Microsoft

WHITE PAPER

Use Case Inference and Value Estimation

How Microsoft defines and infers economic units of work and the value agents contribute

Abstract

The shift to AI and agent-augmented work requires a new framework for measuring productive work done and the corresponding value contributed by the agent infrastructure. Such an understanding characterizes the expansion of work done when in conjunction with agentic systems and helps customers understand and maximize efficiency with usage-based billing. This paper documents the primary two components of such a framework. First, we provide an overview of how Microsoft defines and measures work done, including development and evaluation of a taxonomy of economic units of work. We provide examples of the taxonomy classes and show how our taxonomy maps to the O*NET taxonomy of work activities. Second, we show how we define value of use cases done. Our measures of value are assisted work hours and a corresponding dollar value estimate based on how much of a use case was performed by the agent. Usage reporting based on use cases and their corresponding cost and estimated value can be found in Microsoft Agent 365 Business Analytics.

1 Units of Work

We utilize a taxonomy of use cases to classify work done into a comprehensive, standardized format. Standardization allows for comparison over time and across organizations. Here we describe the taxonomy, including development and evaluation, as well as how our taxonomy maps to the O*NET taxonomy.

1.1 Microsoft Use Case Taxonomy

The Microsoft Use Case Taxonomy consists of 6,323 information work-relevant business process use cases, referred to as use cases in the taxonomy. Three use case levels sit within a broader five-level hierarchy: segment (3), domain (34), category (L1; 231), subcategory (L2; 1,052), and use case (L3; 6,323). Each use case has a single primary path through this hierarchy. For example, *Enterprise Functions* (3) → *Marketing & Brand* (domain) → *Content & Campaign Marketing* (L1) → *Content Strategy & Planning* (L2) → *Content Calendar* (L3). Table 2 at the end of Section 1 shows more examples.

Candidate use cases were derived from three complementary sources: public AI-assistant prompts such as WildChat, which provide topics, keywords, and work use cases; approximately 1 million public job postings available as of 2024; and O*NET occupation data describing the use cases performed in different occupation groups. Taxonomy construction then proceeds through four stages: 1) identify and de-duplicate use cases using the source topics, keywords, and work use cases; 2) place each use case according to its function; 3) separate the work being performed from the action used to perform it; and 4) merge look-alike entries before verification, returning fixes to the relevant source step.

Unlike other studies, such as the Anthropic Economic Index, which classifies conversations directly into O*NET task statements and inherits O*NET's occupational structure (Handa et al., 2025), our taxonomy is built bottom-up from observed AI-assisted work and job postings, with O*NET used only as a downstream mapping layer. The trade-off is future coverage: because O*NET spans the full economy, their approach already extends to all occupations as AI adoption reaches them, whereas our taxonomy is scoped to information work.

Note that placement into the taxonomy uses the industry spread of the supporting job-posting evidence. Use cases with broad cross-industry support are placed under *Enterprise Functions* segment; those concentrated in one industry are placed under a corresponding *Industry Vertical*. Finally, to ensure the taxonomy captures the *what* rather than the *how* of each use case, we use an "intent-swap" test to ensure the use case identity remains stable when only the requested action changes. For example, changing "summarize our vision statement" to "draft our vision statement" changes the intent but preserve the use case.

1.2 Use Case Taxonomy Evaluation

1.2.1 Use Case Taxonomy Structural Evaluation

Prior to the final pruning of the taxonomy, a structural evaluation examined how well categories could be distinguished using the probability distributions produced by JEV-1.13.0 over 1,000 synthetic agent workloads. The study evaluated a near-final version of the taxonomy, including 7,039 leaf categories and 34 domains. Covariance and normalized correlation matrices were used to identify categories whose predicted probabilities rose and fell together, providing a diagnostic of potential overlap within the hierarchy.

The analysis found strong separation at the broadest level: the two largest segments had a probability correlation of −0.89. At the domain level, the correlation matrix was sparse, while at the category level, correlated probability patterns were concentrated within parent groups. A descriptive screen identified 13 potentially entangled sibling pairs across the hierarchy. Qualifying pairs shared a parent, had a probability correlation of at least 0.50, were jointly scored on at least 20 workloads, and had positive probabilities for both categories on at least 10 workloads. These findings suggest that the overlap detected by this screen was predominantly local to related categories rather than distributed across unrelated branches.

Discrimination weakened with depth. Mean whole-path confidence, calculated by multiplying the selected branch probabilities through the hierarchy, declined from 82% at the segment level to 44% at the leaf level; the margin between the first- and second-ranked choices narrowed from 68% to 28%. Leaf-level diagnostics were supported by sufficient observations for only 565 of the 7,039 leaves, leaving most leaves unmeasured.

In summary, while our synthetic agent workloads only match a subset of taxonomy leaf nodes, the taxonomy generally shows good discrimination, with any non-discrimination localized in the same branch of the taxonomy.

1.2.2 Use Case Classification Evaluation Against Human Labeling

Human validation uses a two-pass annotation procedure to construct a reference set of use case labels. The annotation program includes vendor judges and internal volunteers, with a target reference set of 500 conversations made available through an internal Microsoft data program. Customer content is never seen by Microsoft employees nor reviewed by human annotators.

Human judges review the conversation text alongside two independently generated sets of proposed use cases: one from the production classifier (see Section 1.3) and one from an LLM judge. In the first pass, judges mark each proposed use case as correct or incorrect based on the work requested by the user. In the second pass, they correct rejected labels and search the taxonomy for use cases missed by both systems. This second pass is intended to expose omissions that would be hidden by evaluating only the classifier's own proposed labels. Where multiple judges review a conversation, a consensus rule determines the reference labels and inter-judge agreement is recorded.

The human reference is used to evaluate both the classifier and the LLM judge. The evaluation examines whether relevant use cases appear among the classifier's ranked candidates, either as

the top candidate (Hit@1) or in the top 5 (Hit@5). Results shown in Table 1 indicate whether the use case selected by the human annotator was deemed reasonably similar according to a separate judge LLM, where reasonableness approximates a cosine similarity score of 0.7 as determined by earlier testing using embedding similarity. We note that the taxonomy is large at the leaf node level (6,323 use cases), creating a very difficult labeling task for human annotators, who agreed about 60% of the time, and thus should not be interpreted strictly as ground truth. Reasonable exact use case matches were made to human labels about 67% of the time. When rolled up to higher levels of the taxonomy, domain matches to human labels were in the top 5 candidates 89% of the time.

	Hit@1	Hit@5
Use-case	66.90%	77.70%
Subcategory	69.70%	81.70%
Category	71.70%	82.90%
Domain	77.70%	89.20%

Table 1. Results of use case classification for Cowork compared to human labels. Reasonable labels were selected as from the 6,323 use cases as the top match about two-thirds of the time.

1.3 Use Case Inference

Production use case inference follows an approach in which every agent's work is classified into the same taxonomy and we adapt the approach we take depending on the nature of the agent. For conversational agents such as Cowork, we can observe additional context from multiple conversations and from supporting context such as the user role, while for other agents only telemetry emitted directly from session traces are observed. Thus, matching the inference technique to the signal available ensures that we infer the use case in a way that takes best advantage from how each agent works.

For Cowork, the production inference pipeline first processes conversation activity to infer free-form descriptions of the use cases being performed and associates each inferred use case with its supporting turns. Use cases are not constrained to conversation boundaries: a use case can draw on turns from multiple conversations, a conversation can contain multiple use cases, and a single turn can contribute to more than one use case.

To classify use cases, candidate taxonomy-style use cases are generated via LLM from the raw Cowork user text. Embeddings are retrieved for these candidate use cases and then ranked according to cosine similarity with embeddings from the taxonomy itself. A second LLM call is used to pick the best single use case from the top 5 candidates, and its corresponding entry in the taxonomy is selected as the final use case. This entire process, including the final use case classification, runs "eyes off", meaning no customer content is ever seen by Microsoft. In addition to the evaluation against human labeling detailed in Section 1.2.2, in evaluation with

13,570 synthetic prompts, the correct use case is identified in 85.4% of instances and falls in the top 8 candidates in 92% of instances.

The pipeline operates incrementally. The initial inference run processes the preceding seven days of activity. Subsequent runs occur approximately every 24 hours and process activity since the previous run, rather than using fixed calendar-day windows. Following a processing failure, the next run can cover a longer interval to catch up on unprocessed activity.

For other agents, inference operates on agent sessions, where a session ends after 30 minutes of inactivity. Request and response text, including tool results, is extracted from the session traces. To classify a session, each request and response is embedded and compared via cosine similarity against two reference embedding for each use case in the taxonomy, one built from example requests and one from example responses. The request and response similarities are summed, and the use case with the highest combined score is selected as the single use case for the session.

In evaluation with the same 13,570 synthetic prompts, using the request text alone, the correct use case is identified in 61.0% of instances and falls in the top 8 candidates in 90.9% of instances. The correct use case is thus usually retrieved but not always ranked first. To address this, a single LLM call that re-ranks the top candidates has been validated offline and is planned as a fallback for low-confidence sessions. In the same evaluation, this raises identification of the correct use case to between 88.0% and 90.5%, and to at least 97.4% within the top 8 candidates, depending on the model used.

1.2 Mapping to O*NET

For generalization to broader industry standard, we map our taxonomy to the O*NET taxonomy from the U.S. Department of Labor. We create normalized semantic embeddings for each use case description and each individual O*NET DWA title, then compute the cosine similarity between every description and the DWA embeddings, assigning the match as the highest-scoring DWA and its parent IWA. For an initial measure of confidence in the match we also compare cosine similarity with the next strongest candidate DWA that rolls up to a different IWA. The gain in cosine similar for the match over the next closest candidate IWA yields a 'score margin'. We flag matches where the score margin was less than 0.03 or the cosine similarity with the matched use case was less than 0.35.

The median cosine similarity between use cases and the matched IWA was 0.57 (mean = 0.57, indicating a largely normal distribution). 3,941 of the use cases were flagged per the above criteria. However, even in those cases, the median cosine similarity with the matched use cases was 0.55 (mean = 0.55), indicating that the level of the match was comparable to the overall set. In other words, the match was similarly strong, but there was at least one other candidate IWA that was nearly as strong a match.

To improve and validate the prompt used for mapping, we compared mapped results to those achieved through direct classification of Copilot conversations (Counts et al., 2026, and Tomlinson et al., 2025). Using this comparison, we iteratively improved the prompt used for the

LLM-based mapping, achieving 96% accuracy in mapped to directly classified IWA labels in our final prompt.

	Use Case 1	Use Case 2	Use Case 3
Segment	Industry Verticals	Enterprise Functions	Industry Verticals
Domain	Banking, Insurance, & Financial	Marketing & Brand	Professional & Consulting Services
L1	Banking Operations & Technology	Product & Field Marketing	Talent & Capability
L2	Reconciliation & Settlement Operations	Product Marketing & Launches	Knowledge & Methodology Management
L3	Position Reconciliation	Go-to-market Plans	Best Practice Codification
Description	Comparison of internal position records against custodian and depository holdings to confirm accurate securities and cash positions.	Coordinated plans for bringing a product to market, covering target audience, messaging, channels, pricing, and launch.	The formalization of proven practices into documented standards that consultants can consistently apply across engagements.
Mapped O*NET IWA	Reconcile financial data	Communicate with others about business strategies.	Coordinate activities with clients, agencies, or organizations.
Mapped O*NET DWA	Reconcile records of sales or other financial transactions.	Collaborate with others to develop or implement marketing strategies.	Collaborate with outside groups to develop programs or projects.

Table 2. Three example entries from the use case taxonomy, including mapping to O*NET IWA and DWA.

Estimating Value

For each use case instance, we estimate the labor value of the work done by the agent. Labor value is defined as what the work done by the agent would cost when done by a human. Here we provide an operationalization of our measure, including creation and validation of its components. Specifically, the monetary value of a use case performed by an agent is a function of four components, as shown in Figure 1: the hourly rate of occupations that perform that use case, the time required to complete the use case in its entirety, the maximum percent of the use case the agent is capable of contributing, and the use case completion rate. We describe each in turn.

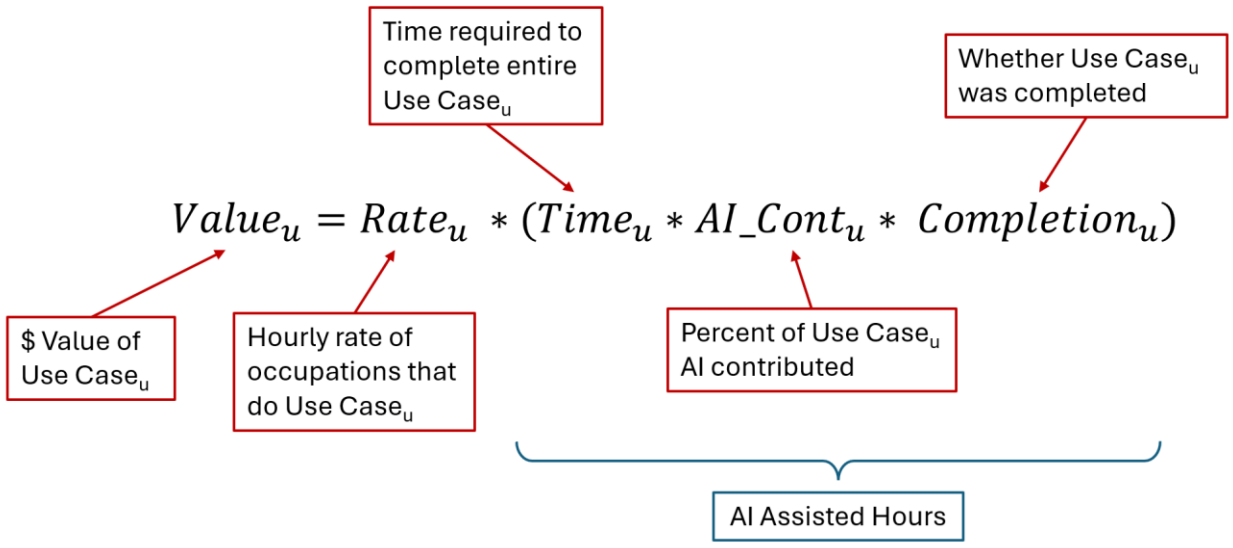


Figure 1. Use case value is a function of the hourly rate of occupations that do the use case multiplied by the number of AI-assisted hours, which are comprised of the time it would take a person to do the use case, the percent of the use case AI contributed, and whether Cowork completed the use case.

2.1 Rate

Rate refers to the average annual salary of the occupations that perform the mapped O*NET IWA, divided by 2,080 work-hours per year. For each use case, we utilize our mapping into the O*NET taxonomy and then traverse to occupations that perform the corresponding O*NET Intermediate Work Activity (IWA). The rate estimates range from $18.68 to $97.93 per hour, with a median of $46.07. All hourly rates and value estimates are in U.S. dollars, though rate information may be overridden with different monetary denominations, which would alter the value estimations.

2.2 Time

To estimate the time required for each of the 6,323 use cases, we developed a three-pronged methodology that involves LLM estimation and human validation (see Figure 2). Details are below, but in summary we conclude that (1) the model and human estimates largely correlated with each other, (2) human workers largely found model estimates to be reasonable, and (3) MAI-Thinking-1 model is the closest model in our experiment.

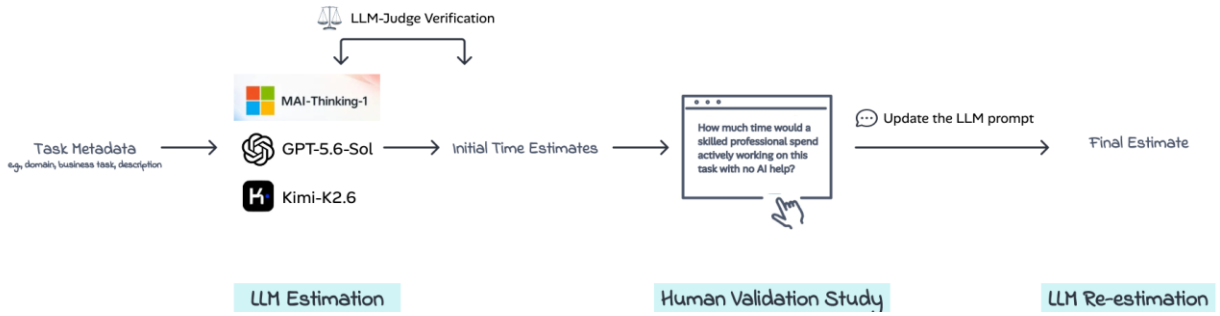


Figure 2: The three-pronged methodology. First, we use three LLMs (i.e., MAI-Thinking-1, GPT5.6-Sol, Kimi-K2.6) to estimate the time by generating 10 realistic scenarios for each use case. We validate each use case in with 1,000 human workers on Prolific and revise the prompt and generated the estimates with MAI.

2.2.1 LLM Estimation

We estimate the time taken for a person to complete the entire use case using an **LLM estimation**. For each of the 7,000 use cases, we synthetically generate 10 example scenarios and then use an LLM to estimate the time required for each of the 10, of which we then take the median value. Reporting the median allows us to reduce the sensitivity to outlier scenarios. This procedure was completed using three different LLMs (GPT 5.6, MAI, and Kimi). The estimates obtained from the three models correlate strongly (MAI-GPT at r = 0.73, Kimi-GPT at 0.77). For the final result we selected the MAI-based estimates, which tended to have the most conservative median estimate for each use case and narrower estimated time range.

2.2.2 Human Validation Study

To validate our jury LLM estimates, we conducted a human validation study and then iterated the LLM prompt used for use case time estimation incorporating the human estimates of total use case time for a subset of the 6,323 use cases. In a **human validation study** conducted on Prolific, we selected at random 1,000 of the 6,323 use cases and obtained three estimates per use case from human raters via the crowdsource platform Prolific. The use case time estimation prompt was then revised to bring the jury LLM estimates closer to the human estimates.

Across the 1,000 sampled use cases, at least one of three human raters judged MAI's estimate reasonable for **949 use cases (94.9%)**. MAI's estimated time range overlapped with at least one rater's final range for **939 use cases (93.9%)**, and overlapped with the use case-level median human range for **765 use cases (76.5%)**. The difference between MAI's median point estimate and the median final human estimate was 37.5 minutes absolute terms and **-17.3%** fractionally.

In other words, our use case time estimates were within a bit over half an hour of human estimates, typically falling below the human estimate by about 17%.

2.2.3 LLM Re-estimation

We ran an **LLM re-estimation** from user feedback and examples where model and human estimate were highly inconsistent. In this stage, we revised the prompt, added few-shot examples, and applied the LLM judge using MAI-Thinking-1 again. Examples of use case time estimates are provided in Table 3.

Domain	Use Case	Description	Estimates
Data, Analytics & AI	Hallucination & Safety	Evaluate risks from fabricated, misleading, or unsafe AI-generated content and identify guardrails and mitigations.	3 hr 45 min (2 hr 53 min–4 hr 38 min)
Data, Analytics & AI	Cross-Validation Strategies	Design validation schemes, such as k-fold and stratified splits, to assess model generalization and guard against overfitting.	1 hr 50 min (1 hr 25 min–2 hr 25 min)
Data, Analytics & AI	Data Mapping	Map source fields to target data structures, including the required transformations.	4 hr (3–5 hr)
Data, Analytics & AI	Management Reporting Packs	Prepare consolidated reports for management review of business performance.	2 hr 45 min (2 hr–3 hr 45 min)
Professional & Consulting Services	Resource Allocation	Allocate people and resources across delivery workstreams, balancing skills, availability, and competing demands.	2 hr 20 min (1 hr 50 min–3 hr 10 min)
Strategy, Planning & Governance	Initiative Prioritization	Rank competing initiatives by value, effort, and strategic fit to guide sequencing and resourcing.	6 hr 30 min (4 hr 30 min–8 hr 30 min)
Product & Design	Roadmap Tradeoffs	Balance competing initiatives, resources, and timing when making roadmap prioritization decisions.	4 hr 30 min (3–6 hr)
Product & Design	Sprint Collaboration	Coordinate product and engineering work during sprints, including standups, use case clarification, and iterative delivery.	2 hr (1 hr 35 min–2 hr 45 min)

Table 3. Examples of the MAI time estimates for business use cases

2.3 AI Contribution

AI and agentic systems may not be capable of completing the entirety of a use case. If a use case takes 10 hours in its entirety, but an agent is only capable of doing 60% of that use case, then the contribution from the agent should not exceed 6 hours. Thus we include a term *AI_Cont* to capture this ceiling. This term is determined by taking the lower of two estimates, one based on and validated empirically against a large-scale analysis of general Copilot usage, the other from an LLM estimate of AI capabilities with respect to each of the 6,323 use cases in the taxonomy. The empirically-validated estimates benchmark against real-world AI capability scores from a Microsoft platform, while the LLM estimates target the new use case taxonomy directly.

The empirically based capability ceiling is derived from Tomlinson et al. (2025) and Counts et al. (2026), each of which identified O*NET IWAs in samples of Copilot conversations. The assumption is that a sufficiently large sample of real-world usage reflects the capabilities of the AI platform. In order to generate a more granular equivalent to the IWA estimates in prior work, we created an AI Applicability estimate of each DWA in the O*NET taxonomy, following the procedure in Eloundou et al. (2024), which has been shown to correlate highly with these empirical AI Applicability scores (see Tomlinson et al., 2024). We validate against the Counts et al. IWA-level scores, finding that when our DWA-level estimates are aggregated to the IWA level, our scores correlate with the empirically derived scores at r = 0.73, while our median work activity-level estimate is 0.46 (on a 0 – 1 scale) compared to 0.44 for the empirically-derived scores.

The direct LLM estimates were obtained using the same MAI LLM. The correlation between the empirically validated and LLM-based estimates was moderate at 0.28, and we thus use the lower of the two for each of the use cases to err on the conservative side of AI capability.

2.4 Use Case Completion

As with use case inference, completion in production follows one of two paths depending on the nature of the agent. We start by showing how this was done for Cowork using a deterministic calculation based on the success (or failure) of the agent to respond to the final turn in a user session, followed by the LLM judge evaluation used for other agents.

2.4.1 Deterministic Calculation for Cowork

For each use case identified within a conversation, we select the last turn associated with that use case and determine whether it produced a final response. A successful final-turn execution receives a score of 1; an observed failure receives a score of 0. Intermediate-turn outcomes are not averaged, so an earlier unsuccessful attempt does not reduce the score when the final associated turn succeeds.

For a given use case or use case, use case completion rate is the arithmetic mean of these final-turn scores across associated conversations. Equivalently, when all outcomes are observed, it is the number of successful final associated turns divided by the number of evaluated conversation–use case instances. A conversation containing multiple use cases can contribute a

separate outcome to each use case; the last turn associated with a use case may therefore differ from the final turn of the conversation as a whole. Use case identification and turn assignment use an LLM, but completion scoring and aggregation are deterministic given those assignments and the execution records.

When an outcome cannot be computed because supporting-turn information is unavailable, we substitute a default score of 0.85. Such cases include conversations for which no use case was inferred and no use case-specific turn mapping is available. These values are imputed rather than observed successes or failures. Low-confidence use case classifications can still receive a measured completion score when their supporting-turn evidence is available. The resulting rate therefore reflects both observed execution outcomes and any default values included in the aggregation.

$$\mathrm{TCR}_u = \frac{\sum_{i \in O_u} s_i + 0.85|M_u|}{|O_u| + |M_u|}$$

Here, O_u is the set of conversation–use case instances with observed final-turn outcomes for use case u, s_i is 1 for successful execution and 0 for failure, and M_u is the set requiring the imputed score of 0.85. Each instance receives equal weight; the rate is defined when at least one instance is included.

2.4.2 Validation Results

We conducted an initial consistency check over October 26–28, 2026, comparing use case-level completion rates in an internal organizational sample with Cowork's session-level last-turn success rate across the broader internal deployment. Mean use case completion rates were 95.3%, 94.2%, and 93.5%, respectively.

2.4.3 Interpretation

The measure captures execution reliability rather than directly assessing whether an output fully satisfies the user's intent. A score of 1 indicates successful final-turn execution, not that the agent completed every component of the underlying business use case or produced an output of sufficient quality. In the value calculation, AI_Cont separately limits the share of the use case that AI is capable of performing; use case completion rate accounts for execution outcomes within that framework.

Further validation requires comparisons on the same conversations and use case assignments, separate reporting of observed and imputed outcomes, and human assessment of whether the intended work was accomplished. LLM-based assessments of use case fulfillment may provide a complementary signal but should not be treated as ground truth without validation against human judgments.

2.4.4 LLM Judge Evaluation for Other Agents

For other agents, we estimate use case completion with an LLM judge applied to a sample of sessions. The judge reviews the entire session, including all turns and tool calls, and returns a pass or a fail. A pass receives a score of 1 and a fail receives a score of 0. To manage cost, a substantial portion of the sessions of each agent are sampled for judging. Once an agent has reached a critical threshold of judged sessions within a rolling 30-day window, the average of its judged scores is imputed to its remaining sessions. Imputed scores are thus values between 0 and 1 rather than observed outcomes, and no completion score or value is reported for an agent below this threshold. Because each session carries a single use case, the completion score enters the value calculation at the session level.

We validated this approach in two ways. First, we compared the judge against human review of 172 production sessions. The judge passed 91.6% of the sessions that the reviewer marked complete (recall), with a precision of 77.2% and an overall accuracy of 77.9%. Second, we compared imputed scores with the scores later assigned by the judge to the same agents. Across 68,090 paired evaluations from 935 agents, the session-weighted mean absolute error was 2.97 percentage points, with error concentrated in low-volume agents.

3. Summary and Conclusion

Understanding the impact of AI and agentic systems requires more than counting conversations, turns, or model usage. It requires understanding the work people are undertaking with these systems, the share of that work the systems contribute, and the value of that contribution. As agents take on increasingly substantial parts of business workflows, these measures become important for identifying where AI expands productive capacity, comparing use across organizations and over time, and helping customers assess the value they receive relative to usage-based costs. Here we outlined two complementary foundations for that understanding: a standardized taxonomy of work and a transparent framework for estimating the labor value of agent contributions.

The Microsoft Use Case Taxonomy organizes 6,323 information-work business use cases into a three-level hierarchy aligned with business domains. Structural evaluation of the taxonomy found strong broad-level separation and predominantly local overlap among related categories, while also showing weaker discrimination at greater depth and limited coverage of leaf categories. The taxonomy is detailed and tailored to modern information work. We thus support generalization to industry standard by mapping to the O*NET work activity taxonomy, with a final mapping achieving 96% agreement with directly classified IWA labels. Finally, during inference for Cowork we associate use cases with their supporting conversation turns to generate a consistent representation of work not restricted to conversation boundaries. Inference in production follows one of two paths depending on the nature of the agent: this conversation-based approach for Cowork, and classification of each session from agent traces for other agents.

The value framework translates classified work into assisted work hours and their corresponding dollar value using four components: estimated unaided human use case time, the hourly rate of

relevant occupations, an AI contribution ceiling, and a use case completion measure. The wage component is anchored in occupations associated with the mapped O*NET activity. Use case-time estimates combine realistic synthetic scenarios, model estimation, and human feedback. In validation on 1,000 sampled use cases, at least one of three raters found the MAI estimate reasonable for 94.9% of use cases, and the model range overlapped the median human range for 76.5% of use cases.

The AI_Cont term uses the lower of an estimate validated against a large sample of Copilot usage and an LLM estimate to constrain the credited contribution conservatively; their moderate correlation of 0.28 indicates that they are not interchangeable measures. For Cowork, the completion component is grounded in final-turn execution records, with initial mean use case completion rates ranging from 93.5% to 95.3% across Microsoft-internal Cowork usage samples. Successful completion does not establish that the intended business work was fully accomplished or that its quality was sufficient, and missing outcomes receive an imputed score of 85%. Rather, the completion term combined with the AI_Cont term give a more holistic picture of the extent to which the desired use case outcome was achieved, albeit still an estimate based on potentially assistable work rather than a direct measurement of realized human time savings. For other agents, completion is instead assessed by an LLM judge on a sample of sessions and imputed for the remainder, with the judge passing 91.6% of the sessions that human review marked complete.

Taken together, the taxonomy and valuation framework provide an interpretable basis for connecting AI activity to recognizable work and its estimated labor-equivalent value. The evidence supports their use for characterizing patterns of work and contribution, while identifying clear priorities for continued validation of classification, use case scope, assistable effort, and successful fulfillment. The resulting dollar values should be understood as estimates of contributed labor value, not as cash savings, net return on investment, or the entirety of the benefits AI may create. Used with these distinctions, the framework can help customers and product teams move from measuring AI usage to understanding what work AI enables, where it contributes most, and how that contribution evolves as agentic capabilities and workflows mature.

Finally, we emphasize that customers know their own work processes best, and the framework is designed so that customers can override our default estimates for rate and estimated use case time, tailoring the valuation to their own compensation structures, workflows, and typical use case scope. This flexibility lets each organization fine-tune the estimates so that the resulting assisted work hours and dollar values more closely reflect the realities of its business. Note that customized inputs will naturally shift the estimates away from the defaults described here, and the validation results reported in this paper apply to those defaults.

References

Counts, S., Chen, Y., Dong, J., Sharma, H., Zaikin, A., Hu, R., Kok, A., Yilmaz, G. O., Suri, S., Tomlinson, K., Jaffe, S., & Wang, W. (2025). AI in the Enterprise: How People Use M365 Copilot Chat. arXiv preprint arXiv:2605.23958. doi:10.48550/arXiv.2605.23958

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science, 384*(6702), 1306–1308. doi:10.1126/science.adj0998.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., & Ganguli, D. (2025). Which economic tasks are performed with AI? Evidence from millions of Claude conversations. arXiv preprint arXiv:2503.04761.

Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). *Working with AI: Measuring the occupational implications of generative AI*. arXiv preprint arXiv:2507.07935. doi:10.48550/arXiv.2507.07935.