HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2026 No. 036

An Artificial Intelligence-Based Process for Content Tagging References in the O*NET Reference List

Prepared for: National Center for O*NET Development
313 Chapanoke Road, Suite 130
Raleigh, NC 27603

Prepared under: HumRRO Subcontract:
1-312-0207142-41224L
RTI Prime Contract O*NET 98-01
(HumRRO Contract ID S 07-08)

Authors: Alexander S. McKay, HumRRO
Dan J. Putka, HumRRO
Phil Lewis, National Center for O*NET Development

Date: April 17, 2026

An Artificial Intelligence-Based Process for Content Tagging References in the O*NET Reference List

Table of Contents

Introduction 1

The Current Project 1

Step 1: Extract Abstracts for O*NET Reference List Publications 3

Step 2: Informational Interviews with Human Resources Research Organization (HumRRO) Analysts & Refining Content Category Definitions 7

Step 3: Develop LLM Prompt Templates for Content Tagging 12

Step 4: Draw Sample of Publications and Run Them Through LLM Content Tagging Process 17

Step 5: Collect HumRRO SME Judgments 19

- Level of LLM-SME Agreement by Abstract Word Count 24
- Summary 27

Step 7: Finalize Prompts and Run All Current Reference List Publications Through LLM Content Tagging Process 28

- Summary 37

Proposed Content Tagging Process for New Publications and Recommendations for Future Use 38

- Additional Considerations for Future Use 39
  - Abstract Extraction Requires Meaningful Time and Planning 40
  - Data Completeness Meaningfully Affects Tagging Quality 40
  - Revisiting the Tagging Process 40

Conclusion 41

References 42

Appendix A: Summary of Success Rates Across Publication Types and Current Content Categories 43

Appendix B: Final Prompt Template for LLM Content Tagging 44

Appendix C: Instructions for Human SMEs 48

Table of Contents (Continued)

List of Tables

Table 1.1. Summary of Abstract Extraction Results by Method 6
Table 2.1. Current and Revised Content Category Definitions 9
Table 2.1. (Continued) 10
Table 2.1. (Continued) 11
Table 3.1. List of LLMs Examined and Usage 13
Table 3.2. Percentage of Publications for Which There Was Full Agreement on Category Assignments across Runs within LLMs and Across Models for LLM Ensembles 15
Table 3.3. Percentage of Publications for Which There Was Agreement between LLM-Based and Ground Truth Ratings 16
Table 4.1. Percentage of Publications for Which There Was Full Agreement on Category Assignments Across Runs of Sonnet 4.5 (N = 180) 18
Table 4.2. Total Yes and No Counts and Percentages Based on Final Majority Rule Data 18
Table 5.1. Percentage of Publications for Which There Was Full Agreement on Category Assignments Across the Four HumRRO SMEs 19
Table 5.2. Total Yes and No Counts and Percentages Based on Final Majority Rule Data 20
Table 6.1. Agreement Between the HumRRO SMEs and LLM Tags 21
Table 6.2. Human vs. LLM Confusion Matrix 22
Table 6.3. Human vs. LLM Yes/No Tag Counts 23
Table 6.4. Number of Tags Assigned to Publications 24
Table 6.5. Agreement Between the HumRRO SMEs and LLM Tags Based on Word Count 26
Table 7.1. Percentage of Publications for Which There Was Full Agreement on Category Assignments Across Runs of Sonnet 4.5 (N = 2,850) 28
Table 7.2. Total Yes and No Counts and Percentages Based on Final Majority Rule Data 29
Table 7.3. Top Five Most Frequent Category Combinations 30
Table 7.4. Co-Occurrence of Categories 32
Table 7.5. Total Yes Counts and Percentages Based on Data Availability 33
Table 7.6. Number of Category Combinations Assigned to Publications Based on Data Availability 34
Table 7.7. Inclusion and Exclusion of Current Content Category in New LLM Tags 36
Table A.1. Summary of Abstract Extraction Results by Publication Type 43
Table A.2. Summary of Abstract Extraction Results by Current Content Categories 43
Table B.1. Variable Input Fields 47

List of Figures

Figure 1.1. Flow Chart of Abstract Extraction 5
Figure B.1. Final Prompt Template for LLM Content Tagging 44
Figure C.1. Instructions for Human SME Tagging 48

An Artificial Intelligence-Based Process for Content Tagging References in the O*NET Reference List

Introduction

The Occupational Information Network (O*NET) is a comprehensive system developed by the U.S. Department of Labor (DOL) that provides information for approximately 900 occupations encompassing over 56,000 jobs within the U.S. economy. This information is maintained in a comprehensive database (National Center for O*NET Development, 2026). To keep the database current, the National Center for O*NET Development (hereinafter referred to as "the Center") conducts ongoing data collection to identify and maintain current information on the characteristics of workers and occupations. As part of its online Resource Center, O*NET maintains and regularly updates a list of research and product-oriented publications that illustrate how widely O*NET data, websites, and tools are used. As of January 2026, over 2,800 references were included in the list. Each reference is currently organized or "tagged" by analysts into one of nine content category areas (e.g., Labor Market Research, Career and Vocational Research, Application Development**).** This report summarizes efforts to effectively leverage artificial intelligence by developing, evaluating, and implementing a Large Language Model (LLM)-based approach to automating the assignment of content tags to references in the O*NET Reference List. In addition, importantly, a new content category is added: "Artificial Intelligence." The updated process also now supports the potential for references to have multiple content area tags.

The Current Project

The objective of this task was to develop and evaluate an LLM-based process (simple prompting) for content tagging references in the O*NET Reference List and introduce a new content category explicitly designed to tag references that involve the use of O*NET in conjunction with Artificial Intelligence (AI), Natural Language Processing (NLP), or Machine Learning (ML) methods or technology. The process we developed will streamline the tagging of references for O*NET, bringing greater efficiency and consistency to the current, solely human-driven process.

To put the current project in context, it is helpful to understand how content categories are currently assigned to references in O*NET's Reference List. Currently, a team conducts a literature search for publications that use O*NET information, data, or tools or customers (e.g., researchers or application developers) submit to the Center potential reference additions. Once relevant publications are identified, the team divides them for further evaluation. Each publication is evaluated by one analyst to determine its relevance to O*NET and categorized into a single content area based on its content. A review of the current process reveals a few potential areas for improvement in assigning content categories to references. First, use of a single analyst may potentially lead to idiosyncratic categorization. Second, publications are classified into a single content area even when multiple areas may be relevant. Forcing a singular classification decision can exacerbate errors in the assigned content area when more than one area applies to a given publication. Nonetheless, we recognize that having human reviewers complete additional verification checks and tagging all publications into multiple categories would be an arduous task that would increase the time and costs to complete the entire task. In light of these observations, our focus for this project was to explore how the content categorization process can be automated using LLMs.

The goal of the alternative process that we propose is to use an LLM to evaluate the publication's reference (which includes author, title, publication year, source, URL), as well as the publication's abstract (should one be available), and determine what content tags apply to it. Thus, this alternative approach shifts the process from "content classification" to "content tagging," where multiple tags can apply to a given publication rather than being assigned to the "best" fitting content area. Further, it is possible to use multiple LLMs (e.g., Claude and Llama models) and/or multiple LLM runs (i.e., basing tagging judgments by multiple runs of the same LLM) in the tagging process, such that each publication is evaluated multiple times to identify idiosyncrasies in judgments resulting from any single LLM or LLM run. This alternative approach offers greater flexibility in the reference tagging process, overcomes limitations of the current human-based process, and likely does so at a fraction of the time and cost.

We split this report up into the seven steps that we used to develop the new process:

- Step 1: Extract Abstracts for O*NET Reference List Publications
- Step 2: Conduct Informational Interviews with the Human Resources Research Organization (HumRRO) Analysts & Refining Content Category Definitions
- Step 3: Develop LLM Prompt Templates for Content Tagging
- Step 4: Draw Sample of Publications and Run Publications Through LLM Content Tagging Process
- Step 5: Collect HumRRO Subject Matter Expert (SME) Judgments
- Step 6: Evaluate LLM Judgments Against HumRRO SME Judgments
- Step 7: Finalize Prompts and Run All Current Reference List Publications Through LLM Content Tagging Process

We provide details on each step in the sections that follow. Upon detailing these steps, we close the report with a summary of the proposed content tagging process and recommendations for future use as publications are added to the O*NET Reference List.

Step 1: Extract Abstracts for O*NET Reference List Publications

For this step, we first attempted to acquire abstracts or brief summaries of each publication in the latest available internal project or working version of the O*NET Reference List. The availability of reference summary information was viewed as a key component to the success of the new automated approach to content tagging references. Neither the working version of the O*NET Reference List nor the published O*NET Reference List on the O*NET Resource Center contains an abstract or summary information for each publication in the list. Across the two versions of the list, the following fields are in common: (1) Reference (which contains the author names, title, and source); (2) Publication Type (e.g., Research and Technical Reports, Journal Articles, Blog Posts, Books and Book Chapters); (3) Publication Year; (4) URL (associated with the reference); and (4) Content Category. In light of the omission of abstracts, we developed a simple script to extract abstracts or summaries from publications. The script was created to "pull" abstracts or summaries for publications using the URL when available.

We started with the most current internal project version of the O*NET Reference List (January 2, 2026). The file contained 2,889 references. We removed 38 entries flagged as "Duplicate" in an administrative column in the file named "Comments." One additional unflagged duplicate entry was removed.[1] The final dataset included 2,850 references, 141 of which lacked a URL (approximately 5%). The majority of the references without URLs were published from 1999 to 2010. We attempted to obtain abstracts or summaries for the remaining 2,709 publications.

We then developed R code to pull abstracts or summaries. The code used a systematic process to scrape external research databases and the internet. This process is summarized in Figure 1.1. It starts by checking each URL for specific information that guides the search to the source most likely to have an available abstract or summary. We describe the process in detail below.

First, the URL was checked for the following text strings: "arXiv," "SSRN," and "NBER." For arXiv, we used their available Application Programming Interface (API) to pull abstracts (arXiv, 2024).[2] For SSRN, we scraped the webpage associated with each reference using multiple Cascading Style Sheet (CSS) selectors. In other words, the webpage was checked in a few ways for a "div" (i.e., division) section that would contain an abstract. For NBER, we used their API for article metadata and, if that failed, the code scraped the webpage (National Bureau of Economic Research, 2026).

Second, when the URL included the text ".pdf," we used text extraction for the PDFs. This process worked by downloading the PDF and extracting the text via an internal document search for the keywords "Executive Summary," "Abstract," or "Summary." There was also additional code that attempted to avoid the table of contents (i.e., rules governing minimum character lengths).[3]

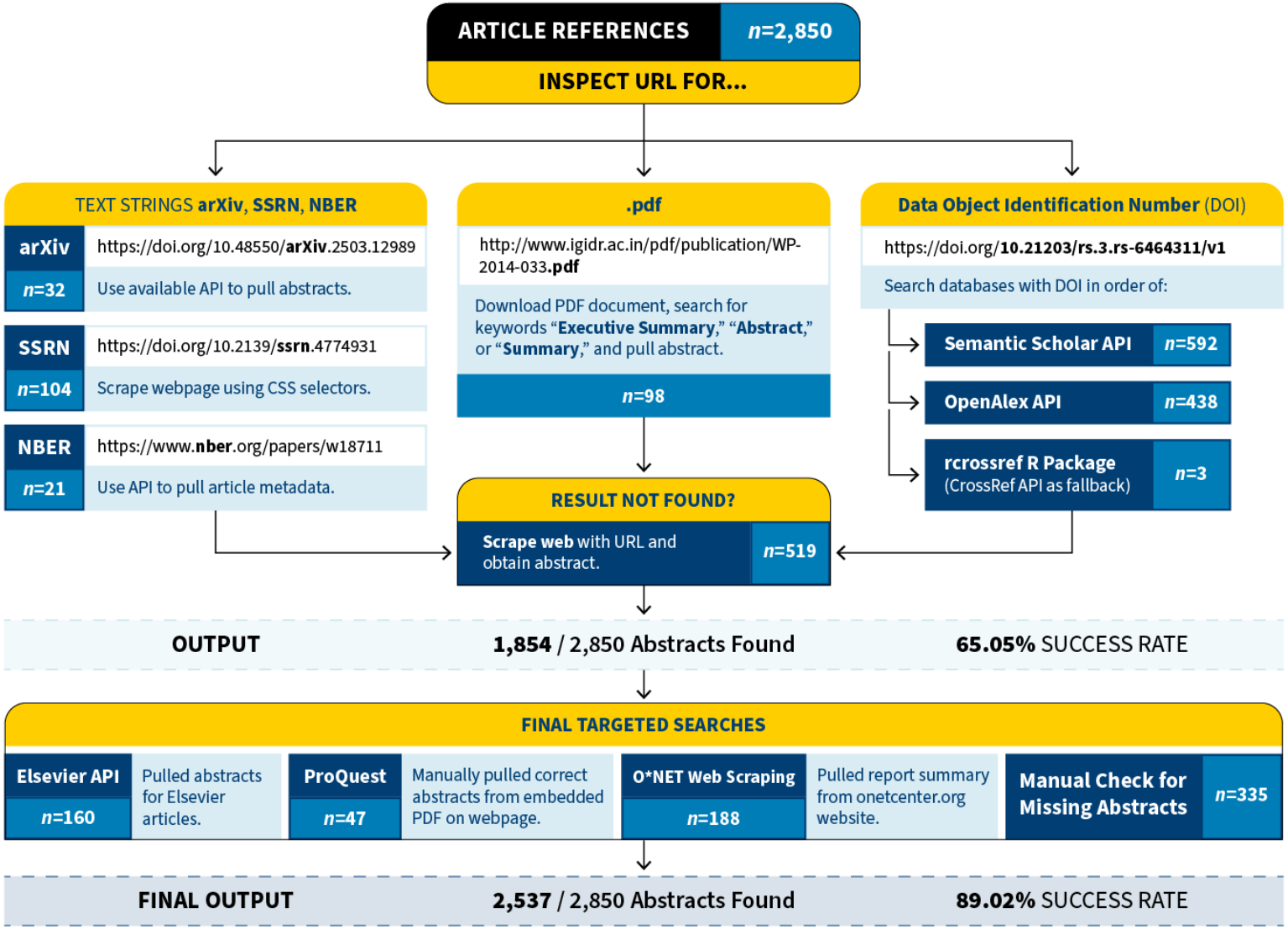
[1]This additional reference was included in the data as a pre-print article and a published journal article.
[2]ArXiv's API Access website requests that users of its API service offer the following acknowledgement when documenting use of their API – "Thank you to ArXiv for use of its open access interoperability" - thus we offer that acknowledgement here.
[3]Although the code successfully reduced the frequency of Table of Contents downloads for articles, multiple abstracts were still pulled with the Table of Contents included. We did not view this as problematic due to maximum character limits, such that both the Table of Contents and abstracts were pulled.

Third, when the URL included a DOI number, multiple research databases were searched for abstracts. The order of databases searched was as follows: Semantic Scholar API (Allen Institute for AI, 2024), OpenAlex API (Priem et al., 2022), rcrossref R package (Chamberlain et al., 2025), and CrossRef API (as a fallback to the rcrossref package; Crossref, 2024). Last, if none of the databases had an available abstract, the code attempted to web scrape one as a final fallback.

The code also included general processing checks to improve the quality of the results. First, the length of the identified abstracts was checked, and was expected to be more than 50 characters (to avoid the table of contents). Second, line breaks and excessive whitespace were removed. Third, abstracts and summaries from PDFs were trimmed to a maximum of 15,000 characters (about 5 single-spaced pages). This code was included to strike a balance between extracting too little text for longer report summaries and extracting too much text, leaving a large portion of the text beyond the summary section. We erred on the side of pulling more text than was likely necessary based on the notion that “more is better than less” for the tagging process. Last, we used polite scraping practices by incorporating delays between requests to avoid overloading servers.

The output included three new columns: (1) a column containing the abstracts, (2) whether the data pull was successful or failed, and (3) the source where the abstract was obtained or a reason if the code failed to pull an abstract (e.g., 404). This code successfully pulled abstracts for 1,854 of the total 2,850 publications (65.05% success rate).

Figure 1.1. Flow Chart of Abstract Extraction



Note. Numbers (*n*) indicate abstracts obtained at each stage of the process.

Upon inspecting the results, we supplemented these searches with four additional targeted searches. First, the O*NET Center used an Elsevier API to pull abstracts for Elsevier articles, as this publisher requires a developer account and API key (Elsevier, 2024). Second, for all 47 ProQuest URLs (i.e., dissertations), the wrong text was extracted from the webpage. These webpages embed the PDF, making text extraction more difficult. For these references, we manually pulled the abstract. Third, for O*NET Program publications with URLs containing "onetcenter.org," we used a targeted R script to scrape publication summaries directly from the O*NET Resource Center webpages. The script identified summary text using a specific CSS selector associated with the O*NET Resource Center's publication layout, successfully retrieving summaries for 188 publications. Last, we manually checked for abstracts for most of the remaining publication URLs. This process involved going to the URL in the database. If there was an abstract, it was copied and added to the final dataset. If the URL failed (e.g., 404 error), we conducted a Google search with the publication's reference to find a working URL. If a working URL was found, we updated the URL and tried to obtain an abstract.

After completing the full search, we successfully obtained abstracts for 2,537 abstracts (89.02% success rate out of 2,850). In Table 1.1, we summarize the success rates of obtaining abstracts based on each retrieval method. In Appendix A, we provide additional tables showing the success rate based on publication types and current content categories.

Table 1.1. Summary of Abstract Extraction Results by Method

Method	Found Count	Found Percent
arXiv	32	1.3
NBER	21	0.8
OpenAlex	438	17.3
PDF	98	3.9
rcrossref	3	0.1
Semantic Scholar	592	23.3
SSRN	104	4.1
Web Scraping	519	20.5
Elsevier API	160	6.3
O*NET Web Scraping	188	7.4
Manual	335	13.2
Manual ProQuest Search	47	1.9

Note. Total count sums to the 2,537 publications where abstracts were pulled. Percentages are based on the total count of abstracts obtained.

Step 2: Informational Interviews with Human Resources Research Organization (HumRRO) Analysts & Refining Content Category Definitions

While pulling abstracts in Step 1, we also completed Step 2 of this effort. We conducted short, informational interviews with four HumRRO analysts tasked with assigning content categories to entries in the O*NET Reference List. These interviews allowed us to capture lessons learned and nuances associated with the content categorization process. The insights we gained in these meetings helped us develop revised content category definitions and prompt templates that we used in subsequent steps.

During these meetings, we asked the analysts questions concerning strengths and limitations of the existing categorization process, decision policies used for categorizing publications, and resources used to inform categorization decisions. These questions were intended to help us understand the pain points experienced in the content categorization process, with the potential to inform revisions to content category definitions.

Although the insights gleaned helped inform definition revisions, they were not the primary consideration for revision. The existing definitions (labeled as “original” in Table 2.1) were designed to support a classification process where a publication was only assigned one content category, as opposed to a tagging process where more than one content tag could be assigned to a given publication. This shift has fundamental implications for updating the definitions, as elements of current definitions that were intended to narrow the scope of a content category to reduce overlap between categories and make them somewhat artificially mutually exclusive could be removed. For example, there is valid conceptual overlap between several sets of content categories (e.g., Industrial-Organizational Research, Career and Vocational Research, and Labor Market Research). Mutually exclusive distinctions between these content areas in revised definitions was no longer critical. Consequently, the updated tagging process would no longer force one-to-one classification assignments of publications to categories.

In light of the shift from classification to tagging and insights from analysts, we revised content category definitions to include a short sentence describing the category and a subset of keywords strongly related to that content area, which would be beneficial for an LLM-driven tagging process. This conclusion was also based on a review of documentation used in the categorization process, which included key descriptors used to categorize publications. Note, to simplify definitions, we also largely eliminated references to O*NET itself in several content categories (see the Focal Content Domain category definitions referenced below and in Table 2.1). Because all references in the list involved O*NET products, tools, or databases in some way (else they wouldn’t be in the list in the first place), references to O*NET could be limited from content-domain focused category definitions to streamline and focus their descriptions.

Beyond the changes described above, we also made some structural changes to the current list of content categories. First, at the request of the O*NET Center, we developed a new content category named “Artificial Intelligence” that focused on research using ML, NLP, and AI in conjunction with O*NET data, products, and/or tools. Showcasing the growing, extensive use of O*NET within efforts focusing on or using artificial intelligence was a key aspect of this project. Second, we put the definitions into two overarching domains. The first category, named “Focal Content Domain,” dealt with the research domain of the publication’s content and included six areas (see Table 2.1). The second category, named “Additional Publication Features,” dealt with other aspects of a publication beyond the research area itself and included four categories

(again, see Table 2.1). Last, we used a keyword-based approach, adding keywords to each definition.

After developing initial revised definitions, we shared these with the O*NET Center for feedback. Adjustments were made based on two rounds of feedback. Table 2.1 includes the original content category names and definitions, revised category names and definitions, the new Artificial Intelligence category definition, and the two overarching domain structures.

Table 2.1. Current and Revised Content Category Definitions

Original Category Label	Revised Category Label	Original Category Definition	Revised Category Definition
Focal Content Domains			
Industrial / Organizational Research	**Industrial-Organizational Psychology Content**	General research conducted using O*NET's products, tools, or databases that does not focus on application development for the purpose of answering a conceptual or theoretical question related to topics such as personnel selection/recruitment, individual assessment, occupationally set training/development, and job analysis.	Content is focused on topics studied in the area of Industrial-Organizational Psychology examples of which include but are not limited to: job search, human resources, personnel selection/recruitment, individual assessment, motivation, leadership, job analysis, job performance and attitudes, and organizational behavior.
Health Research	**Health Content**	General research conducted using O*NET's products, tools, or databases that does not focus on application development for the purpose of answering a conceptual or theoretical question related to topics such as public health, physical safety, epidemiology, and physical/mental illness.	Content is focused on topics studied in the areas of medicine, public/occupational health and workplace safety examples of which include but are not limited to: assessment of medical students/professionals, requirements of medical practice or occupations, public health and safety, epidemiology, occupational health psychology, workplace injuries, illness prevention, occupational stress/strain, mental health, and worker well-being.
Education Research	**Education Content**	General research conducted using O*NET's products, tools, or databases that does not focus on application development for the purpose of answering a conceptual or theoretical question related to topics such as education/work preparedness, academic training, and child/adolescent development.	Content is focused on topics studied in the areas of education and workforce development examples of which include but are not limited to: education/work preparedness, education-to-work transition, academic training, skills development through formal or informal learning, curriculum design, and child/adolescent development.
Career and Vocational Research	**Career and Vocational Content**	General research conducted using O*NET's products, tools, or databases that does not focus on application development for the purpose of answering a conceptual or theoretical question related to topics such as career counseling, vocational interests, and vocational rehabilitation.	Content is focused on topics studied in the areas of career development and vocational counseling including examples of which include but are not limited to: career counseling, vocational interests (e.g., RIASEC), work values, occupational fit, person-job fit, vocational rehabilitation, and career decision-making.

Table 2.1. (Continued)

Original Category Label	Revised Category Label	Original Category Definition	Revised Category Definition
Labor Market Research	**Labor Economics Content**	General research conducted using O*NET's products, tools, or databases that does not focus on application development for the purpose of answering a conceptual or theoretical question related to topics such as descriptive states of industry staffing/skills/knowledge, prescriptive actions on issues/conditions in the labor market, wages/wealth, and longitudinal comparisons of labor market demographics.	Content is focused on topics studied in the area of labor economics examples of which include but are not limited to: occupational staffing patterns, skill/knowledge distribution and gaps in the labor market, wage trends, economic development or characteristics, labor market dynamics, employment projections, and longitudinal comparisons of labor market demographics.
Other Research	**Other Field-Specific Content**	General research conducted using O*NET's products, tools, or databases that does not focus on application development for the purpose of answering a conceptual or theoretical question related to topics not covered in other categories.	Content is focused on topics studied in areas other than I-O psychology, education, health, career and vocational counseling, and labor economics.
Artificial Intelligence	**Artificial Intelligence**	*Not a Current Category*	Publication involves O*NET's products, tools, or database in conjunction with Artificial Intelligence (AI), Natural Language Processing (NLP), or Machine Learning (ML) methods. This includes development or evaluation of AI, NLP, or ML-driven applications that utilize O*NET content or data, or research employing AI, NLP, or ML methods with O*NET content or data.
Application Development	**Application and Product Development**	Research carried out by entities other than O*NET or O*NET contractors on applications developed using O*NET products, tools, or databases. This may include new products, tools, or processes developed.	Publication has an explicit focus on the design, development, or technical evaluation of a new application, software tool, instrument, product, or service that is built using O*NET data or tools.
O*NET Evaluation	**O*NET Evaluation**	Independent research conducted on O*NET's products, tools, or databases for the purpose of evaluating O*NET's products, tools, or databases. This type of research is conducted not for application development, but to document the validity, reliability, or other psychometric, statistical, or content characteristics of O*NET products, tools, or databases.	Publication has an explicit focus on the evaluation of the quality of O*NET's products, tools, or databases and has a URL that does NOT include "onetcenter.org".

Table 2.1. (Continued)

Original Category Label	Revised Category Label	Original Category Definition	Revised Category Definition
O*NET Initiatives and Development	**O*NET Program**	Research conducted or commissioned by O*NET pertaining broadly to O*NET products, tools, or databases and associated components (e.g., algorithms).	Publication is conducted or commissioned by the O*NET Program or U.S. Department of Labor, or research authored by O*NET employees and contractors, pertaining to O*NET products, tools, databases, or associated components (e.g., algorithms). Includes reports, articles, and development work originating from the O*NET Resource Center or O*NET-sponsored project. Only assign this tag to a publication if its URL contains “onetcenter.org”

Step 3: Develop LLM Prompt Templates for Content Tagging

After pulling abstracts, we developed and experimented with various LLM prompt templates and models to assign content category tags to publications. This process was done in conjunction with Step 2 to support revisions to the content category definitions. All data used in this step were excluded from future efforts with SME raters (Step 5). For developing the prompt, we selected 45 publications from the full list, five from each of the nine existing content categories. Our input data from these publications included the publication's reference, URL, and abstract.

To facilitate a prompt-based development process and automate LLM-based tag generation, we employed HumRRO's secure generative AI platform (RRObot), built on the Amazon Web Services (AWS) Bedrock service (Amazon Web Services, n.d.; https://aws.amazon.com/bedrock/). AWS Bedrock is a service that provides organizations with access to a variety of foundation models (e.g., LLMs from third-party providers such as Meta, OpenAI, and Amazon) to drive the functionality of generative AI applications, including HumRRO's RRObot platform (Human Resources Research Organization, 2025). The RRObot platform provides HumRRO staff with a suite of AI tools to facilitate the delivery of high-stakes assessment and talent management solutions. RRObot operates in a completely secure environment (based on adherence to current National Institute of Standards and Technology [NIST]/International Organization for Standardization [ISO]/Cybersecurity Maturity Model Certification [CMMC] compliance standards and boundaries) where client information remains protected. RRObot allows HumRRO staff to interact with LLMs. It provides HumRRO staff with highly accessible, practical means to significantly streamline, enhance, and reimagine how we do our work, leveraging generative AI. For this effort, we leveraged RRObot's batch processing capabilities to upload a file of prompt templates designed to tag publications into content categories.

We followed several steps to develop our final LLM prompt. We first developed and tested initial prompt templates to ensure they produce the desired output (i.e., that the final output file included yes/no tags for each publication across all ten content categories, in the specified format). After our output file returned the expected format, we examined the initial accuracy of the LLM models for this process to help determine which model to use for tagging. The final prompt that we developed is in Appendix B.

Next, the first author (Alexander McKay) served as an initial SME and tagged 45 publications using the proposed process. These publications were tagged using the current nine category definitions and the new definition for the Artificial Intelligence category. Each publication's original content category was not considered in the tagging exercise. After completing these ratings, the 45 publications were tagged by the Llama 4 Maverick model (Meta, 2025). The Llama 4 Maverick model was used because it is a less expensive yet powerful LLM. This made it ideal for quick experimentation, where multiple runs would be performed and tested. We then compared the human and LLM tags, carefully reviewing disagreements. Disagreements were then adjusted based on a review of the publication information by the first author. This process was repeated several times, as LLM ratings can change across runs. The final tags served as "ground truth" for comparing the performance of various models.

The final step was to compare different LLM models and conduct an initial validation study. We used the ground truth tags to assess how well the prompts were working and to compare several AI models with different hyperparameter settings. We tested five models with two different hyperparameter settings for each model. The two hyperparameter settings we tested

were: (1) Top *P* set at 1.0 and Temperature set at 0.2, and (2) Top *P* set at 0.2 and Temperature set at 1.0. Temperature controls how random or "creative" the AI model's responses are: lower values make outputs more focused and predictable, whereas higher values increase variety. Top *P* (also called nucleus sampling) controls how narrowly the LLM focuses on the most probable answer, with lower values making it stick closely to the statistically strongest choice and higher values allowing it to consider a wider range of response probabilities. For each model, we had the LLM tag the publications three times to verify tag consistency across runs. Table 3.1 includes information about the five models. These models were chosen as the number of parameters varies along with cost, and they show different patterns of ratings (e.g., we have found that Llama 4 is more conservative in its ratings, indicating that a tag does not apply to a category, relative to other models). If we could find similar or better-quality results with a less expensive model, we would use that model for tagging all publications. We also tested an "ensemble" model using the first sets of tags from each of the five models to determine whether combining models would be better than any single model. That is, we used the tags from the first of the three runs for each model. This produced 12 tag sets (five individual models and one ensemble model) that we could compare across models, content categories, and ground truth ratings. Here, the 12 "sets" reflect the combinations of five models (plus one ensemble) times two hyperparameter combinations.

Table 3.1. List of LLMs Examined and Usage

LLM	Full LLM Name-Version	Release Date	# of Parameters	Usage	Citation
Llama 4 Maverick	us.meta.llama4-maverick-17b-instruct-v1:0	April 5, 2025	17B	Used for experimentation	Meta (2025)
Claude Haiku 4.5	us.anthropic.claude-haiku-4-5-20251001-v1:0	October 15, 2025	Not disclosed	Used for experimentation	Anthropic (2025)
Claude Sonnet 4.5	us.anthropic.claude-sonnet-4-5-20250929-v1:0	September 29, 2025	Not disclosed	Used for experimentation and final model used for tagging	Anthropic (2025)
Claude Opus 4.5	us.anthropic.claude-opus-4-5-20251101-v1:0	November 24, 2025	Not disclosed	Used for experimentation	Anthropic (2025)
Claude Opus 4.6	us.anthropic.claude-opus-4-6-v1	February 5, 2026	Not disclosed	Used for experimentation	Anthropic (2025)

In the sections that follow, we present two sets of results. First, we examine agreement in LLM content category assignments across runs within each LLM (and across LLMs in the case of the ensemble). Second, we examine agreement between LLM content category assignments and ground truth assignments. All analyses are based on the 45 publications referenced above for initial evaluation.

For the 45 publications examined, Table 3.2 provides the percentages for which (a) all three runs of a given LLM agreed in their assignment (or non-assignment) of a given content category to a publication, and (b) all models in the ensemble agreed in their assignment (or non-

assignment) of a given content category to a publication. The results show that the outputs of each LLM varied little across runs, with overall agreement rates ranging from 95.9% to 100%. Agreement rates for the ensembles were lower, suggesting that models were less aligned in their assignments of content categories to publications than in assignments made by different runs of the same model. This is not necessarily a negative finding, in that it could indicate that different models are providing different nuanced perspectives on category assignment, which, in combination, may have value for the prediction of ground truth ratings, relative to the narrower perspective provided by any single LLM.

We then compared the LLM's results to the ground truth ratings. For this comparison, we used a majority vote approach: within-model comparisons require that at least two out of three runs agree, and for the ensemble model comparisons, at least three out of five models must agree. The results are shown in Table 3.3. The model with the lowest overall agreement was Llama 4 (86.2%), and the model with the highest overall agreement was Sonnet 4.5, regardless of hyperparameter settings. Although there was variation in the percent agreement across content categories between LLM tags and ground truth ratings, it was evident that Sonnet 4.5 with hyperparameters of Top *P* = 0.2 and Temperature = 1.0 was the better model.

Based on these initial pilot results, we decided to use the Sonnet 4.5 model with Top *P* = 0.2 and Temperature = 1.0 to tag all publications in the full dataset. This decision was based on several factors. First, there was perfect agreement across the three runs, highlighting consistency in how it tags publications. Second, it showed the strongest overall agreement with the ground truth ratings, indicating that it performs well. Last, it strikes a balance between cost and accuracy. This model is less expensive than the two Opus models, which demonstrate a similar performance. It also performed better than Haiku 4.5, another less expensive model with fewer parameters. In sum, Sonnet 4.5 balances the tradeoff of performing quite well while being more cost effective than some of the more powerful models. In future steps, we will use content tags from the Sonnet 4.5 model on a larger subset of publications to check model accuracy with SME tags.

Table 3.2. Percentage of Publications for Which There Was Full Agreement on Category Assignments across Runs within LLMs and Across Models for LLM Ensembles

Percentage of Publications for which there was Full Agreement on Category Assignments (across runs for LLMs, across models for ensembles)												
	Hyperparameters: Top *P* = 0.2, Temp = 1.0						Hyperparameters: Top *P* = 1.0, Temp = 0.2					
Category	**Llama 4**	**Haiku 4.5**	**Sonnet 4.5**	**Opus 4.5**	**Opus 4.6**	**Ensemble**	**Llama 4**	**Haiku 4.5**	**Sonnet 4.5**	**Opus 4.5**	**Opus 4.6**	**Ensemble**
Artificial Intelligence	100	100	100	100	100	100	100	100	100	100	100	100
Application & Product Development	99.3	100	100	100	99.3	71.6	100	100	99.3	99.3	99.3	78.7
Career and Vocational Content	91.1	100	100	100	100	85.8	100	100	99.3	99.3	100	92.9
Education Content	94.8	100	100	99.3	100	78.7	99.3	100	100	99.3	99.3	75.1
Health Content	99.3	100	100	100	100	94.7	100	100	100	100	100	94.7
Industrial-Organizational Psychology Content	93.3	100	100	100	100	64.4	100	100	100	100	100	68.0
Labor Economics Content	92.6	100	100	100	100	85.8	100	100	99.3	100	99.3	85.8
O*NET Evaluation	97.0	100	100	100	100	92.9	98.5	100	100	100	100	96.4
O*NET Programs	97.8	100	100	100	100	94.7	100	100	100	100	100	100
Other Field-Specific Content	93.3	100	100	98.5	100	82.2	99.3	100	100	99.3	98.5	85.8
Overall	95.9	100	100	99.8	99.9	85.1	99.7	100	99.8	99.7	99.6	87.7

Note. *N* = 450 total tags across 45 publications. Ensemble results are based on the first run of each of the five models, whereas each individual model was based on three runs. Cell values reflect percent agreement across LLM runs.

Table 3.3. Percentage of Publications for Which There Was Agreement between LLM-Based and Ground Truth Ratings

	Hyperparameters: Top *P* = 0.2, Temp = 1.0						Hyperparameters: Top *P* = 1.0, Temp = 0.2					
Category	**Llama 4**	**Haiku 4.5**	**Sonnet 4.5**	**Opus 4.5**	**Opus 4.6**	**Ensemble**	**Llama 4**	**Haiku 4.5**	**Sonnet 4.5**	**Opus 4.5**	**Opus 4.6**	**Ensemble**
Artificial Intelligence	100	100	100	100	100	100	100	100	100	100	100	100
Application & Product Development	75.6	84.4	88.9	86.7	84.4	86.7	93.3	84.4	88.9	86.7	84.4	86.7
Career and Vocational Content	82.2	91.1	95.6	93.3	91.1	91.1	88.9	91.1	95.6	93.3	91.1	91.1
Education Content	82.2	75.6	86.7	93.3	86.7	86.7	62.2	75.6	84.4	91.1	86.7	84.4
Health Content	97.8	97.8	97.8	97.8	95.6	97.8	97.8	97.8	97.8	97.8	95.6	97.8
Industrial-Organizational Psychology Content	77.8	75.6	82.2	82.2	80.0	84.4	88.9	75.6	82.2	82.2	80.0	82.2
Labor Economics Content	82.2	84.4	91.1	88.9	91.1	91.1	84.4	84.4	91.1	88.9	91.1	91.1
O*NET Evaluation	82.2	86.7	91.1	91.1	91.1	91.1	88.9	86.7	91.1	91.1	91.1	91.1
O*NET Programs	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6
Other Field-Specific Content	86.7	86.7	91.1	84.4	86.7	84.4	86.7	86.7	88.9	84.4	86.7	86.7
Overall	86.2	87.8	92.0	91.3	90.2	90.9	88.7	87.8	91.6	91.1	90.2	90.7

Note. *N* = 450 total tags across 45 publications. Ensemble results are based on the first run of each of the five models, whereas each individual model was based on three runs. The final LLM tags used for comparison with the ground truth are based on majority vote: for within-model comparisons, at least two out of three runs must agree; for the ensemble model comparisons, at least three out of five models must agree. Cell values reflect the percent agreement between LLM and ground truth category assignments.

Step 4: Draw Sample of Publications and Run Them Through LLM Content Tagging Process

After revising the content category definitions (Step 2), developing the final prompt (Step 3), and selecting the final LLM model (Step 3), we sought to examine the effectiveness of the LLM-based process against an expanded set of publications from the reference list to provide a more robust evaluation of the LLM-based category assignments.

To facilitate evaluation of the LLM-based content tagging process, we identified and drew a sample of 180 publications deemed relevant to each of the current nine content categories from the full list of 2,850 publications referenced in Step 2. These 180 publications do not reflect random samples from the full set of publications for two reasons. First, in some exploratory analyses (not reported here), we observed small differences in the number of tags assigned to publications with missing URLs and/or abstracts. Second, we found that some abstracts were more detailed than others, raising the question of whether tag quality might differ by data availability. Thus, we wanted to systematically compare these factors in future steps (Step 6).

We selected 20 publications from each of the current nine content categories, for a total of 180 publications. To select these publications, we calculated the number of words in each abstract. Within each category, we sorted the publication list from largest to smallest based on this count. We then selected the top seven publications in the list, six publications from the middle, and seven from the bottom. Three of the seven at the bottom of the list were publications with missing abstracts. Thus, some publications in the sample of 180 lacked abstracts. This was by design to help us later evaluate how the LLM-based process performed in the absence of abstract information for publications.

We ran this subset of 180 publications along with the final content tagging prompt through RRObot. In the data file, we included the publication reference, URL, and abstract. The cells for publications with missing URLs and abstracts were left empty. We used Sonnet 4.5 with Top *P* = 0.2 and Temperature = 1.0 and conducted three runs. We used a majority vote to determine content categories.

Table 4.1 includes agreement scores across the three runs and ten content categories. Full agreement was defined as all three LLM runs having the same tag for a given category and publication within a category. There was full agreement for eight of the 10 categories, indicating the three runs tagged publications consistently.

Table 4.1. Percentage of Publications for Which There Was Full Agreement on Category Assignments Across Runs of Sonnet 4.5 (N = 180)

Category	*N*	Percent
Artificial Intelligence	180	100
Application & Product Development	179	99.4
Career and Vocational Content	180	100
Education Content	180	100
Health Content	180	100
Industrial-Organizational Psychology Content	180	100
Labor Economics Content	180	100
O*NET Evaluation	180	100
O*NET Programs	180	100
Other Field-Specific Content	179	99.4
Overall	1,798	99.9

Note. *N* = 1,800 possible tags across 180 publications.

We then combined the three LLM runs using a majority rule approach. That is, for a category to be tagged "yes," two of the three runs had to tag it as relevant; otherwise, it would be tagged as "no." Table 4.2 reports the total counts of yes and no tags for the 180 publications (i.e., 1,800 tags across publications and categories). The percentages are based on individual category totals. The top five most tagged categories were Career and Vocational Content, Education Content, Labor Economics Content, Industrial-Organizational Psychology Content, and Health Content.

Table 4.2. Total Yes and No Counts and Percentages Based on Final Majority Rule Data

	Yes		No	
Category	*N*	Percent	*N*	Percent
Artificial Intelligence	22	12.2	158	87.8
Application & Product Development	33	18.3	147	81.7
Career and Vocational Content	73	40.6	107	59.4
Education Content	56	31.1	124	68.9
Health Content	37	20.6	143	79.4
Industrial-Organizational Psychology Content	40	22.2	140	77.8
Labor Economics Content	53	29.4	127	70.6
O*NET Evaluation	6	3.3	174	96.7
O*NET Programs	19	10.6	161	89.4
Other Field-Specific Content	12	6.7	168	93.3
Overall	351	19.5	1,449	80.5

Note. *Note*. *N* = 1,800 possible tags across 180 publications. Percentages are based on individual category totals.

Step 5: Collect HumRRO SME Judgments

We provided four SMEs with the list of 180 publications from Step 4 with the same information the LLM was using: the publication reference, URL, and abstract. Three SMEs were HumRRO staff members with experience assigning content categories to publications in the current Reference List, as well as one additional staff member with experience working with O*NET data. Each SME tagged publications with a yes or no for each of the ten categories based on the information provided (i.e., 1,800 tags × publication combinations). They were given instructions similar to what the LLM received. These instructions are provided in Appendix C. Note that the definitions for Application and Product Development and for O*NET Evaluation differed between the LLM runs and the human tagging process.[4]

Table 5.1 includes full agreement among the four SMEs. Full agreement was defined as all four SMEs having the same tag for a given category and publication within a category. Overall, the four SMEs agreed on 1,428 of the 1,800 tags (79.3%), which was lower than the LLM results. Agreement for six of the 10 categories was above 80% across publications within a category. The four below 80% were Career and Vocational Content (67.8%), Industrial-Organizational Psychology Content (55.6%), Labor Economics Content (71.1%), and Other Field-Specific Content (70.6%). Overall, there was adequate-to-strong agreement among SMEs for most categories. Categories with lower agreement were flagged for further review when comparing the results to the LLM in Step 6.

Table 5.1. Percentage of Publications for Which There Was Full Agreement on Category Assignments Across the Four HumRRO SMEs

Category	Full Agreement *N*	Full Agreement Percent
Artificial Intelligence	168	93.3
Application & Product Development	155	86.1
Career and Vocational Content	122	67.8
Education Content	144	80.0
Health Content	159	88.3
Industrial-Organizational Psychology Content	100	55.6
Labor Economics Content	128	71.1
O*NET Evaluation	154	85.6
O*NET Programs	171	95.0
Other Field-Specific Content	127	70.6
Overall	1,428	79.3

Note. *N* = 1,800 possible tags across 180 publications.

[4]We used different definitions for these two content categories after comparing the results from the SME- and LLM-based tags. We found greater agreement between SMEs and the LLM with these small adjustments.

We then combined the SME tags using a majority rule approach. That is, for a category to be tagged "yes," three of the four SMEs had to tag it as relevant; otherwise, it would be tagged as "no." Table 5.2 reports the total counts of yes and no tags for the 180 publications (i.e., 1,800 tags across publications and categories). The percentages are based on individual category totals. The top five most tagged categories were Industrial-Organizational Psychology Content, Career and Vocational Content, Health Content, Labor Economics Content, and O*NET Programs. Overall, the "yes" tag rates were about half of those observed from the LLM in Step 4. We considered those differences in the analyses reported in Step 6.

Table 5.2. Total Yes and No Counts and Percentages Based on Final Majority Rule Data

	Yes		No	
Category	***N***	**Percent**	***N***	**Percent**
Artificial Intelligence	8	4.4	172	95.6
Application & Product Development	3	1.7	177	98.3
Career and Vocational Content	34	18.9	146	81.1
Education Content	17	9.4	163	90.6
Health Content	24	13.3	156	86.7
Industrial-Organizational Psychology Content	35	19.4	145	80.6
Labor Economics Content	23	12.8	157	87.2
O*NET Evaluation	2	1.1	178	98.9
O*NET Programs	19	10.6	161	89.4
Other Field-Specific Content	8	4.4	172	95.6
Overall	173	9.6	1,627	90.4

Note. *N* = 1,800 tags across 180 publications. Percentages are based on individual category totals.

Step 6: Evaluate LLM Judgments Against HumRRO SME Judgments

To evaluate the quality of the LLM-generated tags, we compared the LLM-based tags from Step 4 with the HumRRO SME tags from Step 5 for the evaluation sample of 180 publications. We used the majority rule tags from both sources to determine the final tags (i.e., the best two out of three for the LLM and the best three out of four for the SMEs). SME tags were treated as the ground truth for these comparisons. It is worth noting, however, that when the new tags were matched against current content categorizations, the LLM demonstrated stronger alignment than the SMEs. This finding was taken into consideration when interpreting the agreement results.

Analysis of LLM-SME Agreement

Table 6.1 reports the agreement statistics between SME and LLM tags across the 180 publications (1,800 tag × publication combinations). An agreement was defined as both the LLM and SME majority vote ratings being the same (i.e., either both have a "yes" tag or both have a "no" tag). We relied on the agreement count and percent to determine the quality of the LLM tags. Overall, full agreement was 87.1% (1,568 of 1,800 tags). Agreement varied considerably across categories. O*NET Programs was the strongest performer, achieving perfect agreement (100% of 180 publications). This was primarily due to the explicit rule that publications with "onetcenter.org" in the URL should be tagged with this content category. Three other categories – Health (92.8%), Other (93.3%), and Artificial Intelligence (92.2%) – also demonstrated strong agreement. In contrast, the Career and Vocational Content and Industrial-Organizational Psychology Content categories showed the lowest full agreement (76.1% each), along with Education Content (78.3%).

Table 6.1. Agreement Between the HumRRO SMEs and LLM Tags

Category	Full Agreement *N*	Full Agreement Percent
Artificial Intelligence	166	92.2
Application & Product Development	150	83.3
Career and Vocational Content	137	76.1
Education Content	141	78.3
Health Content	167	92.8
Industrial-Organizational Psychology Content	137	76.1
Labor Economics Content	148	82.2
O*NET Evaluation	174	96.7
O*NET Programs	180	100.0
Other Field-Specific Content	168	93.3
Overall	1,568	87.1

Note. *N* = 1,800 tags across 180 publications. Percentages reflect the percentage of those publications for which LLM-SME judgements agreed with respect to tagging the publication with the given category.

Table 6.2 presents the confusion matrix comparing human SME and LLM tags across the 10 categories. Across all 1,800 cases, the LLM produced 146 true positives, 1,422 true negatives, 205 false positives, and 27 false negatives. The low false negative count (27) indicates that the LLM rarely missed tags assigned by SMEs to publications. However, the higher false positive count (205) indicates that the LLM tagged approximately 10% of the cases as "yes" when the SMEs did not. We flagged five content categories (Application and Product Development, Career and Vocational Content, Education Content, Industrial-Organizational Psychology Content, and Labor Economics Content) with high false positive rates (greater than 20), with Career and Vocational Content (41) and Education Content (39) showing the highest counts. Notably, the Industrial-Organizational Psychology Content category was the only one with roughly equal numbers of false positives (24) and false negatives (19), suggesting that human SMEs and the LLM diverged somewhat on what should or should not be tagged with this content category.[5] O*NET Programs had no false positives or negatives, and O*NET Evaluation, Other Field-Specific Content, Health Content, and Artificial Intelligence showed low overall error rates.

Table 6.2. Human vs. LLM Confusion Matrix

Category	True Positive	True Negative	False Positive	False Negative
Artificial Intelligence	8	158	14	0
Application & Product Development	3	147	30	0
Career and Vocational Content	32	105	41	2
Education Content	17	124	39	0
Health Content	24	143	13	0
Industrial-Organizational Psychology Content	16	121	24	19
Labor Economics Content	22	126	31	1
O*NET Evaluation	1	173	5	1
O*NET Programs	19	161	0	0
Other Field-Specific Content	4	164	8	4
Overall	146	1,422	205	27

Note. N = 1,800 tags across 180 publications.

We then sought to examine differences in the human SME and LLM tag counts to understand who provided more publications tags overall. Table 6.3 illustrates that the LLM assigned "yes" tags more frequently than the SMEs across most categories (351 vs. 173 overall). The greater number of assigned tags was most pronounced for Career and Vocational Content (73 vs. 34), Education Content (56 vs. 17), Application & Product Development (33 vs. 3), and Labor Economics Content (53 vs. 23). Overall, these results indicate that the LLM was more likely to tag publications as being relevant to these content categories than the human SMEs.

[5]The balanced false positive and false negative rates for the Industrial-Organizational Psychology Content category may reflect the specialized expertise of the SME raters, all of whom have backgrounds in Industrial-Organizational Psychology. Their deeper domain knowledge in this area may have led to more nuanced tagging decisions that diverged more frequently from the LLM compared to the other nine categories.

Table 6.3. Human vs. LLM Yes/No Tag Counts

	Human		LLM	
Category	**Yes**	**No**	**Yes**	**No**
Artificial Intelligence	8	172	22	158
Application & Product Development	3	177	33	147
Career and Vocational Content	34	146	73	107
Education Content	17	163	56	124
Health Content	24	156	37	143
Industrial-Organizational Psychology Content	35	145	40	140
Labor Economics Content	23	157	53	127
O*NET Evaluation	2	178	6	174
O*NET Programs	19	161	19	161
Other Field-Specific Content	8	172	12	168
Overall	173	1,627	351	1,449

Note. N = 1,800 tags across 180 publications.

Table 6.4 further reinforces the pattern of more LLM-based tags at the publication level. On average, the LLM assigned nearly twice as many tags per publication as the human SMEs, M = 1.95 vs. M = 0.96, respectively. Humans were more likely to classify publications with zero “yes” tags (42) relative to the LLM (2), and the LLM tagged more publications with three or more tags (50) compared to the human SMEs (4). Of the 180 publications, 115 (63.9%) received more tags from the LLM than from the SMEs, compared to only 13 (7.2%) where the SMEs assigned more tags. Finally, 52 (28.9%) publications had the same number of tags assigned by the LLM and the SMEs. In sum, these results highlight that the LLM assigned more tags overall and was more likely to tag each publication with more tags than the human SMEs.

Table 6.4. Number of Tags Assigned to Publications

Number of Tags	Human Tags	LLM Tags
0 Tags	42	2
1 Tag	107	70
2 Tags	27	58
3 Tags	4	36
4 Tags	0	13
5 Tags	0	1
M Number of Tags	0.96	1.95
SD	0.69	0.99
Publications with more human tags	13 (7.2%)	
Publications with more LLM tags	115 (63.9%)	
Publications with an equal number of tags	52 (28.9%)	

Note. Counts for the human and LLM columns each sum to 180 publications.

Follow-Up Evaluation of LLM-SME Disagreements

After identifying the number of disagreements, we sought to determine which source was correct: human SMEs or the LLM. This deeper examination into disagreements was based on our findings that when the new tags were matched against current content categorizations, the LLM demonstrated stronger alignment than the SMEs. The first two authors independently reviewed the 232 disagreements. The first author counted 82 tags as correct for the human SMEs and 150 as correct for the LLM. The second author counted 73 tags as correct for the human SMEs and 159 as correct for the LLM. They agreed on 141 of these judgments, disagreeing on 41. As such, rather than disagreement between human and LLM tags suggesting inaccuracies on the part of the LLM, it actually suggested that the bulk of the inaccuracies may have been in the human-based judgments. Additionally, we observed that human judges appeared to be under-tagging relative to the LLM (i.e., human SMEs not assigning a content tag when they perhaps should have). Thus, it appears that the LLM was not being too lenient in its assignment of tags relative to humans (i.e., over-tagging). Rather, the SMEs tended to tag less than perhaps was warranted (e.g., a severity or under-tagging bias on the part of humans, perhaps stemming from many of the SMEs' experience with the original content category assignment process which was a classification rather than tagging task). We consider these results in our recommendations for future use.

Level of LLM-SME Agreement by Abstract Word Count

Finally, we examined agreement across publications between the three different word count categories. As a reminder, we selected 20 publications from each category based on abstract word count. We sorted each category's publication list from largest to smallest based on this count and selected the top seven publications in the list, six publications from the middle, and seven from the bottom. Three of the seven at the bottom of the list were publications with missing abstracts, meaning some of the tags examined were based on them.

Table 6.5 reports these results. An agreement was defined as both the LLM and SME majority vote ratings being the same. The percentages reflect the rate of agreement within a given cell relative to the total cell count (i.e., 56 tags out of 63 for the Artificial Intelligence category when word count was low, yielding an 88.9% agreement rate). Overall, we found that when the abstract word count was higher, there was greater agreement for tags than when it was low (90.2% vs. 83.7%), with moderate word count falling in between (87.6%). This trend was consistent across nearly all categories. The content category with the greatest improvement was Application and Product Development, improving from 71.4% at low word count to 90.5% at high word count. The Labor Economics Content, Career and Vocational Content, and Industrial-Organizational Psychology Content also showed the most notable improvements from low to high word count (76.2% to 87.3%, 73.0% to 81.0%, and 71.4% to 79.4%, respectively). O*NET Programs maintained perfect agreement (100%) across all word count categories, whereas O*NET Evaluation was consistently high, ranging from 95.2% to 98.4%. Conversely, Career and Vocational Content and Industrial-Organizational Psychology Content remained the weakest performing categories regardless of word count. This suggests that agreement challenges for these categories may be less related to abstract length and more reflective of inherent ambiguity in how these topics are classified. Based on the analysis of disagreements by the authors of this report, the disagreements for these categories indicated that the LLM was more likely to be "right."[6]

[6]We also examined differences between publications with and without an abstract separately. However, we did not find any meaningful differences in agreement, likely because of the small sample of publications without an abstract. Therefore, we do not report those results. Notably, any differences are already included in the differences observed between low and high word count abstracts.

Table 6.5. Agreement Between the HumRRO SMEs and LLM Tags Based on Word Count

	Low Word Count		Moderate Word Count		High Word Count	
Category	***N***	**Full Agreement Percent**	***N***	**Full Agreement Percent**	***N***	**Full Agreement Percent**
Artificial Intelligence	63	88.9	54	92.6	63	95.2
Application & Product Development	63	71.4	54	88.9	63	90.5
Career and Vocational Content	63	73.0	54	74.1	63	81.0
Education Content	63	77.8	54	77.8	63	79.4
Health Content	63	92.1	54	88.9	63	96.8
Industrial-Organizational Psychology Content	63	71.4	54	77.8	63	79.4
Labor Economics Content	63	76.2	54	83.3	63	87.3
O*NET Evaluation	63	95.2	54	96.3	63	98.4
O*NET Programs	63	100.0	54	100.0	63	100.0
Other Field-Specific Content	63	90.5	54	96.3	63	93.7
Overall	630	83.7	540	87.6	630	90.2

Note. *N* reflects the number of publications with the given abstract word count level (low, moderate, high). Percentage reflects the percentage of those publications for which LLM-SME judgements agreed with respect to tagging the publication with the given category.

Summary

In summary, agreement between the human SME and LLM tags was strong, with a full agreement rate of 87.1% across the 1,800 tag × publication combinations. Agreement varied across categories, with O*NET Programs achieving perfect agreement and Career and Vocational Content and Industrial-Organizational Psychology Content showing the lowest agreement (76.1% each). Where disagreements occurred, they were predominantly driven by the LLM assigning more "yes" tags than the human SMEs, with 88.4% of disagreements reflecting the LLM tagging a case as "yes" when the SMEs did not. The LLM also assigned nearly twice as many tags per publication on average (M = 1.95 vs. M = 0.96). However, a closer review of the 232 disagreements by the first two authors of this report suggested that the LLM was more often correct, indicating that the LLM's tendency toward additional tagging may reflect the intent of the new multi-tag classification approach rather than lower quality judgments. Finally, agreement was positively associated with abstract word count. Full agreement increased from 83.7% for low word count publications to 90.2% for high word count publications, suggesting that richer publication abstracts support more consistent tagging. Based on these results, we sought to tag all 2,850 publications using the proposed LLM tagging approach.

Step 7: Finalize Prompts and Run All Current Reference List Publications Through LLM Content Tagging Process

This step involved running the complete list of 2,850 publications from Step 1 against the final content tagging prompt (see Appendix B). In the data file, we included the publication reference, URL, and abstract. Publications with missing URLs and abstracts were included, and those cells were left empty. We used Sonnet 4.5 with Top *P* = 0.2 and Temperature = 1.0 and conducted three runs. Using these initial tags, we conducted several analyses. First, we examined agreement across the three LLM runs. Second, we calculated descriptive scores to better understand tag counts, both individual counts and tag co-occurrences. Third, we determined whether tag counts differed across levels of missing data. Last, we examined whether current content categorizations aligned with the LLM-based tags. For all descriptive analyses and those examining alignment with existing categories, we used a majority vote rule for the LLM content category tags.

Analysis of Agreement Across the Three LLM Runs

Table 7.1 includes agreement scores across the three runs and ten content categories. Full agreement was defined as all three LLM runs having the same tag for a given category and publication within a category. There was full agreement for three of the ten categories. For the other seven categories, agreement was above 99.5%, indicating that the three runs consistently tagged publications.

Table 7.1. Percentage of Publications for Which There Was Full Agreement on Category Assignments Across Runs of Sonnet 4.5 (N = 2,850)

Category	*N*	Percent
Artificial Intelligence	2,850	100
Application & Product Development	2,839	99.6
Career and Vocational Content	2,846	99.9
Education Content	2,846	99.9
Health Content	2,849	100
Industrial-Organizational Psychology Content	2,850	100
Labor Economics Content	2,849	100
O*NET Evaluation	2,847	99.9
O*NET Programs	2,850	100
Other Field-Specific Content	2,847	99.9
Overall	28,473	99.9

Note. *N* = 28,500 possible tags across 2,850 *publications*.

Descriptive Analyses of All Final Majority Tags

Table 7.2 reports the total counts of “yes” and “no” tags for all 2,850 publications (i.e., 28,500 tags across publications and categories). The percentages are based on the individual category total. The top five most commonly occurring content tags were Labor Economics Content,

Career and Vocational Content, Industrial-Organizational Psychology Content, Education Content, and Health Content.

Table 7.2. Total Yes and No Counts and Percentages Based on Final Majority Rule Data

	Yes		No	
Category	***N***	**Percent**	***N***	**Percent**
Artificial Intelligence	400	14.0	2,450	86.0
Application & Product Development	373	13.1	2,477	86.9
Career and Vocational Content	813	28.5	2,037	71.5
Education Content	795	27.9	2,055	72.1
Health Content	480	16.8	2,370	83.2
Industrial-Organizational Psychology Content	811	28.5	2,039	71.5
Labor Economics Content	1,227	43.1	1,623	57.0
O*NET Evaluation	66	2.3	2,784	97.7
O*NET Programs	229	8.0	2,621	92.0
Other Field-Specific Content	152	5.3	2,698	94.7
Overall	5,346	18.8	23,154	81.2

Note. Percentages are based on individual category totals.

We then examined what category combinations most commonly occur in the data. Table 7.3 reports the most frequent category combinations by publication count. We show these for publications with only a single tag and for publications with two to five tags. Overall, Labor Economics Content was the most frequently occurring category for publications with a single tag. It was also common for it to co-occur with other tags. Artificial Intelligence was also another tag that consistently occurred with other tag combinations.

Table 7.3. Top Five Most Frequent Category Combinations

Category Combinations	*N*	Percent
Most Frequent Publications with One Tag		
Labor	339	11.9
IO	239	8.4
ONET_P	218	7.7
Health	140	4.9
Career	115	4.0
Most Frequent Publications with Two Tags		
Educ, Labor	164	5.8
Career, Educ	134	4.7
AI, Labor	104	3.7
Health, Labor	81	2.8
Health, IO	77	2.7
Most Frequent Publications with Three Tags		
Career, Educ, Labor	75	2.6
AppProdDev, Career, Educ	34	1.2
AI, Educ, Labor	33	1.2
AI, AppProdDev, Labor	32	1.1
Educ, IO, Labor	29	1.0
Most Frequent Publications with Four Tags		
AI, AppProdDev, Educ, Labor	19	0.7
AI, AppProdDev, Career, Educ	16	0.6
AI, AppProdDev, Career, IO	10	0.4
AI, Career, Educ, Labor	9	0.3
AppProdDev, Career, Educ, Labor	9	0.3
Most Frequent Publications with Five Tags		
AI, AppProdDev, Career, Educ, Labor	14	0.5
AppProdDev, Career, Educ, IO, Labor	7	0.3
AI, Career, Educ, IO, Labor	4	0.1
AI, AppProdDev, Educ, IO, Labor	3	0.1
AI, AppProdDev, Career, Educ, IO	2	0.1

Note. Counts are based on the number of publications (2,850). These do not sum to the total, as this is only a subset of all category combinations. The percentages are based on a denominator of 2,850. AI = Artificial Intelligence; AppProdDev = Application and Product Development; Career = Career and Vocational Content; Educ = Education Content; Health = Health Content; IO = Industrial-Organizational Psychology Content; Labor = Labor Economics Content; ONET_E = O*NET Evaluation; ONET_P = O*NET Programs; Other = Other Field-Specific Content.

Table 7.4 supplements this data by examining co-occurrences among categories. The diagonal shows the total number of "yes" tags for each domain. Off-diagonal scores are the total number of instances in which the two tags co-occur. Notably, the off-diagonal scores can sum to a value greater than the category total but cannot exceed it. For example, there are 400 Artificial

Intelligence tags. The number of any single off-diagonal score for Artificial Intelligence never exceeds that number.

Table 7.4. Co-Occurrence of Categories

Category	Artificial Intelligence	Application & Product Development	Career and Vocational Content	Education Content	Health Content	Industrial-Organizational Psychology Content	Labor Economics Content	O*NET Evaluation	O*NET Programs	Other Field-Specific Content
Artificial Intelligence	400									
Application & Product Development	141	373								
Career and Vocational Content	81	197	813							
Education Content	128	138	374	795						
Health Content	23	35	88	41	480					
Industrial-Organizational Psychology Content	95	108	204	134	133	811				
Labor Economics Content	278	144	216	423	136	207	1,227			
O*NET Evaluation	4	5	26	16	13	22	17	66		
O*NET Programs	9	1	2	0	0	0	0	0	229	
Other Field-Specific Content	29	17	14	23	21	19	34	3	0	152

Note. Counts are based on the number of "yes" tags (5,346). Diagonal scores are the total number of "yes" tags for a given domain. Off-diagonal scores are the total number of instances in which the two tags co-occur.

Descriptive Analyses by Abstract Word Count

We then examined "yes" content category tags based on the amount of missing data. Some publications lacked an abstract and/or URL, making it more difficult to tag them into any content category based on the results in Step 6. Table 7.5 provides descriptive information on the number of "yes" tags for publications with complete data, missing the abstract, and missing the URL and abstract. Overall, the majority of the 5,346 "yes" tags were for publications with complete data (90.4%). The remaining publications were fairly evenly split across the two missing data types.

Table 7.5. Total Yes Counts and Percentages Based on Data Availability

	Complete Data		Missing Abstract		Missing URL and Abstract	
Category	***N***	**Percent**	***N***	**Percent**	***N***	**Percent**
Artificial Intelligence	387	96.8	12	3.0	1	0.3
Application & Product Development	310	83.1	41	11.0	22	5.9
Career and Vocational Content	692	85.1	83	10.2	38	4.7
Education Content	723	90.9	50	6.3	22	2.8
Health Content	465	96.9	10	2.1	5	1.0
Industrial-Organizational Psychology Content	739	91.1	23	2.8	49	6.0
Labor Economics Content	1,138	92.8	56	4.6	33	2.7
O*NET Evaluation	58	87.9	0	0.0	8	12.1
O*NET Programs	196	85.6	19	8.3	14	6.1
Other Field-Specific Content	127	83.6	6	4.0	19	12.5
Overall	4,835	90.4	300	5.6	211	4.0

Note. Counts are based on the number of "yes" tags, summing to a total of 5,346. Percentages are based on individual category totals.

Table 7.6 shows the total number of category combinations for all 2,850 publications and for the two types of missing data. The average number of tags per publication was 1.88 tags. The maximum number of tags per publication was five. The 39 publications that received five tags had complete data. For publications with missing data, the maximum number of content categories for any single publication was four. Thus, the LLM was less likely to tag publications into a greater number of categories (i.e., 4 or 5) when data were missing, particularly when both the URL and the abstract were missing. These results show that most publications fit into multiple categories, which justifies the switch from the current categorization process to the proposed tagging process. Notably, 11 publications (0.4%) received no content category tag. Upon review, these were primarily websites best classified under the Application & Product Development category. Tags were assigned accordingly, so the final dataset contains no publications with zero tags.

Table 7.6. Number of Category Combinations Assigned to Publications Based on Data Availability

	Total Counts		Complete Data		Missing Abstract		Missing URL and Abstract	
Tag Count	***N***	**Percent**	***N***	**Percent**	***N***	**Percent**	***N***	**Percent**
No Tags	11	0.4	1	0.0	7	4.1	3	2.1
Single Tag	1,171	41.1	1,026	40.4	66	38.4	79	56.0
Two Tags	1,043	36.6	932	36.7	66	38.4	45	31.9
Three Tags	450	15.8	406	16.0	30	17.4	14	9.9
Four Tags	136	4.8	133	5.2	3	1.7	0	0.0
Five Tags	39	1.4	39	1.5	0	0.0	0	0.0
Six or More Tags	0	0.0	0	0.0	0	0.0	0	0.0

Note. Counts are based on the number of publications totaling 2,850. The percentages for total counts include 2,850 in the denominator, whereas the percentages for the missing data types are based on each category's total count.

Alignment Between Current Category Tags and LLM-Assigned Tags

We then evaluated the LLM tags by comparing current content categorizations with the new LLM tags. For this analysis, we examined whether the new LLM tags included a “yes” for the category with which each publication is currently tagged. It is important to be cautious when interpreting these results, as any disagreement between the LLM tags and the current content categories may not reflect an issue with the LLM tags but rather fundamental changes to the process. For example, the content definitions changed, and different information was used across the two categorization/tagging processes. Perhaps more importantly, the content category assignment process changed from a “classification” task, where only one category could be assigned to each publication, to a “tagging” task, where more than one category could be assigned to each publication. These differences would be expected to produce nontrivial differences between the LLM-based tags and the current content category assignments. On top of this, a new Artificial Intelligence category was added.

Table 7.7 presents the results of comparing the LLM-based tags to the current content category assignments. In the table, the category counts are for the current process rather than based on the LLM results. Of the 2,850 publications, 2,333 (81.9%) had the current category assigned by the LLM. The current category, assigned by the LLM, varied across publications. The percent of the current category also being assigned by the LLM for Education Research and Health Research was over 90%. Notably, the percentage of the current category assigned by the LLM to Application Development, O*NET Evaluation, and Other Research was low. The LLM's tagging of publications with the O*NET Evaluation category was likely poor, as this judgment is more subjective and based on key information in the broader publication rather than the reference, URL, or abstract. For Application Development, this reason might also apply. For Other Research, this discrepancy might be because multiple categories can be assigned rather than one. These questions require future testing.

Table 7.7. Inclusion and Exclusion of Current Content Category in New LLM Tags

	Current Category Tags	Publication in Current Category was Assigned this Category by LLM		Publication in Current Category was NOT Assigned this Category by LLM	
Category	***N***	***N***	**Percent**	***N***	**Percent**
Artificial Intelligence	*Not a current content category. Cannot be compared.*				
Application Development	258	144	55.8	114	44.2
Career and Vocational Research	292	250	85.6	42	14.4
Education Research	181	174	96.1	7	3.9
Health Research	261	247	94.6	14	5.4
Industrial/Organizational Research	533	420	78.8	113	21.2
Labor Market Research	931	829	89.0	102	11.0
O*NET Evaluation	75	27	36.0	48	64.0
O*NET Initiatives and Development	257	218	84.8	39	15.2
Other Research	62	24	38.7	38	61.3
Overall	2,850	2,333	81.9	517	18.1

Note. Counts in this table are based on the current process.

Summary

Step 7 finalized the LLM-based content tagging process by running all 2,850 publications through the Sonnet 4.5 model across three runs. We then used a majority vote rule to assign final content tags. Agreement across runs was exceptionally high. Across all publications, Labor Economics Content was the most frequently assigned tag (43.1%), and most publications received multiple tags, averaging 1.88 per publication. Notably, the number of tags assigned to publications was influenced by the availability of complete data (i.e., the presence of an abstract and a URL), such that publications with missing data were less likely to receive additional content category tags. When comparing LLM-assigned tags to the current content categorization system, 81.9% of publications were assigned the same category under both approaches. Alignment was highest for Education Research (96.1%) and Health Research (94.6%), whereas categories such as O*NET Evaluation (36.0%) and Other Research (38.7%) showed lower alignment, likely due to the subjective nature of those classifications and fundamental differences between the prior single-category classification process and the new multi-tag approach. Based on these results, we offer guidance and recommendations for future use of this proposed process.

Proposed Content Tagging Process for New Publications and Recommendations for Future Use

The process described in this report will be incorporated into the cyclical Reference List updating work moving forward to assign tags to new publications added to the list. In this section, we provide an overview of the process and note where human input and review are required. The process we describe here assumes that a list of publications has been identified for addition to the latest Reference List and that there is a need to assign content tags to those publications. Furthermore, it is assumed that each publication in that list has a working URL. If a working URL is not readily available, an analyst can conduct a web search to find one using the reference information (e.g., author names, title, publication outlet). If a URL cannot be found, the LLM can tag the publication using the available reference information, and the resulting tags should be reviewed for accuracy.

First, with the aforementioned list in hand, use the R code and procedures from Step 1 in this report to obtain abstracts for each publication in the list. Starting with a dataset that includes the references and URLs of the new publications, the R code can be run to obtain the abstracts. If the code fails to return an abstract for a publication, the team executing the process can default to a manual search using the link or the reference itself. This might be necessary for ProQuest publications (e.g., dissertations). To obtain abstracts for Elsevier articles, the O*NET Center can run a specific R code file when provided with a dataset of Elsevier publications that includes the reference and URLs.

Second, the team responsible for conducting the LLM tagging process can use the content category definitions from Step 2 and the final LLM prompt from Step 3 (see Appendix B) to generate provisional content category tags for each new publication in the list using an Excel dataset that includes the references for the new publications to be added along with their URLs and abstracts as input. The input/prompt should be passed through Sonnet 4.5 three times with hyperparameter settings Top $P = 0.2$ and Temperature = 1.0 to generate tags. Running the input/prompt combination will yield three Yes/No tagging judgments for each content category × publication combination.

Third, using the output described above, a provisional content category judgment for each content category × publication combination should be calculated based on the pattern of Yes/No judgments across the three runs for that combination. If at least two of the three runs are “yes,” then the content category should be provisionally designated as a “yes” for the given combination. If not, then that content category should be designated as a “no” for the given combination.

The process described above will result in 10 provisional tagging judgments for each new publication (one for each content category). At this point, the provisional tagging judgments should be subject to human review, with a focus on those tagging judgments that the current effort suggests may be most in need of further review. Specifically, based on the current work, we suggest the following review:

- Publications with “onetcenter.org” in the Link column are properly tagged as “yes” for “O*NET Programs.” If something is tagged with a “no,” it should be changed to “yes.” Then, manually check all publications tagged with “O*NET Programs” using the “AI, Machine Learning, NLP” filter on the Reports and Documents page of the O*NET

Resource Center website. If the report is listed on the webpage with that filter, it is necessary to change any “no” tags to “yes” in the Artificial Intelligence column.

- Publications with missing URLs, missing abstracts, or very short abstracts. We define very short abstracts as those with 100 words or fewer. This corresponds to the bottom quartile of abstract word counts across the 2,850 publications in this report, excluding those with missing abstracts. This threshold is somewhat arbitrary and may require revision in the future.

- Publications receiving zero content category tags. These publications can be identified by counting the number of content category tags assigned per publication and flagging those with a count of zero. Flagged publications can then be manually reviewed and assigned the best-fitting category by an analyst. Note that such cases were rare in the current tagging process.

- The following content categories: (a) O*NET Evaluation, (b) Other Field-Specific Content, (c) Application and Product Development, (d) Career and Vocational Content, and (e) Industrial-Organizational Psychology Content. As noted in Step 6, the extent of agreement between LLM and SME-assigned content tags varied across content categories. O*NET Evaluation consistently showed low agreement with SMEs across multiple steps, likely because the tagging decision for this category depends on information that is typically not available in the publication’s reference or abstract (e.g., whether the publication’s primary focus is evaluating O*NET’s psychometric properties). A single human reviewer should do a brief inspect all publications tagged “yes” for this category and consider supplementing with a brief manual check of each publication. Other Field-Specific Content also showed lower agreement with SMEs, which is partly expected given that this is a residual category and publications tagged with it may require additional judgment. Application and Product Development showed improvement after the definition was revised between Steps 3 and 4. However, inclusion rates remained lower than most other categories, likely because tagging decisions for this category may depend on specific details about an application’s design or development that are not captured in a title or brief abstract. Lastly, Career and Vocational Content and Industrial-Organizational Psychology Content showed the lowest agreement between the LLM and human SMEs (76.1% each), regardless of abstract word count. Our review of disagreements in these categories generally supported the LLM’s judgments, but human oversight remains advisable. We also recommend reevaluating the need for the brief manual checks in these categories in the future to determine whether they remain necessary.

Upon completing the LLM tagging and human review process described above, the new set of publications is complete with the tagging information for each publication. The tagging information in this case consists of 10 columns of data, one for each content category. Each column takes on values of Yes or No, indicating whether the given content category applies to the given publication, based on the process above (post-human review).

Additional Considerations for Future Use

The process described in this report represents a meaningful improvement over the current human-driven, single-category classification approach. The shift from a classification approach to a multi-tag content tagging approach, combined with the use of an LLM, offers greater efficiency and consistency compared to a human-based approach and aligns well with tagging

judgments made by human SMEs. At the same time, several considerations are important when implementing and refining this process in future cycles. We organize these recommendations into five areas.

Abstract Extraction Requires Meaningful Time and Planning

Identifying missing abstracts that could not be obtained via executing scripts was a time-intensive part of the current process. Although our R code successfully extracted abstracts for about 65% of publications automatically, reaching an 89% success rate required manual review, including manual searches, a targeted Elsevier API query, and manual extraction of ProQuest dissertations. Future users of this process should plan for this part of the process, particularly for publications in categories with historically lower URL and abstract availability (e.g., Application Development, O*NET Initiatives and Development; see Appendix A). Allocating time for abstract retrieval before the tagging step is strongly recommended. Expanding or improving the automated extraction code (e.g., better handling of publisher-specific access restrictions) could reduce this burden over time.

Data Completeness Meaningfully Affects Tagging Quality

The results in Steps 4 through 7 consistently showed that the quality and quantity of available data (i.e., the presence and length of abstracts) impacted tagging accuracy and the number of tags assigned. Publications with longer, more detailed abstracts showed stronger agreement between the LLM and SMEs (90.2% for high word count vs. 83.7% for low word count), and publications with missing abstracts received fewer tags overall. For publications with missing URLs or abstracts, we recommend additional human review of LLM-assigned tags. A practical approach would be to automatically flag these publications in the final dataset (e.g., with a data completeness indicator) so reviewers can prioritize them. We were unable to identify a universal word count threshold below which abstracts should be flagged for review, as this varied across categories. We recommend that future users inspect the distribution of abstract lengths within each cycle to establish a reasonable review threshold.

Revisiting the Tagging Process

LLMs and the broader AI landscape evolve rapidly. The Sonnet 4.5 model selected for this effort performed well under the conditions tested, but future model updates or entirely new models may offer improvements in accuracy. That said, the level of alignment we observed between LLM tags and human judges with Sonnet 4.5 was high. As such, the more pressing issue is updating the process above if Sonnet 4.5 is deprecated. Should Sonnet 4.5 not be available for a future tagging effort, an alternative LLM will need to be identified and evaluated for use in the process above. At a minimum, we would envision the following evaluation process:

- Identify one or more LLMs to consider as replacements for Sonnet 4.5 in the tagging process.
- Take the 180 publications used in Step 6 as an evaluation set and re-run the LLM-based tagging process using the replacement model. Compare tagging judgments based on the replacement LLM to judgments made by Sonnet 4.5 (available for the 180 publications from Step 6) and SME judgments (also available for the 180 publications from Step 6). This comparison should focus on whether the replacement LLM achieves a

level of agreement with SME judgments that meets or exceeds that of the Sonnet 4.5 process.

- Take the 2,850 publications used in Step 7 as an evaluation set and re-run the LLM-based tagging process using the replacement model. Compare tagging judgments based on the replacement LLM to judgments made by Sonnet 4.5 (available for the 2,850 publications from Step 7). This comparison should focus on whether the replacement LLM achieves a satisfactory level of agreement with Sonnet 4.5 judgments, evaluating whether the judgments mirror those provided by Sonnet 4.5 (and thus providing further evidence that it can serve as a suitable replacement).

Beyond being prepared to deal with LLM-related changes, content category definitions may need to be updated as the research landscape and the Center's priorities evolve. Any changes to the definition should be reflected in the prompt template before running a new tagging cycle. Changes to the prompt should also be validated on a subset of publications before being applied to the full list.

Conclusion

In conclusion, this report demonstrates that an artificial intelligence-based content tagging approach offers a viable, efficient, and scalable alternative to the traditional human-driven classification process used for the O*NET Reference List. By leveraging a structured LLM-prompt design, refined content category definitions, and a multi-run consensus strategy, the proposed method achieves high levels of consistency and strong alignment with SME judgments while substantially increasing tagging flexibility through multi-category assignment. Importantly, the introduction of a dedicated Artificial Intelligence content category enables the O*NET Program to systematically capture and highlight the growing body of research leveraging Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning (ML) in conjunction with O*NET data, products, and/or tools, ensuring the Reference List remains aligned with emerging technological trends. Although some variability remains across specific content areas and data completeness levels, the findings indicate that LLM-generated tags are not only comparable to human judgments but, in many cases, may better reflect the intent of a multi-tag framework. With appropriate human oversight focused on known edge cases, this hybrid approach provides a robust foundation for ongoing reference list maintenance and positions the O*NET Program to adapt to future advancements in AI while improving both the quality and efficiency of content tagging.

References

Allen Institute for AI. (2024). *Semantic Scholar Academic Graph API*. Semantic Scholar. Retrieved January 20, 2026, from https://www.semanticscholar.org/product/api

Amazon Web Services. (n.d.). Amazon Bedrock: Build generative AI applications and agents at production scale. https://aws.amazon.com/bedrock/

Anthropic. (2025). *Claude Haiku 4.5* [Large language model]. https://claude.ai/

Anthropic. (2025). *Claude Opus 4.5* [Large language model]. https://claude.ai/

Anthropic. (2025). *Claude Opus 4.6* [Large language model]. https://claude.ai/

Anthropic. (2025). *Claude Sonnet 4.5* [Large language model]. https://claude.ai/

arXiv. (2024). *arXiv API access*. Cornell University. https://info.arxiv.org/help/api/index.html

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2025). *rcrossref: Client for various CrossRef APIs* [Computer software]. rOpenSci. Retrieved January 20, 2026, from https://doi.org/10.32614/CRAN.package.rcrossref

Crossref. (2024). *REST API documentation*. Crossref. Retrieved January 20, 2026, from https://www.crossref.org/documentation/retrieve-metadata/rest-api/

Elsevier. (2024). *Elsevier developer portal*. Elsevier B.V. Retrieved January 30, 2026, from https://dev.elsevier.com/

Human Resources Research Organization (2025). M*eet RRObot: HumRRO's secure, proprietary AI platform.* https://www.humrro.org/corpsite/article/meet-rrobot-humrros-secure-proprietary-ai-platform/

Meta. (2025). *Llama 4 Maverick (17Bx128E)* [Large language model]. Hugging Face. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E

National Bureau of Economic Research. (2026). *NBER working papers and chapters metadata*. National Bureau of Economic Research. Retrieved January 20, 2026, from https://www.nber.org/research/data/nber-working-papers-and-chapters-metadata

National Center for O*NET Development (2025). *O*NET 30.0 Database.* O*NET Resource Center. Retrieved January 1, 2026, from https://www.onetcenter.org/database.html

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv. https://doi.org/10.48550/arXiv.2205.01833

Appendix A: Summary of Success Rates Across Publication Types and Current Content Categories

Table A.1. Summary of Abstract Extraction Results by Publication Type

Publication Type	Abstract Found	Abstract Missing	URL Missing
Blog Posts	0	6	1
Books and Book Chapters	172	12	13
Journal Articles	1,322	19	6
Magazine Articles	2	1	0
Miscellaneous	11	13	0
Newspaper Articles	9	6	0
Papers and Presentations	123	13	82
Research and Technical Reports	596	29	16
Theses and Dissertations	171	2	4
Unpublished Manuscripts	36	0	0
User Guides and Training Materials	46	11	2
Webinars	2	0	1
Websites	47	60	16

Note. We distinguish between references where the abstract was missing and the URL was missing for more precise missing data counts.

Table A.2. Summary of Abstract Extraction Results by Current Content Categories

Current Content Category	Abstract Found	Abstract Missing	URL Missing
Application Development	162	63	33
Career and Vocational Research	276	11	5
Education	166	11	4
Health Research	257	4	0
Industrial/Organizational Research	484	9	40
Labor Market Research	874	38	19
O*NET Evaluation	52	7	16
O*NET Initiatives and Development	211	24	22
Other Research	55	5	2

Note. We distinguish between references where the abstract was missing and those where the URL was missing to obtain more precise counts of missing data.

Appendix B: Final Prompt Template for LLM Content Tagging

Figure B.1. Final Prompt Template for LLM Content Tagging

Your task is to assign one or more content tags to a given publication.
You will first read the publication's reference, URL, and abstract:

Publication Reference:{PublicationRef}
Publication URL:{PublicationURL}
Publication Abstract: {PublicationAbstract}

Next, you will read names and descriptions of the two sets of content tags. One set of tags regards the focal use of O*NET in the given publication and a second set regards additional publication features of the publication.

Content Tag Set 1: Focal Content Domain

1. IO (Industrial-Organizational Psychology Content): Content is focused on topics studied in the area of Industrial-Organizational Psychology examples of which include but are not limited to: job search, human resources, personnel selection/recruitment, individual assessment, motivation, leadership, job analysis, job performance and attitudes, and organizational behavior.

2. Health (Health Content): Content is focused on topics studied in the areas of medicine, public/occupational health and workplace safety examples of which include but are not limited to: assessment of medical students/professionals, requirements of medical practice or occupations, public health and safety, epidemiology, occupational health psychology, workplace injuries, illness prevention, occupational stress/strain, mental health, and worker well-being.

3. Educ (Education Content): Content is focused on topics studied in the areas of education and workforce development examples of which include but are not limited to: education/work preparedness, education-to-work transition, academic training, skills development through formal or informal learning, curriculum design, and child/adolescent development.

4. Career (Career and Vocational Content): Content is focused on topics studied in the areas of career development and vocational counseling including examples of which include but are not limited to: career counseling, vocational interests (e.g., RIASEC), work values, occupational fit, person-job fit, vocational rehabilitation, and career decision-making.

5. Labor (Labor Economics Content): Content is focused on topics studied in the area of labor economics examples of which include but are not limited to: occupational staffing patterns, skill/knowledge distribution and gaps in the labor market, wage trends, economic implications of automation on work, labor market dynamics, employment projections, and longitudinal comparisons of labor market demographics.

6. Other (Other Field-Specific Content): Content is focused on topics studied in areas other than industrial-organizational psychology, education, health, career and vocational counseling, and labor economics.

Content Tag Set 2: Additional Publication Features

7. AI (Artificial Intelligence): Publication involves use of Artificial Intelligence (AI), Natural Language Processing (NLP), or machine learning (ML) methods.

8. AppProdDev (Application & Product Development): Publication has an explicit focus on the design, development, or technical evaluation of a new application, software tool, instrument, product, or service that is built using O*NET data or tools.

9. ONET_E (O*Net Evaluation): Publication has an explicit focus on the evaluation of the quality of O*NET's products, tools, or databases and has a URL that does NOT include "onetcenter.org".

10. ONET_P (O*NET Program): Publication is conducted or commissioned by the O*NET Program or U.S. Department of Labor, or research authored by O*NET employees and contractors, pertaining to O*NET products, tools, databases, or associated components (e.g., algorithms). Includes reports, articles, and development work originating from the O*NET Resource Center or O*NET-sponsored project. Only assign this tag to a publication if its URL contains "onetcenter.org".

EXPECTED OUTPUT FORMAT:

For each content tag within each set, you will output a decision (YES or NO) and a reason (1-2 sentences explaining your decision) using XML tags.

Example structure:
AI: YES
AI_Reason: This article discusses machine learning applications with occupational data.

ONET_P: NO
ONET_P_Reason: This article's URL does not contain onetcenter.org.

And so on for all 10 content tags.

IMPORTANT FORMATTING RULES:
- Each decision tag must contain ONLY YES or NO (e.g., <AI>YES</AI> or <AI>NO</AI>)
- Each reason tag must contain 1-2 complete sentences explaining your decision
- Do not nest tags incorrectly
- Do not add extra wrapper tags like <Article> or <Response> unless specified
- Ensure all tags are properly closed with matching opening and closing tags
- Output all 10 categories in the order specified below
- Do not include any additional text outside the XML tags

CATEGORY EVALUATION GUIDELINES:

1. Do NOT indicate "YES" for a given tag UNLESS you can confidently conclude that tag applies to the publication based on the contents of its reference, abstract, or URL.

2. If the URL contains "onetcenter.org", assign a NO to all the following content areas: IO, Health, Educ, Career, Labor, Other, AppProdDev, and O*NET_E. This should override any guidance offered above.

OUTPUT REQUIREMENTS:

Output your categorization using the following XML tags in this exact order:

For the first category (Industrial-Organizational Psychology Content):
<IO>YES or NO</IO>
<IO_Reason>Your 1-2 sentence explanation here.</IO_Reason>

For the second category (Health Content):
<Health>YES or NO</Health>
<Health_Reason>Your 1-2 sentence explanation here.</Health_Reason>

For the third category (Education Content):
<Educ>YES or NO</Educ>
<Educ_Reason>Your 1-2 sentence explanation here.</Educ_Reason>

For the fourth category (Career and Vocational Content):
<Career>YES or NO</Career>
<Career_Reason>Your 1-2 sentence explanation here.</Career_Reason>

For the fifth category (Labor Market Content):
<Labor>YES or NO</Labor>
<Labor_Reason>Your 1-2 sentence explanation here.</Labor_Reason>

For the sixth category (Other Content):
<Other>YES or NO</Other>
<Other_Reason>Your 1-2 sentence explanation here.</Other_Reason>

For the seventh category (Artificial Intelligence):
<AI>YES or NO</AI>
<AI_Reason>Your 1-2 sentence explanation here.</AI_Reason>

For the eighth category (Application & Product Development):
<AppProdDev>YES or NO</AppProdDev>
<AppProdDev_Reason>Your 1-2 sentence explanation here.</AppProdDev_Reason>

For the ninth category (O*Net Evaluation):
<ONET_E>YES or NO</ONET_E>
< ONET_E_Reason>Your 1-2 sentence explanation here.</ ONET_E_Reason>

For the tenth category (O*NET Funded):
<ONET_P>YES or NO</ONET_P>
<ONET_P_Reason>Your 1-2 sentence explanation here.</ONET_P_Reason>

Note: This prompt template uses two types of formatting syntax: (1) Curly braces {} indicate variable placeholders that are dynamically replaced with specific content for a given publication reference, URL, and abstract combination to be rated (e.g., {PublicationRef} is populated with the reference to be tagged; and (2) Angle brackets <> define XML tags that structure the LLMs output format, enabling programmatic extraction of the YES or NO tag for each category domain and the given publication.

Table B.1. Variable Input Fields

Input Field	Description	Cell Content
PublicationRef	Publication Reference	Includes the publication's reference with information such as the author name(s)/organization, publication year, title, and source information.
PublicationURL	Publication URL	Includes the publication's URL.
PublicationAbstract	Publication Abstract	Includes the publication's abstract, summary, or executive summary.

Appendix C: Instructions for Human SMEs

Figure C.1. Instructions for Human SME Tagging

NOTE: Some article abstracts/summaries get cut off in the cells. Make sure to check for this occurring!

Your task is to tag publications into one or more of 10 content areas. It is possible that publications will be tagged into more than one content area.

The Excel file provided includes the publication's reference, URL, and abstract. You will use this information to tag the publications into the content areas. *You should NOT visit the URL for any publication. Only use the information provided to tag the publication. For some publications, there is insufficient information (e.g., missing abstracts). For these, tag the publication based on the information available.*

Do NOT indicate "YES" for a given tag UNLESS you can confidently conclude that tag applies to the publication based on the information available (i.e., reference, abstract, or URL).

Below are names for the 10 content categories and their descriptions, which are split up into two sets. *Note that these are CONTENT categories that might encompass more than just RESEARCH on a topic.*

Content Tag Set 1: Focal Content Domain

1. IO (Industrial-Organizational Psychology Content): Content is focused on topics studied in the area of Industrial-Organizational Psychology examples of which include but are not limited to: job search, human resources, personnel selection/recruitment, individual assessment, motivation, leadership, job analysis, job performance and attitudes, and organizational behavior.

2. Health (Health Content): Content is focused on topics studied in the areas of medicine, public/occupational health and workplace safety examples of which include but are not limited to: assessment of medical students/professionals, requirements of medical practice or occupations, public health and safety, epidemiology, occupational health psychology, workplace injuries, illness prevention, occupational stress/strain, mental health, and worker well-being.

3. Educ (Education Content): Content is focused on topics studied in the areas of education and workforce development examples of which include but are not limited to: education/work preparedness, education-to-work transition, academic training, skills development through formal or informal learning, curriculum design, and child/adolescent development.

4. Career (Career and Vocational Content): Content is focused on topics studied in the areas of career development and vocational counseling including examples of which include but are not limited to: career counseling, vocational interests (e.g., RIASEC), work values, occupational fit, person-job fit, vocational rehabilitation, and career decision-making.

5. Labor (Labor Economics Content): Content is focused on topics studied in the area of labor economics examples of which include but are not limited to: occupational staffing patterns, skill/knowledge distribution and gaps in the labor market, wage trends, economic implications of automation on work, labor market dynamics, employment projections, and longitudinal comparisons of labor market demographics.

6. Other (Other Field-Specific Content): Content is focused on topics studied in areas other than industrial-organizational psychology, education, health, career and vocational counseling, and labor economics.

Content Tag Set 2: Additional Publication Features

7. AI (Artificial Intelligence): Publication involves O*NET's products, tools, or database in conjunction with Artificial Intelligence (AI), Natural Language Processing (NLP), or machine learning (ML) methods. This includes development or evaluation of AI, NLP, or ML-driven applications that utilize O*NET content or data, or research employing AI, NLP, or ML methods with O*NET content or data.

8. AppProdDev (Application & Product Development): Publication involves the development of applications, software, instruments, products, tools, or services using O*NET's products, tools, or databases.

9. ONET_E (O*Net Evaluation): Publication involves the evaluation of the quality of O*NET's products, tools, or databases for the purpose of documenting validity, reliability, psychometric characteristics, statistical properties, or content characteristics.

10. ONET_P (O*NET Program): Publication is conducted or commissioned by the O*NET Program or U.S. Department of Labor, or research authored by O*NET employees and contractors, pertaining to O*NET products, tools, databases, or associated components (e.g., algorithms). Includes reports, articles, and development work originating from the O*NET Resource Center or O*NET-sponsored project. Only assign this tag to a publication if its URL contains "onetcenter.org".

Expected Output: For each category, you will output a decision (YES or NO).

Remember: Do NOT indicate "YES" for a given tag UNLESS you can confidently conclude that tag applies to the publication based on the information available (i.e., reference, abstract, or URL).