U.S. Internet

The Battle for the Persistent Consumer AI Agent Begins

Apple, Google, OpenAI, Meta and startups like Instinct have laid out different visions for how each sees the persistent consumer agent space playing out. The criteria for winning likely include: 1) trust, 2) low-cost compute, 3) access to personal connectors, and the prize is upwards of ~$10T.

U.S. Internet
POSITIVE

U.S. Internet
Ross Sandler
+1 415 263 4470
ross.sandler@barclays.com
BCI, US

Alex Hughes
+1 212 526 3069
alexander.hughes@barclays.com
BCI, US

Michael DiSanto
+1 212 526 1054
michael.disanto@barclays.com
BCI, US

Alexander Kessinger
+1 212 526 1324
alexander.kessinger@barclays.com
BCI, US

The Key Take-Away: The biggest turf battle in consumer internet has kicked off in '26 around persistent consumer AI agents. The importance of this space is not to be taken lightly, we see the market caps of most of the Consumer Mag-4 (Apple, Google, Meta, and OpenAI - private; not covered) reacting to how market share plays out over the coming years. Unlike enterprise, persistent consumer agents are likely to be free for the end user and funded by a combination of ads and referral fees on transactions. With substantial ad stacks and $300+ in US ARPU, this puts GOOGL and META in early pole position from a business model standpoint. The cost of hosting free consumer agents runs about $4/DAU per month at scale with connectors in place, ~10x more expensive compared to ChatGPT's (owned by OpenAI) inference cost per free user of sub-$0.50 per month. Meta and Google are likely most prepared to finance this heavy cost and have large distribution advantages from mobile, and it's yet to be determined if Apple and OpenAI are willing to spend upwards of $40B+ in extra annual compute to win at the free tier. OpenAI's new agentic product Dots looks very compelling, but more geared towards work and only currently in higher paying tiers. Beyond the compute cost, we believe the race comes down to which company can earn the trust of the user around security and encryption, and relatedly, can convince the user to "connect" their agent to the most downstream personalized services.

Key Conclusions in Our Report Include:

1. **Low(er) Cost Compute And High Ad Revenue Per DAU Matters:** Persistent Consumer Agents cost ~10x more than chatbots. This shows up in two primary areas: a) the virtual machine, and b) the model API calls. With a large base of advertisers who know what cost-per-transaction is, Google and Meta have a huge advantage. Standardization around merchant fees and APIs (like affiliate marketing decades ago) will help companies like Instinct (private; not covered) generate revenue earlier in their lifecycle.

2. **Trust and Security Will Be Paramount:** Meta's Muse was constructed by some of the core encryption experts from WhatsApp, which we think over time may help consumer confidence ramp. Apple and Google are in the best starting position around this issue. Startups likely need to get this right out of the gate to establish trust. Instinct recently disclosed that 40% of users provide a credit card within three weeks of onboarding, a good

Barclays Capital Inc. and/or one of its affiliates does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.

Please see analyst certifications and important disclosures beginning on page 11.
Completed: 30-Sep-26, 02:11 GMT Released: 30-Sep-26, 04:10 GMT Restricted - External

sign for establishing trust early in the process. Storing private credentials and making sure these are not exposed unless the user grants permission will be critical to earning trust.

3. **Access to Personal Connectors:** The first to win "Connectors" stands to benefit. Persistent agents are only as good as what the consumer exposes them to. Connectors to things like email, calendar, payment, logins, shopping, travel, etc, are a way to get one's AI agent to accomplish a wider range of tasks. We don't see consumers connecting many agents to these downstream services, so getting there first may have an advantage. Right now, Muse and ChatGPT are well positioned.

4. **Architecture Matters:** Below we illustrate what we have seen from Apple, Meta, and OpenAI around how the persistent agent stack works. Persistent agents need to run most compute in the cloud, not locally, to accomplish most tasks. Sub-agents may be working continuously around the clock when the user is offline. Apple could offload some lighter weight features to the on-device AI models, but we think this would represent a very small percentage of activity. Related to #1 above, to win, these consumer companies are likely to have to foot the bill of running VMs and APIs in the cloud on behalf of users for long periods of time.

The Persistent Agent Field in 2026

Consumer persistent AI agents are likely to become a major category going forward. In recent weeks we have seen a number of startups introducing Instinct-like offerings, and we would guess there are more on the way. This space likely has room for a few new entrants, but we think that the prize (we estimate upwards of $10T in market cap up for grabs) is too big for the largest consumer internet companies to ignore.

For this reason, we anticipate competitors need a combination of product, distribution, models, compute, and of course everything offered for free with transactions supporting the business model.

The current field is led by Muse and Instinct, but we would guess Gemini Spark, Grok Bot and Siri are likely to be heavily in the mix in '27.

FIGURE 1. **The Leading Persistent Agent Services Today**

	Muse	Instinct	OpenAI Dots	Siri	Gemini Spark	Grok Bot
Where It Runs	Virtual Machine	Virtual Machine	Virtual Machine	Device + Cloud	Virtual Machine	Virtual Machine
Truly Persistent?	✓	✓	✓	✗	✓	✓
Price	Free, Pro Tiers	Free	$100/mo +	Free	$19.99/mo	$20.00/mo
Distribution	Muse app, Meta AI glasses/charm, WhatsApp, Instagram	Currently Invite Only	ChatGPT, Codex	Siri App and iOS	Chrome, Android, Gmail, Workspace, YouTube	Grok app, X, Cursor
Gaps	Trust	Funding/compute scale	Expensive, only available via OpenAI apps	Not persistent	Not free, only accessible via Google apps	Not free, limited adoption

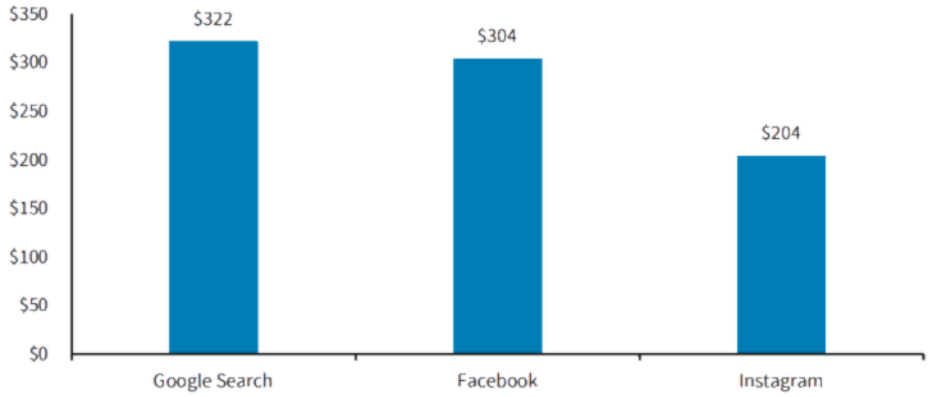
Source: Barclays Research, Company Reports

We think ad revenue per DAU is one of the best ways to think about the future state of the persistent agent revenue model. Digital advertising is decades old, and direct response marketers are at the point where buying ads on various platforms is down to an exact science where every click and conversion can be measured. Ads purchased on Google Search and Meta's Instagram or Facebook already have an embedded cost-per-acquisition framework across all direct response categories. Shopify's merchants know "for every $1 in ad spend, $X

dollars in gross profit" is likely to be achieved as the user navigates to their website or app and makes a purchase. The same can be said for online travel, insurance, food delivery, and just about any service that runs on the web.

Persistent agents charging a transaction fee around downstream gross profit dollars on a category-by-category basis is more or less how the direct response ad industry already operates. Hence, we can use ad revenue per DAU as a proxy for what kind of revenue a Muse agent could facilitate once all the supplier relationships are established (many of which are already built into Meta's ad manager stack). Interestingly, both Google and Meta have gone down the route of "native checkout" inside surfaces like Search and Instagram in years past, and have come up with flows that merchants have come to appreciate (iFrame browser loads inside apps, etc.), so the leap to transactions inside Muse or Gemini Spark should be rather simple to set up.

FIGURE 2. **U.S. Revenue Per DAU for the Largest Ad Stacks**



2026E
Source: Barclays Research, Company Reports

The same is likely true for companies like Instinct. As industry norms are established around transaction flows inside persistent agents (as we saw with affiliate marketing commissions 20 years ago), it likely makes it easier for startups to generate revenue earlier in their lifecycle. Instinct mentioned recently that 50% of its transactions are for travel bookings, which speaks to the appetite for a new consumer experience in these various transaction categories.

The Hosting Cost of Persistent Agents

There are two primary inference cost buckets for persistent agents: 1) the VM, and 2) the API calls. We can observe from contract pricing on AWS that a single instance with 2 CPU cores and 8 GB of memory running around the clock for a month would cost around $30 after taking out the hyperscaler markup. This is likely what it costs (ballpark) for META to run the same compute stack for Muse. A single instance can host many VMs; we've seen estimates ranging from 5 to 100 VMs per physical instance. This makes sense as Muse only "connects" to a user's VM when it is working, and it sleeps like a human would when it's not in service. As we describe below in the architecture section, the VM is where the user's profile, credentials, and memory are stored to make the agent more helpful over time.

If we assume META is running 8 users per instance, the cost per DAU for the VM would come in just below $4 per month, illustrated below.

FIGURE 3. **VM Cost per User Comes Down via Oversubscription**

Virtual Machine Cost	1 User	8 Users	24 Users	100 Users
Cloud Hourly Revenue / User	$0.062	$0.008	$0.003	$0.001
# Concurrent Users	*1*	*8*	*24*	*100*
Instance Name	*AWS m8a.large*	*AWS m8a.large*	*AWS m8a.large*	*AWS m8a.large*
Cloud Hourly Pricing	*$0.062*	*$0.062*	*$0.062*	*$0.062*
Instance Commitment	*3 Years*	*3 Years*	*3 Years*	*3 Years*
CPU Type	*AMD EPYC 5*	*AMD EPYC 5*	*AMD EPYC 5*	*AMD EPYC 5*
vCPU	*2*	*2*	*2*	*2*
Memory (GB)	*8*	*8*	*8*	*8*
Monthly VM Cost / User	**$29.29**	**$3.66**	**$1.22**	**$0.29**
% Cloud Margin	*35%*	*35%*	*35%*	*35%*

Source: Barclays Research, Company Reports

The other portion of compute required for persistent agents is the input/output tokens per user request. Free users on ChatGPT send around 7 queries per day[1], which costs around $0.41 per month to serve (using '26 inference cost per WAU and backing out the paying subs), illustrated below.

Tasks that involve using the browser, or navigating multiple sites with multiple tool calls, tend to consume far greater volumes of tokens than those via connectors. So the query mix can impact token volume per request over time, likely with a wide range between power and casual users. The query volume for a persistent agent service like Muse is likely higher than the free users on ChatGPT, but for purposes of this report we are using similar inference costs per user to calculate total costs of around $4 per user per month.

FIGURE 4. **Compute Costs for Persistent Agents Likely Higher than Chatbots**

Total AI Usage Costs	ChatGPT - Avg.	ChatGPT - Paid	ChatGPT - Free	Muse - Free
Chatbot Inference Cost	$9.1B	$4.5B	$4.6B	
Total OpenAI Inference Cost	*$16.0B*			
ChatGPT Inference % of Total	*57%*			
Total Subscription Revenue		*$15.1B*		
Chatbot Subscription % Margin		*70%*		
Monthly Inference Cost / User	$0.76	$4.72	$0.41	$0.41
Avg. WAUs	*1,000m*	*80m*	*920m*	
% of Total WAUs	*100%*	*8%*	*92%*	
Total Monthly Cost / User	**$0.76**	**$4.72**	**$0.41**	**$4.07**
Virtual Machine & Other	*$0.00*	*$0.00*	*$0.00*	*$3.66*

Source: Barclays Research, Company Reports

Muse likely has slightly lower margins in the steady state future when it's monetizing at full, compared to META's social apps (Facebook and Instagram). This makes sense as compute costs for traditional mobile apps are lower than something like Muse for obvious reasons. We do think there is a reasonable chance that Muse monetizes at a higher rate per user than Instagram when it's fully developed, as a personal assistant would have far greater intent than lean back entertainment use cases.

[1] OpenAI CFO Sarah Friar: IPO, AI Rivalries, New Device, and Spending $100B+ on Compute

FIGURE 5. **Hypothetical Steady State Monthly Unit Economics for Social Apps vs. Muse (US)**

Future Agentic Unit Econ (US Only)	FoA	Muse
Monthly Revenue per User	$25.64	$25.64
Less: OpEx per User	($11.88)	($14.02)
Depreciation/Compute Cost	($1.94)	($4.07)
Other OpEx	($9.95)	($9.95)
Monthly OI per User	$13.76	$11.62
% Margin	*54%*	*45%*

Peak FoA margin 2024; other FoA metrics 2026E
Source: Barclays Research, Company Reports

The Architectures We've Seen Thus Far

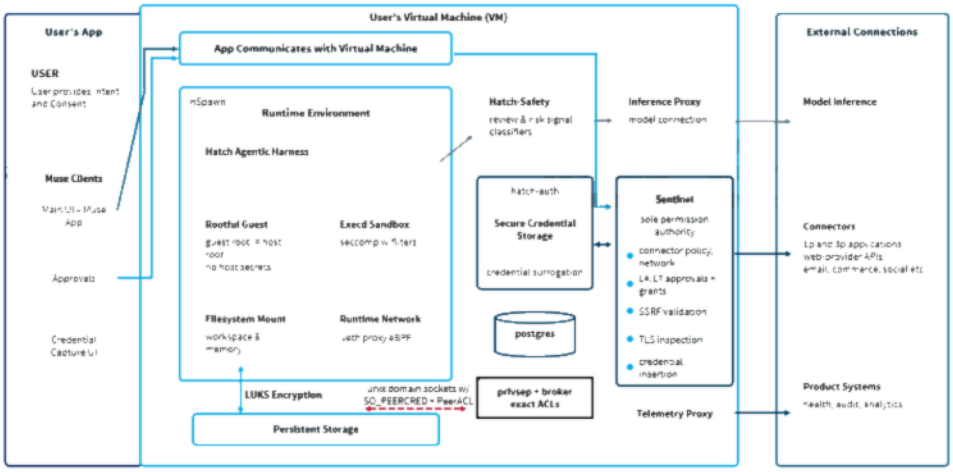
Noted above, the winning architecture for persistent agents will involve a combination of features:

1. Connectors
2. Trust/Security (Sentinel, encryption)
3. The model and the virtual machine (VM)
4. Agent/sub-agent architecture running around the clock, and agents growing more personalized over time

This architecture is far different than what was experienced previously in the consumer chatbot era where context and persistent memory was limited and inference was handled by a single model running on GPUs.

Persistent agents live on their own device, in this case a virtual machine (in the case of Muse's predecessors like OpenClaw, a dedicated Mac Mini). This architecture allows agents and sub-agents to orchestrate and execute tasks autonomously through a hybrid of token generation, computer use and tool calling. The VM runs the agentic harness, centralizes context and credentials inside a sandbox, and connects out to model inference and the end applications. "Connectors" are APIs that integrate downstream applications into the agentic flow natively. Below we illustrate how META built Muse's architecture in Figure 6.

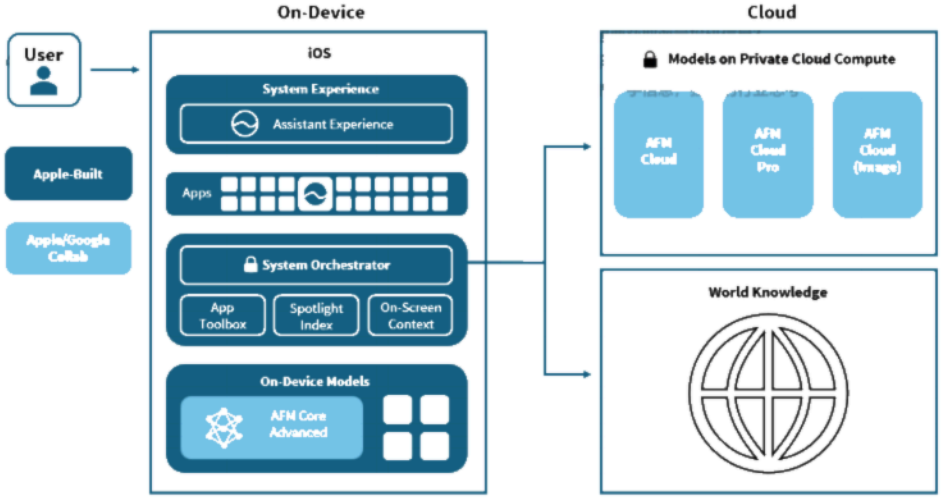
FIGURE 6. **META's Muse Architecture**



Source: Barclays Research, Company Reports

Apple Intelligence (link) took a different approach to Siri AI's architecture. Apple moved some of the inference on-device running locally on its next-gen A20 chipset. This is likely only capable of handling more basic AI tasks like text completion and image-gen; anything requiring a larger model or involving tool calling needs to happen in the cloud in the company's Private Cloud Compute (PCC). On-device has its advantages in terms of safe local storage of user credentials, a point we expect Apple to emphasize when its service is more broadly available.

FIGURE 7. **Apple's Siri AI Supports Agentic Functions but Isn't Fully Persistent in Current Form**



Source: Barclays Research, Company Reports

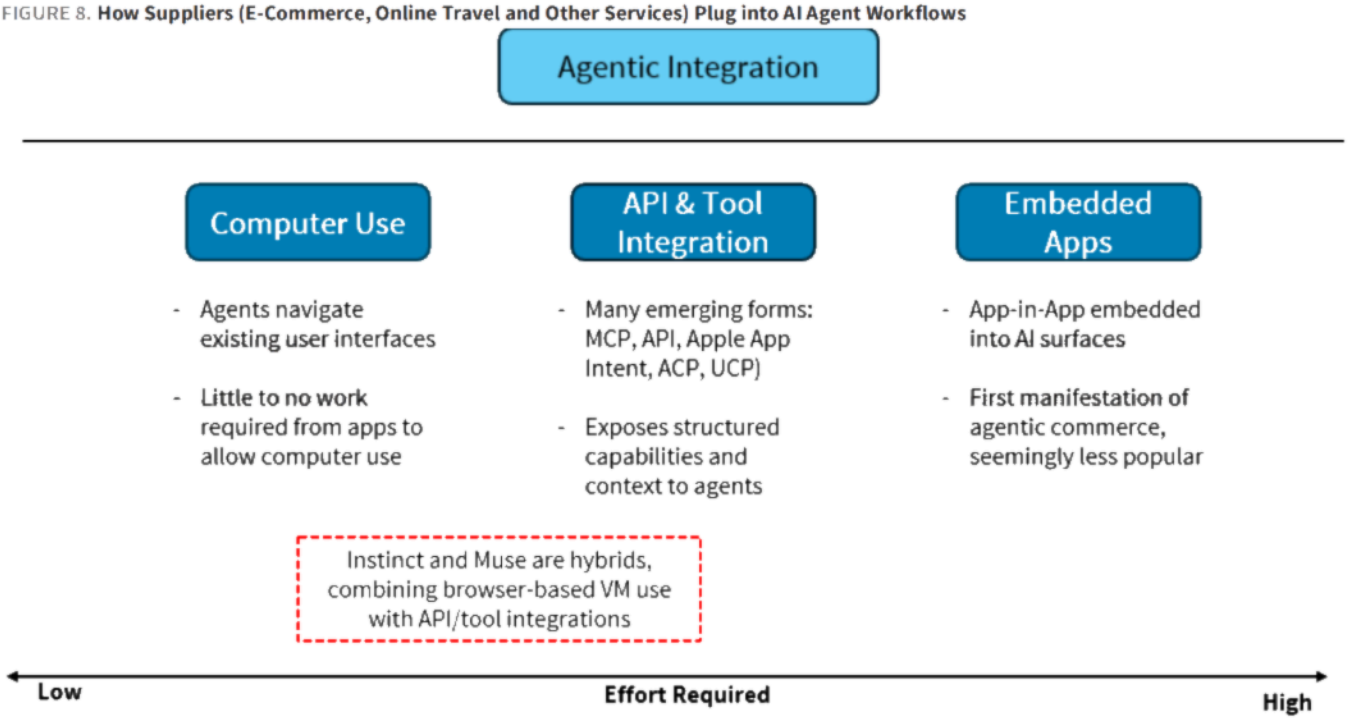
How Apps Connect to Consumer Agents

Third-party applications that handle transactions (like DoorDash or Shopify) can integrate with consumer agents in many ways, including API, computer/browser use, tool integrations, and embedded app-in-app experiences. These require varying levels of effort, but can remove significant friction (and token spend) for agents. For example, an agent could navigate each step of Vrbo's user interface inside a browser, or EXPE could expose the listings via direct API.

Even if a third-party application or website isn't directly integrated, the agent can access most services through computer use on a built-in browser inside the VM. There are downsides, as agents can clog up web and app servers, and degrade the user experience for other human traffic. For example, a recent Resy user was banned after his Instinct hammered ~200 requests/hr to book a table at 4 Charles Prime Rib (link). This method is useful when there are no direct integration alternatives, but over time we think connections via API/MCP data pipelines are likely to expose third party services to agents.

Shopify has recently integrated via this route with Muse, exposing billions of listings in an AI-friendly Catalog. Some applications go a step further and build bespoke experiences inside AI applications via SDKs, as seen early on with ChatGPT and various merchants.

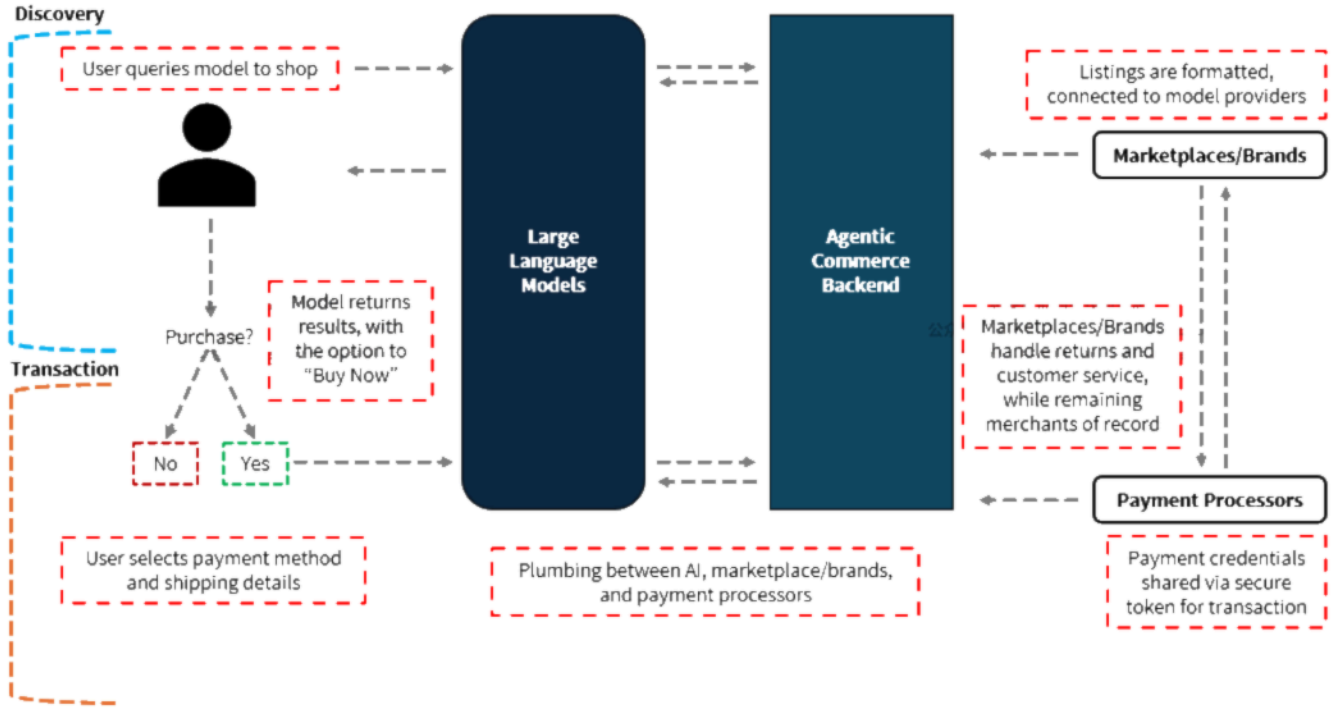
FIGURE 8. **How Suppliers (E-Commerce, Online Travel and Other Services) Plug into AI Agent Workflows**



Source: Barclays Research, Company Reports

Google/Shopify's Universal Commerce Protocol (UCP) and Stripe/OpenAI's Agentic Commerce Protocol (ACP) are examples of API and tool integrations, and are illustrated below in Figure 9. The end-user interacts primarily with the persistent agent in the native AI app (ie., Muse), and the bulk of post-transaction actions fall on the merchant or its platform (i.e., Shopify). Where the transaction happens is less consequential (natively, in a browser iFrame, via Stripe Link, etc.); the goal is to reduce the amount of steps and friction for both the agent and the merchant.

FIGURE 9. **E-Commerce Protocols Being Embraced By Persistent Agents & Merchants**



Source: Barclays Research, Company Reports

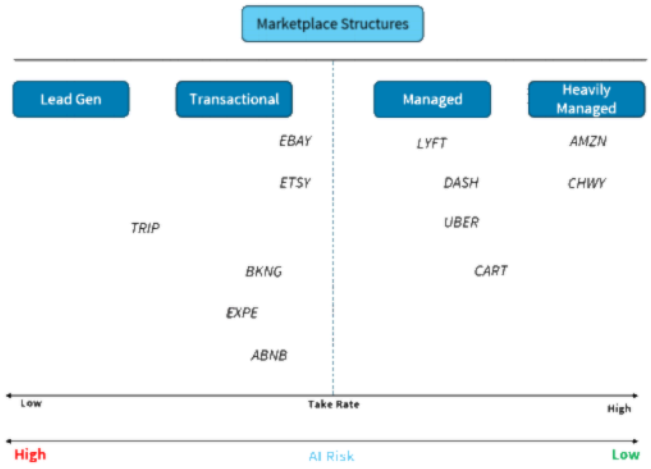
How This Changes the Rules of Engagement

If persistent consumer agents gain meaningful traction, we'd expect further turmoil in the marketplace layer as discovery and transactions reroute, renewing disintermediation concerns across consumer internet. Noted above, there is a reason why 50% of Instinct's $1B GMV run-rate is travel bookings - consumers are likely seeking alternatives to some of the drudgery involved in planning and booking travel, despite OTA commentary to the contrary. We expect suppliers to adapt, and to the degree that persistent AI agents end up being 1B+ Western market users, those suppliers who resist integrating will likely find themselves losing market share to those who embrace.

The logical question is which suppliers are most at risk and which are likely to preserve their unit economics in this future state where AI is conducting discovery and fulfillment of transactions. At a high level, the more painful it is to fulfill a transaction, the more defensible a supplier's business model likely is. Painful could mean heavy use of real world physical assets to complete a transaction successfully (costly warehouses, drivers, etc.). The less painful, the more likely there are multiple suppliers already set up to facilitate said transactions (lead generators and asset light marketplaces).

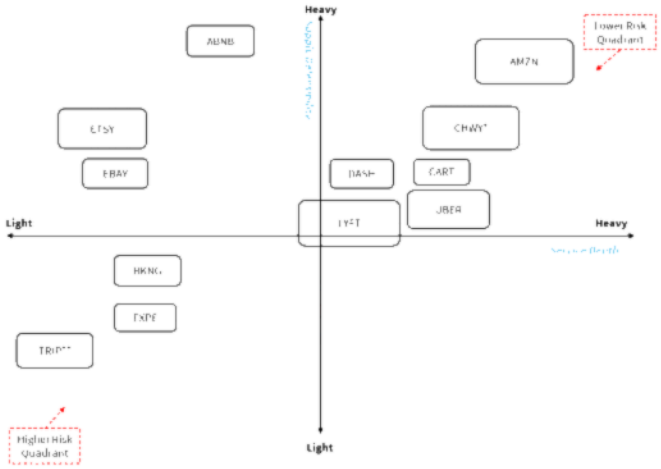
We look at our companies through the lens of how much work does the marketplace do vs. how much do the sellers do. We also look at those who serve as demand generation only, vs. integrated in the transaction process.

FIGURE 10. **Depth of Services Matters....**



Source: Barclays Research, Company Reports

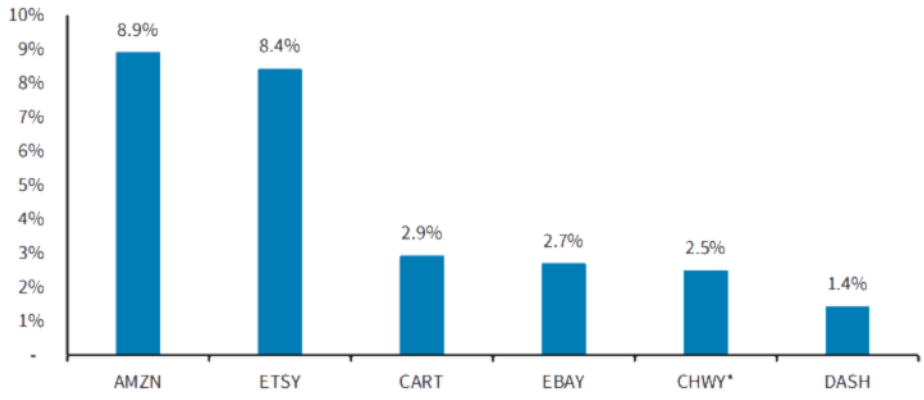
FIGURE 11. **.... As Does Supply Defensibility in Defining Disintermediation Risk**



Boxes sized by take rate, *CHWY Gross Margin, **Experiences Take Rate
Source: Barclays Research, Company Reports

Another dynamic to consider is where the business model generates most of its operating profit: 1) via fulfilling transactions, or 2) via on site advertising (ads as % of GMV). Agents not only create a race to the bottom on consumer prices, but take away ad revenue which often times represents the bulk of operating profit in e-commerce. Heavily managed marketplaces will be in a position to adjust their fee structure to make up for lost ad revenue. But those who don't have leverage on their supplier fees are likely to lose out.

FIGURE 12. **Ads Typically Represent a Large Percentage of Operating Income**



FY26E Advertising Revenue % of GMV, *Revenue for CHWY
Source: Barclays Research, Company Reports

It's no surprise that these marketplaces have come under pressure once again as persistent agents contest their ability to generate profit at the same level as in the past. Below we show the criteria investors should think about around defensibility of the supplier or marketplace business model plugging into persistent agents.

FIGURE 13. **Marketplace Disruption (Or Lack Thereof) Heatmap**

Company	Marketplace Structure				Supply			Demand		Advertising Revenue % of GMV	YTD Return
	Structure	Transaction Complexity	Purchase Frequency / Consideration	Price Point of Underlying Good	Scale	Unique Products?	Concentrated	Scale	Replicable		
AMZN	**Heavily Managed**	High	Weekly/Monthly	Less than $50	Large	No	No	Large	Strong habituation	9%	7%
UBER	Heavily Managed/Managed	High	Weekly/Monthly	Less than $100	Large	No	No	Large	Strong habituation		-15%
DASH	Heavily Managed/Managed	High	Weekly/Monthly	Less than $50	Large	No	No	Large	Strong habituation	**1%**	-17%
CHWY*	**Heavily Managed**	High	Monthly	Less than $100	Large	Mixed	Some Big Suppliers	Moderate	Scaled peers	3%	-45%
LYFT	Managed	High	Weekly/Monthly	Less than $50	Large	No	No	Moderate	Scaled peers		-22%
CART	Heavily Managed/Managed	High	Weekly	$50-125	Large	No	Major Grocers	Large	Strong habituation	3%	-3%
ETSY	Transaction/Managed	Moderate	3-4x per year	Less than $50	Moderate	Yes	No	Moderate	Unclear	8%	25%
ABNB	Transaction/Managed	Moderate	1-2x per Year	High	Large	Yes	No	Large	Unclear		16%
EBAY	Transaction/Managed	Moderate	3-4x per year	Varies	Moderate	Yes	No	Moderate	Unclear	3%	21%
BKNG	Transaction/Managed	Moderate	1-2x per Year	High	Large	Mixed	Hotel Chains, Airlines	Large	Unclear		-24%
EXPE	Transaction/Managed	Moderate	1-2x per Year	High	Large	Mixed	Hotel Chains, Airlines	Large	Unclear		-6%
TRIP	Lead Gen/Transaction	Low	1-2x per Year	High	Moderate	Mixed	Hotel Chains, Airlines	Low	Unclear		-42%
										NASDAQ:	**15%**

Priced as of 9/29/26 Close
Source: Barclays Research, Company Reports

