REALM: A Dataset of Real-World LLM Use Cases

Anonymous ACL submission

Abstract

Large Language Models (LLMs), such as the GPT series, have driven significant industrial applications, leading to economic and societal transformations. However, a comprehensive understanding of their real-world applications remains limited. To address this, we introduce **REALM**, a dataset of over 94,000 LLM use cases collected from Reddit and news articles. **REALM** captures two key dimensions: the diverse applications of LLMs and the demographics of their users. It categorizes LLM applications and explores how users' occupations relate to the types of applications they use. By integrating real-world data, **REALM** offers insights into LLM adoption across different domains, providing a foundation for future research on their evolving societal roles.

1 Introduction

Large Language Models (LLMs) have revolutionized AI with advanced natural language understanding and reasoning capabilities (Guo et al., 2025; Jaech et al., 2024; Team et al., 2024). These advancements have spurred new applications and transformed workflows across various industries. As LLM capabilities expand, researchers in fields like social science, economics, and human-computer interaction (HCI) are examining their social implications, anticipating both benefits and disruptions, and addressing risks tied to their rapid adoption.

Currently, there remains limited understanding of how LLM are deployed and utilized in real-world scenarios despite widespread adoption. Existing studies generally fall into two primary categories. The first uses occupation-activity typologies to measure LLM's influence on various tasks, aiming to assess task-exposure rates and impact on occupations (Eloundou et al., 2023; Felten et al., 2018; Brynjolfsson et al., 2018; Webb, 2020). However, these studies often lack empirical grounding, weakening their validity. The second category analyzes large datasets from social media and other public sources to evaluate sentiment and trends around LLM applications (Leiter et al., 2023; Miyazaki et al., 2024; Koonchanok et al., 2024; Naing and Udomwong, 2024), but these studies are hindered by a lack of comprehensive taxonomies and rely on simplistic methods, such as keyword filtering, leading to low precision and recall.

To conduct an empirical study with a comprehensive taxonomy, we present a new dataset **Re**al-world **A**pplications of Large **L**anguage **M**odels (**REALM**), documenting real-world use cases of LLMs. **REALM** comprises over 94,000 instances, including 15,000 sourced from Reddit discussions and 79,000 from news articles, spanning from June 2020 (the release date of GPT-3) to December 2024. Our analysis focuses on two primary dimensions: the objectives of LLM use case based on a comprehensive taxonomy (Theofanos et al., 2024), and the occupations of current or potential end-users, categorized using the O*NET database (Handel, 2016). Our contributions are threefold: (1) we employ a systematic and reliable taxonomy to link use cases to occupations (2) we provide initial statistical analyses derived from the dataset, demonstrating its potential for analyzing trends, informing policies, and supporting cross-domain research (3) we provide accessible via an interactive dashboard and an API for structured data processing.

2 Related Work

Studying the impact of LLM-based applications has emerged as a prominent research topic, including both structured analyses based on existing occupation-activity taxonomies and empirical investigations utilizing social media big data.

Some research examines the societal and economic impacts of LLMs through structured taxonomies, employing either GPT-based or human

annotations to assess potential LLM usage. For example, (Eloundou et al., 2023) estimates that 19% of U.S. workers could have 50% or more of their tasks exposed to LLMs using a task-exposure rubric derived from the O*NET database. This study builds on a broader body of work that quantifies AI exposure rates (Felten et al., 2018; Webb, 2020; Brynjolfsson et al., 2018). However, these analyses often rely on predefined rubrics and expert opinions for annotation, which can introduce biases and fail to reflect real-world dynamics.

In contrast to structured taxonomy-based approaches, empirical research leverages(Naing and Udomwong, 2024) social media data to explore real-world applications of LLMs. For instance, (Leiter et al., 2023) analyzes sentiment trends and employs topic modeling to investigate the impacts of LLMs across domains such as education and technology, but utilizing a highly simplified taxonomy. (Miyazaki et al., 2024) and (Koonchanok et al., 2024) incorporate occupation-level analyses based on O*NET and Indeed to examine how different professions perceive and interact with LLMs. (Miyazaki et al., 2024) further analyzes user-provided prompts and applies topic modeling to understand user interactions with LLMs. However, these studies often depend on keyword filtering or hashtag-based data collection methods, which are imprecise and fail to provide a comprehensive taxonomy due to the necessity of omitting many keywords that may have multiple meanings. Beyond social media, other research has examined LLM adoption through alternative data sources with varying focuses including scientific literature (Gao and Wang, 2024) and large-scale user interaction logs with LLMs (Zheng et al., 2023).

Building upon the aforementioned works, **REALM**introduce a taxonomy-driven approach coupled with a semantically rich filtering process. We systematically link LLM applications to specific occupations and use cases using both social media and news platforms with high-performing data collection and cleaning pipeline. This enables a precise and nuanced understanding of how different professional sectors are integrating LLMs into their workflows, offering a robust foundation for future research.

3 Dataset Construction

The construction of the **REALM** follows a structured pipeline for collecting and processing high-quality data, as illustrated in Figure 1. A dedicated

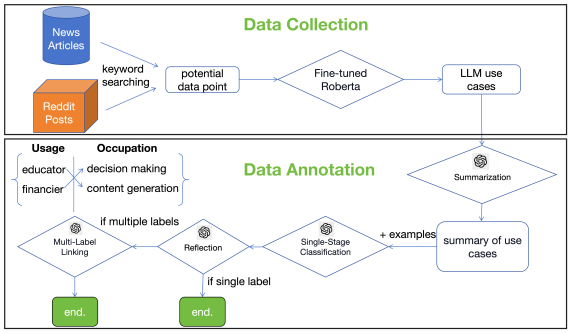


Figure 1: REALM Dataset Construction Pipeline

dashboard **https://realm-e7682.web.app/** is also built to facilitate data exploration and visualization. Construction details are presented below.

Data Collection Data in **REALM** is sourced from Reddit (via Academic Torrents[1]) and news articles (retrieved using NewsAPI[2]) spanning from June 2020 (when GPT-3 was released) to December 2024. This process narrows down billions of datapoints to a curated dataset of 94,000 entries, containing 15,000 Reddit posts and 79,000 news articles. The real-time, user-generated content from Reddit provides diverse perspectives, while NewsAPI offers professionally curated articles, ensuring diversity.

The data collection process begins with keyword-based extraction to isolate LLM-related content, using a comprehensive list of model names obtained from GitHub[3]. Then we perform essential data cleaning to ensure the quality of the collected content. This includes removing duplicates, unusable tweets, stop words, and non-textual elements, ensuring the data is well-prepared for further processing. We employ a fine-tuned RoBERTa model to filter irrevelant data, ensuring relevance by identifying descriptions of LLM use cases or potential applications. This filtering process prioritizes recall over precision to include all pertinent data, achieving high recall scores of 0.94 for news articles and 0.95 for Reddit posts after human validation. Following the initial filtering, a four-stage annotation pipeline, detailed in Section 3, is applied to further refine the dataset by enhancing precision.

Data Annotation The annotation for **REALM** involves labeling LLM use case types and occupa-

[1]Academic Torrents is a distributed system for sharing academic datasets and resources. More information is available at https://academictorrents.com

[2]https://newsapi.org/

[3]https://github.com/Hannibal046/Awesome-LLM

LLM Use Categorization				Occupation Categorization			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
Content Creation	0.75	0.83	0.79	Management	0.87	0.62	0.72
Content Summarization	0.70	0.88	0.78	Business and Financial Operations (Bus & Finance)	0.82	0.70	0.76
Decision Making	0.71	0.74	0.72	Computer and Mathematical (Comp & Math)	0.72	0.86	0.78
Detection	0.68	0.73	0.70	Architecture and Engineering (Arch & Eng)	0.80	0.75	0.77
Digital Assistants	0.69	0.79	0.74	Life, Physical, and Social Science (Life & Science)	0.73	0.81	0.77
Discovery	0.74	0.78	0.76	Community and Social Service (Comm & Social)	0.69	0.82	0.75
Image Analysis	0.80	0.85	0.82	Legal	0.78	0.85	0.81
Information Retrieval	0.79	0.78	0.78	Educational Instruction and Library (Edu & Library)	0.80	0.79	0.79
Personalization	0.70	0.73	0.71	Arts, Design, Entertainment, Sports, and Media (Art & Media)	0.81	0.77	0.79
Prediction	0.78	0.83	0.80	Healthcare Practitioners and Support (Health & Care)	0.70	0.79	0.74
Process Automation	0.68	0.89	0.77	Sales and Related (Sales)	0.77	0.75	0.76
Recommendation	0.72	0.74	0.73	Office and Administrative Support (Office & Admin)	0.76	0.82	0.74
Robotic Automation	0.79	0.83	0.81	Production	0.73	0.81	0.77
Vehicular Automation	0.82	0.86	0.84	Others	0.64	0.80	0.71

Table 1: Performances of our pipeline for LLM Use categorization and Occupation categorization tasks separately.

tions. We use (1) LLM Use taxonomy(Theofanos et al., 2024) and (2) the Occupation taxonomy. Further details on the taxonomies are provided in Appendix A.1. To simplify analysis, 9 occupation groups with less than 1% frequency are merged into "Others", and healthcare-related roles are consolidated into "Healthcare" (Koonchanok et al., 2024) to obtain a total of 14 occupations. We utilize a four-module annotation pipeline using GPT-4o-mini to streamline the process, consisting of Summarization, Classification, Reflection and Multi-Label Linking module. Further information about these modules can be found in Appendix A.2. Table 1 presents the precision, recall, and F1 score of the four-stage pipeline on a validation dataset of 1000 data points.

Performance Validation To evaluate the performance of the whole pipeline, two expert annotators labeled a total of 3,000 data points. See Appendix A.3 for more information.

4 Analysis of REALM Dataset

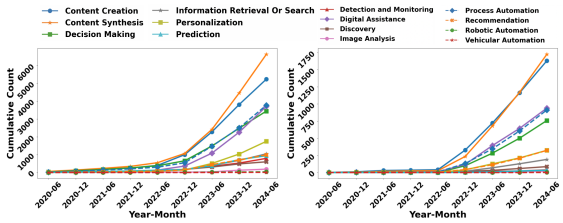
In this section, we present preliminary analysis results derived from the **REALM** dataset. As depicted in Figures 2a and 2b, the adoption and application of LLMs experienced a significant surge from between 2022 and 2023, coinciding with the release of GPT-3.5 in November 2022.

Figures 2a illustrate the number of LLM use cases categorized by application type within News and Reddit sources, respectively. These figures reveal that Content Creation and Content Synthesis are the most prevalent LLM use cases, including writing assistance, summarization, and code generation. Additionally, applications related to Decision-making, Digital Assistance, and Process Automation have experienced substantial growth.

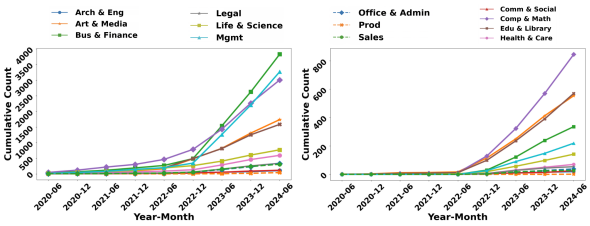
The Decision-making category includes activities such as recommending business strategies, optimizing marketing campaigns, and suggesting financial investment strategies. Process Automation involves streamlining routine and repetitive tasks. The increasing interest in these use cases indicates that users are perceiving LLMs as intelligent and reliable collaborators, thereby building trust in their capabilities to help daily work. This trend signifies that LLMs are evolving beyond their traditional role as generative AI, actively supporting and enhancing industrial workflows. However, certain areas such as Robotics Automation, Recommendation Systems, and Vehicular Automation have not yet extensively integrated LLM technologies.

Figure 2b illustrates the utilization of LLMs across various occupations. In News sources, discussion were initially dominated by professionals in Computer Science and Mathematics. However, over time, people in Business and Management have assumed a more prominent role in these discussions. In contrast, as Reddit has a more technically inclined user base, its discussion is consistently dominated by individuals from computer-related fields. But there has been a growing interest in LLMs among professionals in Art & Media, Product Development and Education.

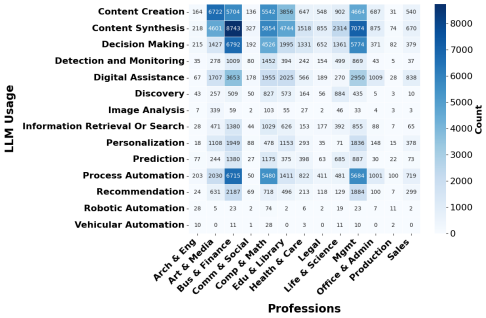
Figures 2c and 2d demonstrate varying levels of applicability and adoption of LLM technologies across different professional domains: individuals in Art, Business, Computer Science, and Management are the most engaged with LLMs across both News and Reddit sources. Their primary activities include content creation, content synthesis, decision-making, digital assistance, and process automation. Additionally, professionals in Computer Science, Mathematics, Physical Sciences, and Social Sciences are notably involved in discovery-oriented tasks, highlighting the potential of LLMs to advance scientific research. In contrast, occupations in the legal and sales sectors, as well as roles centered on manual or operational tasks (food preparation, transportation, and farming) exhibit



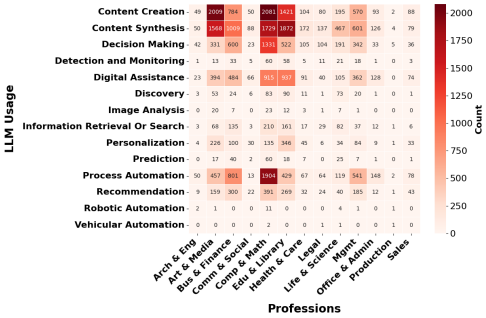
(a) News Article(left) and Reddit Posts(right) Count by LLM Use



(b) News Article(left) and Reddit Posts(right) Count by Occupation



(c) LLM Use for Different Professions: News Articles



(d) LLM Use for Different Professions: Reddit Posts

limited engagement with LLMs.

5 Discussion and Conclusion

5.1 Potential Usage of REALM

Analyzing Occupational Usage Patterns. **REALM** facilitates an in-depth analysis of LLM use cases across various occupations, including education, healthcare, and finance. For example, within the education sector, LLMs are frequently discussed for tasks like "lesson planning, quiz creation, rubric design, and administrative support,"[4] helping teachers reduce preparation time and dedicate more attention to student interactions. Some highlight the need to teach students "how to ask the important questions," as effectively using AI requires clear problem formulation and critical thinking skills that many students still need to develop. Some discussed academic integrity as "students now rewrite AI outputs with multiple LLMs to make it appear human,"[5] illustrating new challenges. New education methods are developed by redesigning assignments, such as requiring in-class writing or creating AI-focused tasks to foster responsible usage. This detailed information complements frameworks like *Particip-AI* (Mun et al., 2024), which evaluates AI risks and benefits through workshops and surveys, by providing large-scale, real-world cases.

[4] https://www.reddit.com/r/edtech/comments/1d676dc/genai_and_education/

[5] https://www.reddit.com/r/Teachers/comments/1d6hnqe/how_to_detect_if_ai_wrote_an_essay_when_the/

Studying Bias and Fairness in LLM Applications. **REALM** serves as a valuable resource for examining bias and fairness in LLM use cases. Unlike existing datasets that are designed for specific tasks using counterfactual inputs, biased prompts, or masked token evaluation (Gallegos et al., 2024), the biases present in **REALM** emerge naturally from user-generated content. For instance, biases like religious stereotyping and racial profiling are directly reported by users during their engagement with LLM-based systems.

Tools for Developers: Building Robust and User-Aligned LLMs. Our dataset also provides practical value to LLM developers in two key areas: enhancing robustness and improving user alignment. **REALM** offers information about how LLMs are utilized in real-world contexts, thus allowing developers to anticipate trends and proactively design protection mechanisms for evolving vulnerabilities. Additionally, **REALM** captures firsthand user feedback, providing insights into usability issues and emerging needs.

5.2 Conclusion

This paper introduces **REALM**, a novel dataset for analyzing real-world LLM use cases across professional domains. It contains data from Reddit and news articles, providing a comprehensive view of LLM usage while acknowledging limitations such as uneven representation. We aim to advance the understanding of LLMs' evolving societal roles by this public dataset.

Limitations

While **REALM** provides valuable insights into the utilization of LLMs across various professions and tasks, it is subject to several limitations. Firstly, the dataset is exclusively derived from Reddit and news articles, excluding other significant platforms such as Twitter, Facebook, and LinkedIn. Secondly, the inherent characteristics of the selected sources contribute to skewed distributions within the dataset. Reddit's user base is predominantly technical, and news articles often emphasize business-related topics, leading to an overrepresentation of fields such as Computer Science and Business, while less prominent sectors like Transportation and Farming remain underrepresented. Addressing these biases and establishing accurate denominators for normalization purposes lie beyond the current scope of **REALM**.

References

E. Brynjolfsson, T. Mitchell, and D. Rock. 2018. What can machines learn, and what does it mean for occupations and the economy?

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models.

E. W. Felten, M. Raj, and R. Seamans. 2018. A method to link advances in artificial intelligence to occupational abilities.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey.

Jian Gao and Dashun Wang. 2024. Quantifying the use and potential benefits of artificial intelligence in scientific research.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Michael J Handel. 2016. The o* net content model: strengths and limitations. *Journal for Labour Market Research*, 49(2):157–176.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ratanond Koonchanok, Yanling Pan, and Hyeju Jang. 2024. Public attitudes toward chatgpt on twitter: Sentiments, topics, and occupations.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. Chatgpt: A meta-analysis after 2.5 months.

Kunihiro Miyazaki, Taichi Murayama, Takayuki Uchiba, Jisun An, and Haewoon Kwak. 2024. Public perception of generative ai on twitter: An empirical study based on occupation and usage.

Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2024. Particip-ai: A democratic surveying framework for anticipating future ai use cases, harms and benefits.

Shwe Zin Su Naing and Piyachat Udomwong. 2024. Public opinions on chatgpt: An analysis of reddit discussions by using sentiment analysis, topic modeling, and swot analysis.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Mary Theofanos, Yee-Yin Choong, and Theodore Jensen. 2024. Ai use taxonomy: A human-centered approach. Technical report.

M. Webb. 2020. The impact of artificial intelligence on the labor market. Working paper, Stanford University.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset.

A Appendix

A.1 Taxonomy Selection

For data labeling in **REALM**, we select two taxonomies to provide a comprehensive understanding of the use cases. The LLM Use taxonomy categorizes LLM applications based on human goals and outcomes, while the Occupation taxonomy, adapted from the O*NET database, maps these use cases to specific job roles.

The LLM Use taxonomy, as proposed in (Theofanos et al., 2024), categorizes LLM applications based on human goals and outcomes. This taxonomy is selected due to its strong foundation in real-world AI use cases derived from industry, making it highly relevant for our dataset. It was developed

through a rigorous process of refining common AI use activities, balancing abstraction with specificity, and validating its coverage through multiple AI research repositories, including NIST's AI Community of Interest (COI) Project Catalogue. Although the taxonomy is not specific to technological techniques, its structure is well-suited for categorizing large language model relevant cases. We make minor adjustments to better fit our data labeling task, ensuring it effectively classifies the use cases in our dataset.

For the Occupation taxonomy, we opt for the O*NET database [6], which is widely recognized for its authority and regularly updated to ensure its timeliness. While we consider alternative resources, such as Indeed, we find that their taxonomy levels do not align well with our data, leading to too much variability and making them unsuitable for the labeling task. O*NET provides additional granularity in occupation levels, however, when we attempt to apply this more detailed classification, we encounter difficulties in extracting sufficient information from reddit posts or news articles to support more specific occupational labels. This resulted in poor labeling consistency, and many finer occupation categories were not adequately covered. A more detailed approach could potentially require case studies, workshops, or additional resources, which we leave for future work.

A.2 Details of the Four-Module Annotation Pipeline

Below are the descriptions of the four modules in the data annotation phase shown in Figure 1.

Summarization Module generates summaries for each datapoint, which include news articles or Reddit posts along with their top five comments. The summaries focus on describing the LLM use case and the user professions. Additionally, this step filters out irrelevant cases from the high-recall dataset.

Classification Module leverages GPT-4o-mini with few-shot examples to categorize datapoints. Based on task definitions, category explanations, representative examples, and the summaries produced by the Summarization Module. The module also outputs a confidence score indicating the LLM's certainty regarding the generated label, facilitating the identification of classifications that may require further review.

[6] https://www.onetonline.org/

Reflection Module reassesses and refines classifications for datapoints that receive low-confidence scores from the Classification Module. By reflecting on the previously assigned label using the title, summary, and initial classification, the Reflection Module generates a revised label.

Multi-Label Linking Module establishes connections between occupations and LLM usage types through pairwise mappings based on contextual information. This module determines which usage types correspond to specific occupations, enabling a detailed understanding of LLM adoption across various professional domains.

A.3 Data Annotation Details

The data annotation process consists of two tasks that serve the data collection and data annotation phase separately. Two expert annotators collaboratively labeled the data, achieving a high inter-annotator agreement (IAA).

The first task involves annotating 2,000 data points selected from the set of Reddit posts and news articles after keyword filtering. These data points were categorized into "Use Case" and "Non-Use Case" labels to train a RoBERTa model, which acted as an automated classifier, thus optimizing resource use and time efficiency. The IAA for this task reached 0.89, and the resulting RoBERTa model achieved high recall scores, as discussed in Section 3.

The second task focuses on a subset of 1,000 data points, selected from those identified as 'Use Case' after the RoBERTa model's filtering process, with 500 points from news articles and 500 from Reddit posts. The data points are labeled for "LLM Use Categorization" and "Occupation Categorization" to validate the accuracy of our data annotation pipeline. This task yields IAA scores of 0.87 for LLM Use and 0.85 for Occupations.

Through expert supervision and a context-aware pipeline, the quality of our dataset is ensured.

A.4 Prompts

In the data annotation phase, we use GPT-4o-mini for labeling. Below are the prompts for Reddit posts (the prompts for news articles are similar, with minor adjustments made to accommodate the data characteristics and structure). The content within the curly braces in the prompts represents the input for different data points. We only include the prompts for the Summarization and Single-

Stage Classification modules, as they are most critical for understanding and applying our taxonomy.

The Summarization Module not only serves to generate summaries but also improves the precision score of our LLM use case identification (compensating for the sacrifice made by the RoBERTa model to achieve a higher recall score). The definition of use cases in Listing 1 has been independently verified and achieved a precision score of 98% on the validation set.

```
SYSTEM
You are an expert in analyzing and summarizing use cases of
      large language models. You will be provided with the
      title, content and top 5 comments of a reddit post.
      Based on the human goals and outcomes, summarize in
      detail how LLM is used and who the end users are, paying
       attention to their profession or role (including
      comments). If the human goals or occupation information
      of end users is not clear, just say you don't know. Your
       summary should integrate both the usage and occupation
      information clearly in at least 3 sentences.

If you think this is not a use case of large language model,
      just return the phrase 'not a use case' without other
      words.
These two senerios should all be considered a use case:
1.Cases that describe or mention large language models being
      directly used, including through APIs, or products
      integrated with large language models being used by end-
      users.
2.Cases that include the introduction, release, or
      advertisement of products/software/tools/systems
      developed on top of large language models, mostly
      speculative or experimental and may not yet have end-
      user adoption.

USER
Subreddit:  {subreddit}
Title: {title}
Content: {content}
Top_comments: {top_comments}

Please provide a summary explaining both how the LLM is used
      and detailing the occupation or roles of the end users
      who benefit from or are expected to use the LLM. Use at
      least 3 sentences, do not make assumptions. Or if you
      think this is not a use case of large language model,
      return the phrase 'not a use case' without other words.
```

Listing 1: Prompts for Summarization Module

In Listing 2, the definition of each LLM use category is modified based on descriptions in the original paper to achieve better classification results.

```
SYSTEM
You are an expert assistant specializing in analyzing and
      categorizing how LLM is used in a reddit post with its
      top comments. Here is a large language model use case
      summary of a reddit post and its comments focusing on
      end-users and how users use them. Based on the human
      goals and outcomes of the usage process of LLM stated in
       the summary, please use the following taxonomy to
      categorize them. Analyze based solely on the text
      provided, without making any assumptions, such as
      infering usage for certain senarios or occpations.

1. Content Creation: The LLM generates new and original
      content such as narrative, software code, synthetic data.
       This involves creativit where the LLM acts as a creator
       or producer.

2. Content Synthesis: The LLM combines and/or summarizes parts,
       elements, or concepts into a coherent whole. This
      involves integrating diverse pieces of information to
      form a complete and comprehensive output.

3. Decision Making: The LLM assists in selecting a course of
      action from among possible alternatives to arrive at a
      solution. This involves evaluating options, predicting
      outcomes, and recommending the best course of action
      based on data analysis and algorithms.

4. Detection and Monitoring: The LLM identifies the existence
      or presence of something through careful analysis and
      observation, and monitors processes, quality, or states
      over time. This involves continuous or periodic
      assessment to gain insights and ensure standards.

5. Digital Assistance: The LLM acts as a personal assistant,
      understanding and responding to commands and questions,
      and performing requested tasks in a conversational
      manner. This involves interaction with users through
      natural language processing and task execution.

6. Discovery: The LLM aids in finding, recognizing, or
      uncovering something new. This involves exploratory
      analysis and innovation to reveal new insights, patterns,
       or entities.

7. Image Analysis: The LLM processes and interprets textual
      descriptions related to images to extract meaningful
      information. This involves recognizing patterns, objects,
       and relevant details within visual data.

8. Information Retrieval Or Search: The LLM efficiently finds
      and retrieves relevant information on specific topics of
       interest. This involves searching large databases,
      documents, and online resources to provide accurate and
      relevant data.

9. Personalization: The LLM customizes content, services, or
      products to match an individual's preferences, behaviors,
       or characteristics. This involves tailoring user
      experiences based on interaction history and personal
      data.

10. Prediction: The LLM forecasts the likelihood of future
      outcomes based on historical data and trends. This
      involves using statistical models and machine learning
      algorithms to predict future events or behaviors.

11. Process Automation: The LLM automates repetitive tasks,
      removes bottlenecks, reduces errors, and increases
      efficiency. This involves streamlining processes and
      saving time by automating routine activities.

12. Recommendation: The LLM suggests or proposes a manageable
      set of viable options to aid decision-making. This
      involves analyzing user preferences and context to offer
       personalized recommendations.

13. Robotic Automation: The LLM is used to assist or control
      physical machines to automate, improve, or optimize
      tasks. This involves combining LLM capabilities with
      robotics for complex physical tasks.

14. Vehicular Automation: The LLM supports or manages the
      automation of transportation systems. This involves
      using LLMs in conjunction with autonomous vehicles for
      navigation, safety, and decision-making.

15. Unknown: This category refers to cases where the LLM's
      role or specific functionality is ambiguous or not
      provided. While it's clear that an LLM is used, there is
       insufficient information to classify its use into any
      of the known categories.  Notice, This category is
      mutually exclusive with others.

Here are two examples you can refer to:
Example 1: xxx; Reason: xxx
Example 2: xxx; Reason: xxx

USER
Subreddit: {subreddit}
Article summary: {summary}
The output format should be as follows, do not generate other
      extra things:
category: list the categories fit the datapoint, separate them
       with "/", such as : Content Synthesis/Personalization.
explanation: explain the reason for your classification.
confidence: set the confidence to 1 if you are confident in
```

```
      your judgment. If the information is insufficient or you
       cannot make a reliable judgment, set the confidence to
      0.
```

Listing 2: Prompts for Single-Stage Classification Module (LLM Use)

In Listing 3, the items listed under "this group includes" are based on the finer-grained occupational divisions from the O*NET database.

```
SYSTEM
You are an expert assistant specializing in analyzing and
      categorizing the current or potential end-user
      occupation characteristics. Here is a large language
      model use case summary of a news article focusing on end-
      users and how users use them. According to the summary
      and title, please classify the summary based on the
      definitions and keywords of the following occupational
      groups. Notice: focus on the end user rather than
      developers or roles of LLM itself, and do not make
      assumptions. A summary can include various types of end
      users. Some have specific occupations, while others are
      less defined. If a data point indicates a specific
      occupation, provide that occupation instead of labeling
      it as "Unknown." Labeling it as "Unknown" implies that
      the other labels are invalid.
Definitions for each occupation group are as follows:

1. Management:
This group includes occupations that involve planning,
      directing, and coordinating business activities at
      different levels of management.
This group include:
        1.      Top Executives
        2.      Advertising, Marketing, Promotions, Public
      Relations, and Sales Managers
        3.      Operations Specialties Managers
        4.      Other Management Occupations
Keywords: Leadership, Strategic planning, Resource allocation,
       Policy development

2. Business and Financial Operations:
This group includes occupations that focus on business
      operations and financial management. People in these
      roles are involved in analyzing business strategies,
      managing financial risks, ensuring compliance, and
      improving organizational efficiency.
This group include:
        1.      Business Operations Specialists
        2.      Financial Specialists
Keywords: Risk management, Compliance, Financial analysis,
      Business strategy

3. Computer and Mathematical:
This group includes occupations related to computer science,
      IT, and mathematics. People in these fields are engaged
      in software development, network management, data
      analysis, cybersecurity, and mathematical research.
This group include:
        1.      Computer Occupations
        2.      Mathematical Science Occupations
Keywords: Software development, Data analysis, Cybersecurity,
      Mathematical modeling

4. Architecture and Engineering:
This group includes occupations that focus on the design,
      construction, and maintenance of buildings,
      infrastructure, and machinery. Notice: this does not
      include building ai models, engineers here are experts
      in real estate.
This group include:
        1.      Architects, Surveyors, and Cartographers
        2.      Engineers
        3.      Drafters, Engineering Technicians, and Mapping
       Technicians
Keywords: Building design, Infrastructure, Technical plans,
      Safety standards

5. Life, Physical, and Social Science
This group includes all scientific researches such as biology,
       physics, chemistry, and social sciences. Professionals
      in this category are researchers who contribute to
      diverse areas, including social science analysis,
      environmental conservation, behavioral analysis, and
      public policy development.
This group includes:
        1.      Life Scientists
        2.      Physical Scientists
        3.      Social Scientists and Related Workers
        4.      Life, Physical, and Social Science Technicians
Keywords: Science, Scientific research, Environmental studies,
       Behavioral science, Policy development, Experimentation

6. Community and Social Service:
This group includes occupations focused on providing social
      services, counseling, and community support. If cases
      are just indicating a community-oriented perspective,
      then it should not fall in this category.
This group include:
        1.      Counselors, Social Workers, and Other
      Community and Social Service Specialists
        2.      Religious Workers
Keywords: Social work, Counseling, Community support, Social
      change

7. Legal:
This group includes occupations related to the legal system,
      where professionals provide legal representation,
      support judicial processes, and interpret laws. These
      roles require a strong understanding of legal principles,
       analytical thinking, and excellent communication skills.

This group include:
        1.      Lawyers, Judges, and Related Workers
    2.  Legal Support Workers
Keywords: Legal representation, Judicial process, Law
      interpretation, Advocacy

8. Education, Training, and Library:
This group includes occupations that focus on education,
      training, and providing access to information. These
      people are responsible for teaching, training, and
      assisting others in gaining knowledge and skills.
      Students are also included in this category.
This group include:
        1.      Postsecondary Teachers
        2.      Preschool, Primary, Secondary, and Special
      Education School Teachers
        3.      Other Teachers and Instructors
        4.      Librarians, Curators, and Archivists
    5.  Other Educational Instruction and Library Occupations
Keywords: Teaching, Curriculum development, Educational
      resources, Information access

9. Arts, Design, Entertainment, Sports, and Media:
This group includes content creators, writers and editors, and
       also occupations in creative fields such as design,
      performing arts, social media, and entertainment.
This group include:
        1.      Art and Design Workers
        2.      Entertainers and Performers, Sports and
      Related Workers
        3.      Media and Communication Workers
        4.      Media and Communication Equipment Workers
Keywords: Social media, Artistic design, writers, writers,
      editors, editors

10. Healthcare Practitioners and Support
This group includes occupations involved in diagnosing,
      treating, and supporting patient care.
This group includes:
        1.      Health Diagnosing and Treating Practitioners
        2.      Health Technologists and Technicians
        3.      Nursing, Psychiatric, and Home Health Aides
        4.      Occupational Therapy and Physical Therapist
      Assistants and Aides
        5.      Other Healthcare Practitioners and Support
      Occupations
Keywords: Diagnosis, Treatment, Patient care, Therapy,
      Caregiving, Medical expertise, Rehabilitation

11. Sales and Related:
This group includes occupations involved in selling goods and
      services, maintaining customer relationships, and
      managing sales transactions.
This group include:
        1.      Supervisors of Sales Workers
        2.      Retail Sales Workers
        3.      Sales Representatives, Services
        4.      Sales Representatives, Wholesale and
```

```
      Manufacturing
        5.      Other Sales and Related Workers
Keywords: Sales, Advertising, Customer relations, Product
      promotion, Negotiation

12. Office and Administrative Support:
This group includes occupations that provide clerical and
      administrative support to businesses and organizations.
      People in these roles handle tasks such as scheduling,
      and financial record-keeping.
This group include:
        1.      Supervisors of Office and Administrative
      Support Workers
        2.      Communications Equipment Operators
        3.      Financial Clerks
        4.      Information and Record Clerks
        5.      Material Recording, Scheduling, Dispatching,
      and Distributing Workers
        6.      Secretaries and Administrative Assistants
        7.      Other Office and Administrative Support
      Workers
Keywords: Clerical work, Customer service, Administrative
      tasks, Record keeping, Secretarial services, Dispatching,
       Financial clerks

13. Production:
This group includes occupations related to the manufacturing
      and production of goods. The roles include operate
      machinery, manage production processes, and ensure
      quality control in various industries.
This group include:
        1.      Supervisors of Production Workers
        2.      Assemblers and Fabricators
        3.      Food Processing Workers
        4.      Metal Workers and Plastic Workers
        5.      Printing Workers
        6.      Textile, Apparel, and Furnishings Workers
        7.      Woodworkers
        8.      Plant and System Operators
        9.      Other Production Occupations
Keywords: Manufacturing, Quality control, Machinery operation,
       Production processes

14. Others:
If categorized in this class, the occupation of the end users
      is clearly defined and should be one of these categories:
       Protective Service; Food Preparation and Serving
      Related; Building and Grounds Cleaning and Maintenance;
      Personal Care and Service; Farming, Fishing, and
      Forestry; Construction and Extraction; Installation,
      Maintenance, and Repair; Transportation and Material
      Moving; Military Specific.
If it doesn't fit into above occupational categories, you
      should categorize it into 15. Unknown, not 14. Others

15. Unknown:
This category is used for occupations that cannot be clearly
      identified within previous groups. It should only be
      used if there is insufficient information in the article
       to determine any specific occupation for the end user,
      or if the LLM's usage is so broad that it does not
      pertain to a particular occupational context.

Here are two examples you can refer to:
Example 1: xxx; Reason: xxx
Example 2: xxx; Reason: xxx

USER
Subreddit: {subreddit}
Article summary: {summary}
The output format should be as follows, do not generate other
      extra things:

category: list the categories fit the datapoint, separate them
       with "/", such as : Management/Business and Financial
      Operations.
explanation: explain the reason for your classification.
confidence: Set the confidence to 1 if you are confident in
      your judgment. If the information is insufficient or you
       cannot make a reliable judgment, set the confidence to
      0.
```

Listing 3: Prompts for Single-Stage Classification Module (Occupation)