
20. S. Yao, H. Chen, J. Yang, K. Narasimhan. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. NeurIPS 2022.
21. Vercel. The rise of the AI crawler. December 2024.

## Appendix A: Themes, attributes and app reviewers in the full crawl

All figures are from the 1,168,182-store main sample unless stated, and were recomputed by a separate number check with no mismatches. Full tables: themes.csv (every Theme Store theme), theme-versions.csv, additional-property-names.csv, product-keys.csv and reviewers-*.csv in the published aggregates.

- **Every Theme Store theme.** 743,854 stores report the Theme Store ID of one of 346 listed themes; 328 themes have at least five stores. Among the 52 themes with at least 2,000 stores, mean grades run from 13.5 (Blockshop) to 37.8 (Warehouse), no Product JSON-LD from 0.6% (Prestige) to 54.2% (Flow), and GTIN on every page from 1.6% (Stiletto) to 21.5% (Hyper).

Stores on paid themes (255,023 stores on 304 themes) print no Product JSON-LD 9.4% of the time; stores on free themes matched by ID (488,799 on 24 themes) 1.8%.

- **Theme versions.** Within a theme, the version matters for identifiers. Dawn stores on versions 2 to 5 print a GTIN on 0.5% to 0.9% of pages on average, on versions 6 to 14 about 7%, on version 15 11.7% and on version 16 16.6%. Impulse stores on version 1 print no Product JSON-LD 69.1% of the time, against about 1% on versions 2 to 10. Version numbers are each theme's own and are not comparable across themes.
- **Theme demo stores.** In the earlier run's scan of the 345 Theme Store demo stores, six printed no Product JSON-LD (Fashionopolism, Retina, Responsive, Loft, Maximize and Force). In the full crawl, 55% to 87% of real stores on those themes print none, against 14.1% of all stores.
- **Product attributes.** 1.2% of stores put an additionalProperty on a sampled product page and 1.3% give any attribute detail. The most common additionalProperty names are size, color pattern, customization available, color, age group, target gender and material. Of all stores (including the 14.1% with no Product JSON-LD), 51.2% print an sku key, 10.1% a gtin key (12.9% any gtin, gtin8, gtin12, gtin13 or gtin14 key) and 41.9% a category. 6.3% print aggregateRating; ratings are not one of the 31 facts.
- **App reviewers.** 17,551 stores matched by name and country to reviews of SEO, metafield and bulk-editor apps. Among stores from the Tranco 100,001 to 1 million band, the 1,065 reviewers average 38.6 against 34.3 for the 25,347 that reviewed none of these apps, print no Product JSON-LD 5.5% against 8.9%, show no GTIN 49.2% against 61.1%, and have two or more Product blocks 27.4% against 17.0%. This describes those stores, not the effect of an app.
- **Earlier run.** The October 12 to 14 run of 40,746 stores leaned toward higher-traffic stores, so its numbers are not comparable with the main results. Its barcode pass found 41.0% of stores with a valid barcode in Shopify, 30.7% printing any GTIN, and 13.0% with a valid barcode the page left out (as reproduced by a later number check).

# Appendix B: Outcome of every address

| Outcome | Addresses |
|---|---|
| Graded | 1,221,409 |
| Points at Shopify but the secure connection failed | 186,324 |
| Alias of a store already counted | 197,763 |
| Closed or unpaid store (HTTP 402) | 122,117 |
| Password page | 84,109 |
| Page not found (404) | 78,598 |
| No products | 41,562 |
| Frozen store (HTTP 423) | 10,968 |
| Other HTTP errors | 10,001 |
| Unreachable | 7,061 |

| Outcome | Addresses |
|---|---|
| Not Shopify, no readable product pages, or refused | 6,928 |

## Appendix C: Theme groups

- **Current free themes:** Dawn, Horizon, Spotlight, Refresh, Craft, Sense, Crave, Studio, Ride, Taste, Colorblock, Publisher, Origin, Trade, Fabric, Atelier, Savor, Tinker, Heritage, Vessel, Pitch, Dwell, Ritual, Fresh, Agency.
- **Oldest free themes:** Supply, Minimal, Simple, Brooklyn, Narrative, Venture, Boundless, Express, Pop, Jumpstart, Launch, New Standard.

## Appendix D: Country and industry tables

| Country | Stores | Mean grade (31 facts) | Facts printed (of 31) | No Product JSON-LD | GTIN on every page |
|---|---|---|---|---|---|
| United States | 503,695 | 27.6 | 8.6 | 16.6% | 5.4% |
| United Kingdom | 100,762 | 28.9 | 9 | 13.8% | 7.1% |
| Canada | 49,406 | 27.1 | 8.4 | 17.6% | 5.4% |
| Germany | 44,785 | 31.4 | 9.8 | 8.7% | 12.3% |
| India | 41,102 | 29.5 | 9.2 | 11.4% | 2.3% |
| Australia | 41,101 | 28.5 | 8.9 | 14.9% | 6.2% |
| France | 36,363 | 29.9 | 9.3 | 9.9% | 7.7% |
| Japan | 28,298 | 29.1 | 9.1 | 10.5% | 7.3% |
| Netherlands | 24,836 | 30.9 | 9.6 | 8.9% | 14.4% |
| Italy | 23,768 | 29.9 | 9.3 | 9.9% | 8.6% |
| Spain | 18,771 | 30 | 9.4 | 9.6% | 8.8% |
| Hong Kong | 17,378 | 30.8 | 9.6 | 9.9% | 5.0% |

| Industry | Top-ranked stores | Mean grade (95% half-width) | Mean grade, all ranks |
|---|---|---|---|
| Apparel and accessories | 6,262 | 35.9 (0.3) | 31.3 |
| Health and beauty | 2,209 | 34.8 (0.5) | 30.8 |
| Home and garden | 2,016 | 35.3 (0.6) | 30.4 |
| Electronics | 885 | 35 (0.8) | 31.9 |
| Arts and entertainment | 880 | 33.8 (0.8) | 29.4 |
| Food, beverages and tobacco | 819 | 33.5 (0.9) | 29.5 |
| Sporting goods | 740 | 37.3 (0.9) | 32.2 |
| Vehicles and parts | 492 | 33.9 (1.2) | 30.4 |
| Toys and games | 488 | 33.7 (1.1) | 30.4 |
| Media | 470 | 32.3 (1.3) | 28.9 |

| Industry | Top-ranked stores | Mean grade (95% half-width) | Mean grade, all ranks |
|---|---|---|---|
| Hardware | 426 | 34.6 (1.2) | 31.5 |
| Furniture | 375 | 36.7 (1.2) | 32.3 |

The industry means above come from the industry pass (Section 5.4), a sample; the UCP pass replaces it for category figures (Section 6.6).

## Appendix F: The 31 product facts

Share of the 1,168,182 main-sample stores that print each fact in Product JSON-LD on at least one sampled product page (up to five). Stores with no Product JSON-LD count as not printing any. Within each group facts are ordered by share. The grade is the number of facts a store prints divided by 31, times 100.

| Group | Fact | Stores | Share of stores |
|---|---|---|---|
| The basics (8) | Name | 1,003,680 | 85.92% |
| The basics (8) | Brand | 999,356 | 85.55% |
| The basics (8) | Availability | 999,068 | 85.52% |
| The basics (8) | Price currency | 997,912 | 85.42% |
| The basics (8) | Price | 997,630 | 85.4% |
| The basics (8) | Image | 990,905 | 84.82% |
| The basics (8) | Product URL | 974,702 | 83.44% |
| The basics (8) | Description | 971,914 | 83.2% |
| Finding the exact product (6) | Variants and options | 642,291 | 54.98% |
| Finding the exact product (6) | SKU | 604,436 | 51.74% |
| Finding the exact product (6) | GTIN (barcode number) | 186,526 | 15.97% |
| Finding the exact product (6) | MPN | 74,812 | 6.4% |
| Finding the exact product (6) | Manufacturer | 6,544 | 0.56% |
| Finding the exact product (6) | Model | 1,150 | 0.1% |
| Category, options and specs (11) | Category | 488,871 | 41.85% |
| Category, options and specs (11) | Weight or dimensions | 23,945 | 2.05% |
| Category, options and specs (11) | Extra named attributes | 13,665 | 1.17% |
| Category, options and specs (11) | Color | 9,059 | 0.78% |
| Category, options and specs (11) | Size | 8,496 | 0.73% |
| Category, options and specs (11) | Material | 4,645 | 0.4% |
| Category, options and specs (11) | Audience (gender, age group) | 3,395 | 0.29% |

| Group | Fact | Stores | Share of stores |
|---|---|---|---|
| Category, options and specs (11) | Pattern | 1,173 | 0.1% |
| Category, options and specs (11) | Country of origin or assembly | 913 | 0.08% |
| Category, options and specs (11) | Category-specific detail | 486 | 0.04% |
| Category, options and specs (11) | Certifications | 121 | 0.01% |
| The offer (6) | Price valid until | 177,994 | 15.24% |
| The offer (6) | Item condition | 84,947 | 7.27% |
| The offer (6) | Seller | 78,274 | 6.7% |
| The offer (6) | Return policy | 19,691 | 1.69% |
| The offer (6) | Shipping details | 19,027 | 1.63% |
| The offer (6) | Warranty | 159 | 0.01% |

The size, color, material and other specs rows show what stores print on the page. We cannot see a store's metafields from outside, so they may undercount stores that fill in metafields but never show them to agents. "Variants and options" counts stores that tell agents a product comes in options, and the size and color rows count stores that print the actual values.

**Left out of the grade** (full list in the published tables, left-out-fields.csv): aggregateRating and review (reviews are not graded); releaseDate; isRelatedTo, isSimilarTo, isAccessoryOrSparePartFor and isConsumableFor; alternateName; general page fields (keywords, productID, mainEntityOfPage, title, inLanguage, video, logo, slogan, award, sameAs, disambiguatingDescription, additionalType, other dates, potentialAction, speakable); keys that are not schema.org names (bed_size, package_contents, option, barcode and similar); and keys printed by fewer than 20 stores.

# Appendix G: Category and missing-fact tables from the UCP pass

Source: the UCP pass of November 2 to 3, 2026, joined to the main crawl's pages (Section 5.4). Product shares use the sampled products with a UCP answer. Percentages are rounded to the nearest 5 points. Published as CSV files in the tables archive (attributes-defined-by-top-category.csv, attributes-defined-vs-printed.csv, six-facts-missing.csv, six-facts-missing-by-fact.csv, co-missing-matrix.csv, summary.json).

### G1. Attributes the taxonomy defines, by top level category (the 12 largest categories).

| Top level Shopify category | Mean attributes defined |
|---|---|
| Apparel & Accessories | 11.6 |
| Home & Garden | 9.6 |
| Health & Beauty | 13.7 |
| Food, Beverages & Tobacco | 7.4 |
| Arts & Entertainment | 8.1 |

| Top level Shopify category | Mean attributes defined |
|---|---|
| Sporting Goods | 5.8 |
| Media | 3.5 |
| Electronics | 13.6 |
| Vehicles & Parts | 8.1 |
| Toys & Games | 7.4 |
| Animals & Pet Supplies | 4.4 |
| Hardware | 4.0 |

**G2. Which of the six merchant facts are missing together (answered and graded products).** Each row is the set of products missing that fact.

| Given this is missing | No category | No product type | No tags | No Product JSON-LD | No GTIN | No SKU | No description | No image |
|---|---|---|---|---|---|---|---|---|
| (base rate, all products) | 35% | 45% | 40% | 15% | 90% | 55% | 20% | under 5% |
| category | 100% | 50% | 40% | 25% | 90% | 55% | 30% | under 5% |
| product type | 40% | 100% | 60% | 15% | 90% | 65% | 20% | under 5% |
| tags | 35% | 65% | 100% | 15% | 95% | 70% | 25% | under 5% |
| Product JSON-LD | 55% | 40% | 35% | 100% | 100% | 100% | 100% | under 5% |
| GTIN | 35% | 45% | 40% | 15% | 100% | 60% | 25% | under 5% |
| SKU | 35% | 50% | 50% | 25% | 95% | 100% | 35% | under 5% |
| description | 50% | 45% | 45% | 70% | 100% | 90% | 100% | 5% |
| image | 55% | 65% | 65% | 20% | 95% | 75% | 60% | 100% |