The AI Self-Service Shift: Mapping Invisible Consumer Labor in ChatGPT

Gabriele Agosta
Politecnico di Torino
Torino, Italy
gabriele.agosta@studenti.polito.it

Shayan Bagheri
Politecnico di Torino
Torino, Italy
shayan.bagheri@studenti.polito.it

Shahrzad Shivaei
Politecnico di Torino
Torino, Italy
shahrzad.shivaei@studenti.polito.it

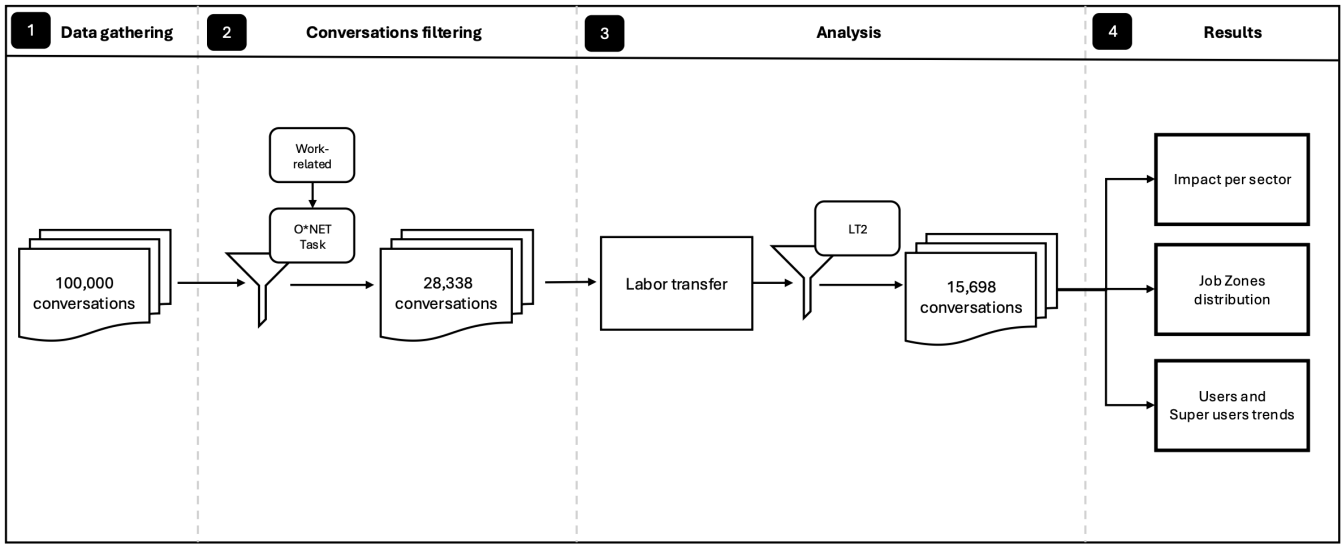


Figure 1: The proposed automated pipeline processes human-AI interaction logs to quantify the shift of professional tasks toward consumers. The workflow begins with data gathering (Step 1) to extract a sample of conversations, which is then passed through a dedicated filter (Step 2) to isolate exclusively work-related interactions. These conversations are subsequently mapped to standard O*NET professional occupations and specific occupational tasks (Step 3) based on their textual context. Finally, the pipeline evaluates the occurrence and nature of labor transfer (Step 4), enabling a comprehensive structural and longitudinal analysis of how AI-assisted self-service impacts different economic sectors.

Abstract

The widespread adoption of Large Language Models like ChatGPT is driving an unprecedented transformation of the labor market, making the quantification of these shifts essential to understand the changing dynamics of the modern workforce. While current research focuses heavily on workplace automation and job replacement, an unexplored economic shift is the transfer of professional work to unpaid consumers through AI-assisted self-service. In this work, we propose an automated pipeline to formalize, map, and measure this labor transfer. Processing a sample of ~100,000 real-world conversations, our pipeline screens for work-related usage, maps interactions to the O*NET database across ~19,000 occupational tasks, and categorizes labor transfer into three tiers. We further conduct a longitudinal analysis spanning two years of user interactions. Results show that occupations requiring higher education account for a large share of AI-assisted consumer labor. We also identify the occupations and tasks most affected by this shift, finding that nearly 70% of the critical tasks in Computer and Mathematical occupations are now being performed by consumers with AI assistance.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Large Language Models, Human-AI Interaction

1 Introduction

The diffusion of Large Language Models (LLMs) like ChatGPT has challenged traditional economic assumptions regarding labor and automation. While early research focused on how AI impacts workers by accelerating or substituting specific occupational tasks within traditional job boundaries, a critical dimension remains unobserved: the migration of professional tasks entirely into the hands of unpaid consumers. Enabled by conversational AI, everyday individuals are

Authors' Contact Information: Gabriele Agosta, Politecnico di Torino, Torino, Italy, gabriele.agosta@studenti.polito.it; Shayan Bagheri, Politecnico di Torino, Torino, Italy, shayan.bagheri@studenti.polito.it; Shahrzad Shivaei, Politecnico di Torino, Torino, Italy, shahrzad.shivaei@studenti.polito.it.

executing specialized duties themselves, creating a massive, unrecorded ecosystem of invisible consumer self-service labor.

Historically, technology-driven self-service was strictly bounded by a lack of specialized training, confining consumer labor to low-complexity administrative chores; generative AI breaks this pattern by acting as an on-demand expert, allowing laypersons to independently navigate high-stakes professional domains.

To formalize, map, and measure this transition, we introduce a classification pipeline designed to map natural conversation logs directly to formal occupational taxonomies by processing a comprehensive set of real-world interactions spanning a two-year horizon. Through this approach, our work yields three primary contributions:

- We present a pipeline that reliably anchors unstructured human-AI interaction logs to standard work taxonomies.
- We identify the preparation-level required by the tasks performed by self-service labor.
- We examine labor transfer across different sectors, identifying the most affected ones and the criticality of the tasks being transferred.

2 Related Work

Prior economic assessments of LLMs heavily concentrate on measuring workplace exposure within formal employment boundaries, often isolating non-professional usage into a separate domain.

[Eloundou et al. 2023] analyzed ChatGPT usage and found that roughly 80% of the U.S. workforce could see at least 10% of their tasks affected. In particular, they distinguished between raw LLM capabilities (affecting 15% of tasks) and the amplified impact of LLM-powered software (affecting 47%-56% of tasks). Personal usage was also measured, with activities such as creative writing and information retrieval identified as the most common consumer applications. However, the study did not explore how these personal interactions might intersect with professional tasks.

[Handa et al. 2025] leveraged large-scale conversation logs from Claude.ai to investigate the economic tasks executed by end-users. Their findings revealed that LLM deployment is heavily concentrated in the *Computer and Mathematics* occupational family. Furthermore, by partitioning task-level impacts into automation versus augmentation, they observed that the vast majority of workplace interactions served to augment, rather than entirely automate, formal employment duties.

In [Frey 2026], the author argues that while historical technological waves primarily automated manual chores, generative AI is also expanding into highly specialized domains that historically required years of professional training. We expand upon this insight by introducing a way of identifying and quantifying this unobserved economic shift.

3 Proposed Solution

We introduce an automated classification pipeline structured into four sequential stages designed to process, filter, and map unstructured interaction data into labor metrics (Figure 1).

The classification tasks within this pipeline use OpenAI's `gpt-5-mini` [Singh et al. 2026]. To ensure reproducibility, the model is queried with a temperature parameter set to 0.

3.1 Conversation Filters

We begin with WildChat-4.8M [Zhao et al. 2024], restrict the corpus to English conversations, and draw a 7% random sample, yielding approximately 100,000 conversations. We then apply a work-related filter using a prompt (Appendix B) that asks whether a conversation involves an occupational task.

3.2 Task Mapping

Conversations successfully identified as work-related are subsequently mapped to professional occupations and specific occupational tasks using the O*NET database [National Center for O*NET Development 2026]. This classification operates as a two-step hierarchical process.

In the first step, the pipeline extracts potential professions from the conversation context. The LLM evaluates the user prompt against a formatted list of O*NET occupation titles, selecting the top N most relevant professions ($N = 5$). In the second step, the pipeline narrows its scope exclusively to the specific tasks associated with those five identified occupations, from which it extracts the top N most plausible tasks ($N = 5$) based on the conversation's content.

To filter out noisy or weak mappings, we implement a consensus filter structured as follows:

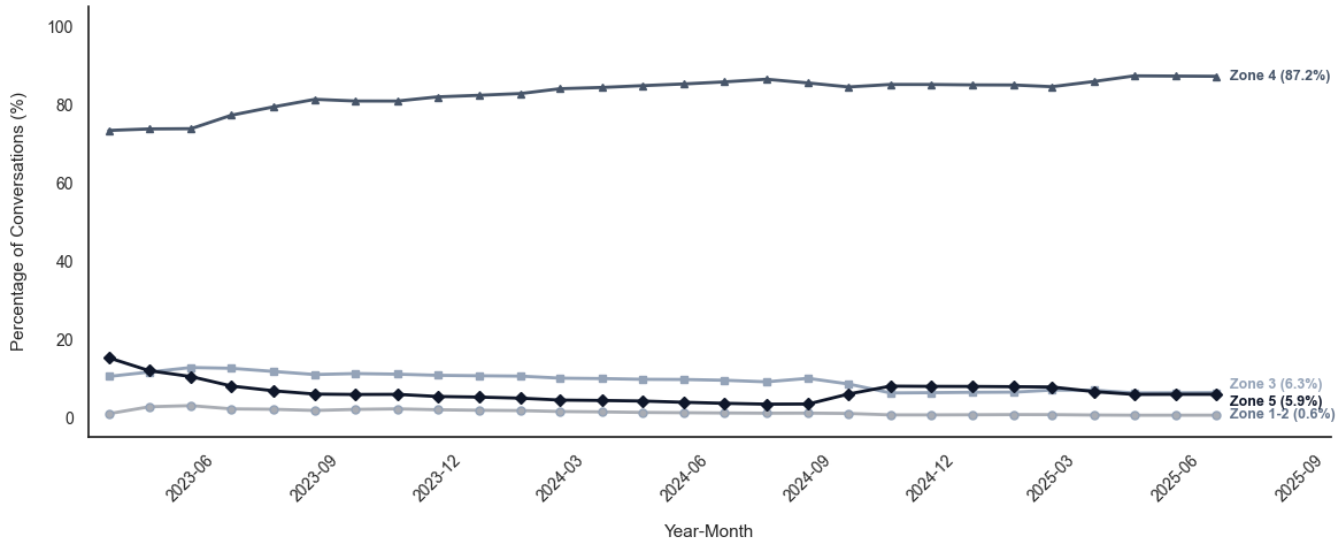
(1) *Aggregation*: Among the top N selected tasks, the pipeline aggregates and counts the frequency of each underlying parent profession assigned to a given conversation.
(2) *Majority Threshold*: A conversation is retained in the final dataset if and only if an absolute majority ($\geq 50\%$) of these tasks belong to the same primary O*NET profession.
(3) *Task Assignment*: If the majority threshold is met, the conversation is definitively mapped to the highest-ranked task belonging to that majority profession.

This hierarchical strategy effectively overcomes the limitations of injecting the entire O*NET taxonomy into a single prompt, which would otherwise require an impractical context window. Furthermore, by evaluating a restricted subset of candidates at each stage, we avoid spreading the model's attention over more than 19,000 distinct occupational tasks.

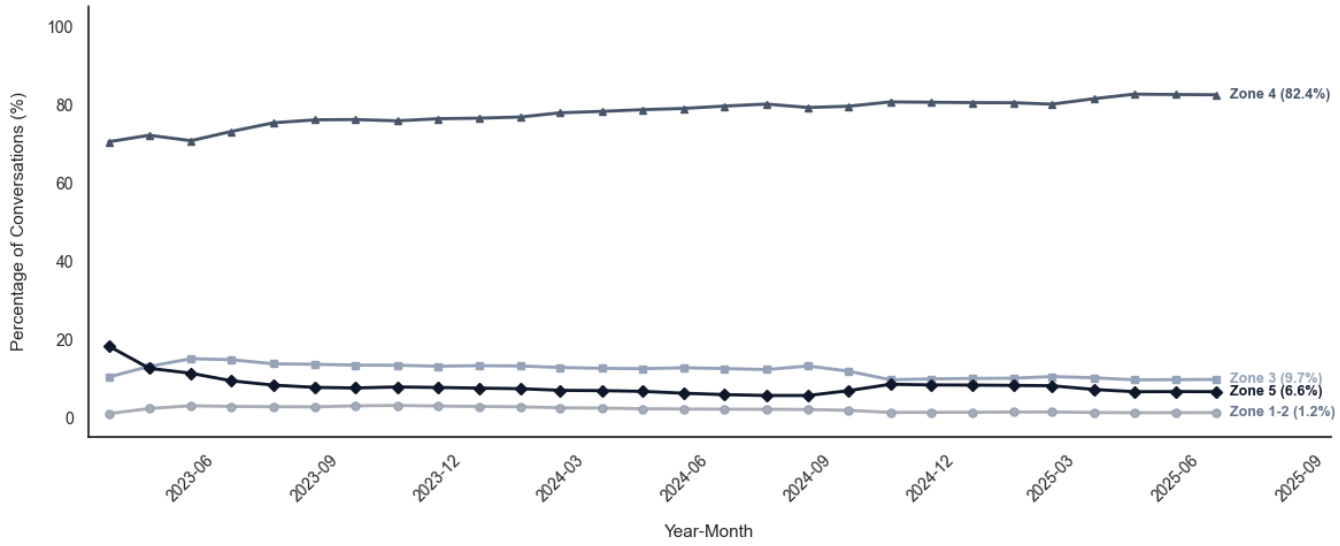
3.3 Labor Transfer

The final and core stage of our pipeline evaluates the occurrence of labor transfer. To formalize and quantify this shift, the pipeline analyzes each successfully mapped O*NET conversation, determining the user's implicit role and categorizing the interaction into one of three tiers:

- **LT2**: Assigned when an everyday consumer utilizes the model to execute or substantially prepare a concrete task that is traditionally outsourced to a paid professional (e.g., auditing a billing error).
- **LT1**: Assigned when the user undertakes a fragment of professional-like labor, but the economic stakes or circumstantial evidence remain low or ambiguous (e.g., basic translation or minor proofreading).
- **LT0**: Applied when the interaction is strictly confined to general education, casual entertainment, or open-ended



(a) Job Zones trends of the regular users



(b) Job Zones trends of the super-users

Figure 2: Longitudinal evolution of full labor transfer across O*NET Job Zones. Figure 2a shows the percentage distribution of the Job Zones of the tasks assigned to the conversations of regular users. Figure 2b focuses only on the super-users, defined as the 1% of users with the highest number of conversations. In both cases, the percentage of Job Zone 4 dominates, followed by Job Zone 3. However, among super-users there is a slightly higher percentage of Job Zone 5, mostly at the beginning of the time period. This means that despite the difference in the number of conversations, super-users and regular users have a similar distribution of Job Zones.

brainstorming. This label also captures scenarios where the user is an active professional leveraging the model for domain-specific job augmentation.

4 Evaluation

The goal of our pipeline is to determine how ChatGPT affects workers and their tasks. To ascertain whether it meets that goal, we ask:

(1) How accurately does our pipeline identify work-related conversations?
(2) Are the task-conversation mappings correct?
(3) Are the labor transfer labels accurate?

4.1 Metrics

To evaluate the pipeline's robustness, we measured its performance across our three core components. For the work-related filter (1), we measured *accuracy*, *True Positive Rate* (TPR), and *False Positive Rate* (FPR) against a manually annotated subset of 100 random conversations. For the occupation-task mappings (2), we assessed accuracy using the agreement rate between the human annotators and the LLM. Finally, for the labor transfer labeling (3), we measured classification alignment using *Cohen's kappa* (κ).

4.2 Setup

Datasets. To evaluate our method on realistic user-chatbot interactions, we leveraged the WildChat dataset [Zhao et al. 2024]. Specifically, we utilized the `WildChat-4.8M` variant, a large-scale corpus of real-world conversations between human users and OpenAI models (e.g., GPT-3.5-Turbo, GPT-4o, and o1) between 2023-04 and 2025-08. These interactions were collected via the platform *Hugging Face Spaces*, offering free ChatGPT access to users who voluntarily opted into anonymous data collection. This specific version contains exclusively non-toxic conversations. From this corpus, we sampled a subset of $\sim 100,000$ English-language conversations.

To create a link between ChatGPT conversations and real-world professional activities, we map user interactions to occupational tasks using the O*NET database [National Center for O*NET Development 2026]. Developed by the U.S. Department of Labor, O*NET provides a framework that covers the entire spectrum of the U.S. economy, tracking over 900 distinct occupations that are further broken down into more than 19,000 unique, textually defined work tasks. For our analysis, we extract four key dimensions from this database:

(1) *Occupations*, identified through the Standard Occupational Classification (SOC), a federal coding system used to classify job titles.
(2) *Tasks*, which provide the granular textual descriptions of job duties.
(3) *Task Importance*, which defines the relative significance of each task within a given job title.
(4) *Job Zones*, which categorize occupations into five levels based on the required education, experience, and on-the-job training.

Baselines. For the work-related conversation filter (1), we adopt a majority-class classifier that always predicts *not work-related*, reflecting the dominant label in our annotated subset. For the occupation-task mapping (2), we compare against the methodology of [Handa et al. 2025], who apply a comparable task-mapping approach to Claude.ai conversations. Finally, for the labor transfer labeling (3), we adopt a random classifier as baseline and interpret the resulting kappa score against the reference ranges of [McHugh 2012], where $\kappa > 0.80$ is considered to indicate strong agreement.

4.3 Execution

Validation. To evaluate accuracy, TPR, and FPR for (1), we drew a random sample of 100 conversations from WildChat and manually annotated each as work-related or not. We then ran the pipeline on the same 100 conversations and computed the three metrics against our annotations; the majority-class baseline accuracy was derived from the same subset. For (2), the agreement rate was computed between the annotators' assignments and the pipeline's output. For (3), κ was computed between the annotators' labels and the pipeline's output; a random classifier was also evaluated on the same annotated subset.

Analysis. Following the validation of the individual pipeline components, we processed the 28, 338 conversations that passed both filtering stages. For each, the pipeline generated an occupation-task-conversation mapping and assigned a labor transfer label. We subsequently enriched these results with *ONET Job Zones* and *local timestamps* to perform a temporal decomposition of labor transfer across professional levels. Furthermore, we utilized *O*NET Task Importance scores* and *O*NET-SOC Codes* to analyze the distribution of labor transfer and the criticality of the tasks involved, defining a "critical" task as one falling within the 75th percentile of importance for its respective profession.

Furthermore, to investigate the behavior of the most active participants, we identified "super-users", defined as the 1% of users with the highest number of conversations. While the majority of users engage in only a single conversation (Appendix A, Figure 4), super-users interact with the model with significantly higher frequency. We extracted these interactions to compare their usage patterns and Job Zone distribution against the broader user base, providing a longitudinal view of how high-intensity users evolve their utilization of ChatGPT over time.

4.4 Results

Validation. For the work-related conversation filter (1), the model achieved an accuracy of 0.77, a True Positive Rate (TPR) of 0.91, and a False Positive Rate (FPR) of 0.03, substantially outperforming the majority-class baseline accuracy of 0.50. The high TPR in particular confirms that the filter reliably captures professional workflows while generating very few false positives.

For the hierarchical occupation-task mappings (2), the human evaluators' agreement with the model's final task assignment reached 92.31%, surpassing the 86% agreement reported by [Handa et al. 2025].

Finally, for the three-tier labor transfer labeling (3), the pipeline achieved a Cohen's kappa of 0.82, substantially above the random baseline ($\kappa = 0.06$), and also exceeding the threshold for strong agreement as defined by [McHugh 2012].

Job Zones. As displayed in Figure 2, the displacement of professional labor is not uniformly distributed across education levels; rather, it is heavily concentrated at the upper end of the skill spectrum. The most prominent feature of the timeline is the absolute dominance of Job Zone 4. Conversely, lower-preparation tiers (Job Zones 1 and 2) are practically absent from the self-service ecosystem. This indicates that consumers are not turning to generative AI

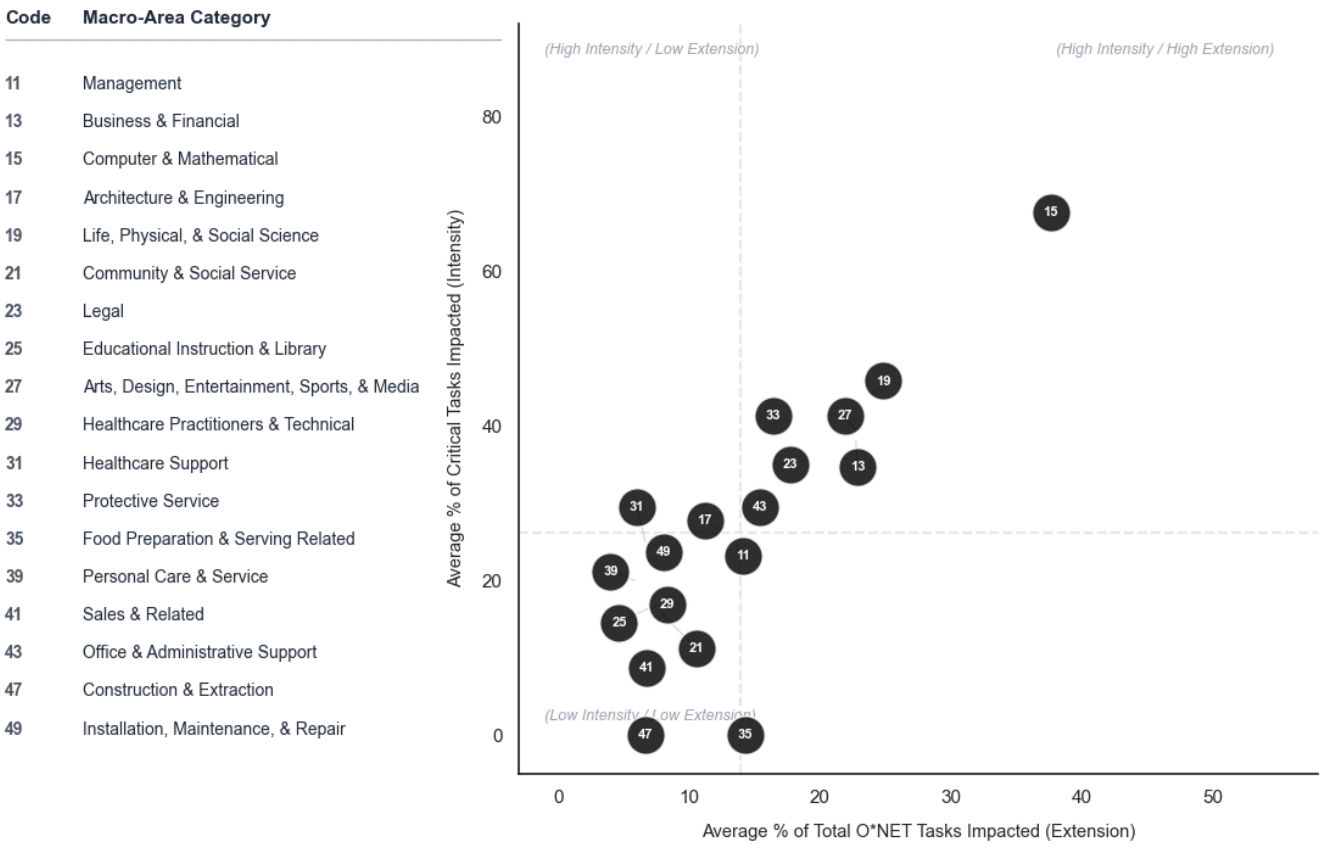


Figure 3: Labor Transfer over O*NET-SOC Macro-Categories. Each coordinate represents an O*NET-SOC Macro-Category; the x-axis measures Extension - the average percentage of total occupational tasks within that category executed by consumers via ChatGPT - while the y-axis measures Intensity - the percentage of those consumer-executed tasks that are categorized as "critical", falling within the 75th percentile of task importance. White collar professions under categories such as "Computer & Mathematical" and "Life, Physical, & Social Science" are the most affected. Conversely, manual domains like "Construction & Extraction" are the least impacted.

to bypass basic, low-complexity labor; instead, they are stepping directly into the roles of highly trained knowledge workers. Notably, early in the platform's lifecycle, approximately 20% of tasks fell within Job Zone 5, but this percentage steadily declined over time, stabilizing at around 6% in the last six months. This stabilization implies that as the technology matured, user expectations normalized around a consistent, highly predictable set of core professional activities. Furthermore, a granular temporal decomposition reveals that this pattern stands not only across months but also throughout the day. As shown in Figure 6 (Appendix A), the distribution of Job Zones across different hours of the day remains consistent, mirroring the overall macro-trends.

*O*NET-SOC Codes.* By analyzing the intersection of task extension and intensity within the *LT*2 dataset, we can identify exactly which professional markets are experiencing active consumer hollowing. As shown in Figure 3, the "Computer & Mathematical" category exhibits both the highest extension and intensity, with a vast portion of its critical duties being directly executed by consumers. This stems from the nature of the output: because software code is entirely digital and self-contained, a consumer can fully realize a programming task without needing a human intermediary. Further down, the "Legal", "Business & Financial Operations", and "Life, Physical, & Social Sciences" display elevated task intensity and moderate overall extension. This pattern means that when consumers look to bypass professionals in these domains, they target highly critical tasks, such as drafting custom contracts or analyzing tax liabilities. Meanwhile, manual trades in areas such as "Construction & Extraction" position themselves at the bottom-left of the plot, showing that physical dependencies create a hard floor for labor transfer.

5 Limitations

First, the WildChat [Zhao et al. 2024] dataset consists of users who opted into free ChatGPT access via Hugging Face. This population may skew toward highly tech-literate individuals, potentially overrepresenting tech-heavy domains like Computer and Mathematical

occupations. Second, our three-tier labor transfer taxonomy (LT0-LT2) relies entirely on visible conversational context. Because we do not observe real-world outcomes, we cannot definitively verify if a user successfully executed a task or if they completely bypassed a paid professional in practice.

6 Conclusion

In this work, we formalized and measured a largely unrecorded economic shift: the transfer of specialized professional tasks to unpaid consumers through AI-assisted self-service. By developing a pipeline that anchors unstructured conversation logs to the O*NET taxonomy, we analyzed the dynamics of this invisible labor ecosystem across two years. Our findings show that generative AI allows consumers to independently step into highly-skilled roles within Job Zone 4. This is most intense in Computer and Mathematical occupations, where 70% of the tasks can be executed directly by consumers.

References

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. arXiv:2303.10130 [econ.GN] https://arxiv.org/abs/2303.10130

Carl Benedikt Frey. 2026. A.I. Claims to Make Our Lives Easier. Does It? *The New York Times* (11 May 2026). https://www.nytimes.com/2026/05/11/opinion/ai-jobs-chores-work.html Opinion | Guest Essay.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. 2025. Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. arXiv:2503.04761 [cs.CY] https://arxiv.org/abs/2503.04761

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

National Center for O*NET Development. 2026. *O*NET OnLine*. https://www.onetonline.org/

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Y. Guan, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomek Korbak, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2026. OpenAI GPT-5 System Card. arXiv:2601.03267 [cs.CL] https://arxiv.org/abs/2601.03267

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Bl8u7ZRlbM

A Plots

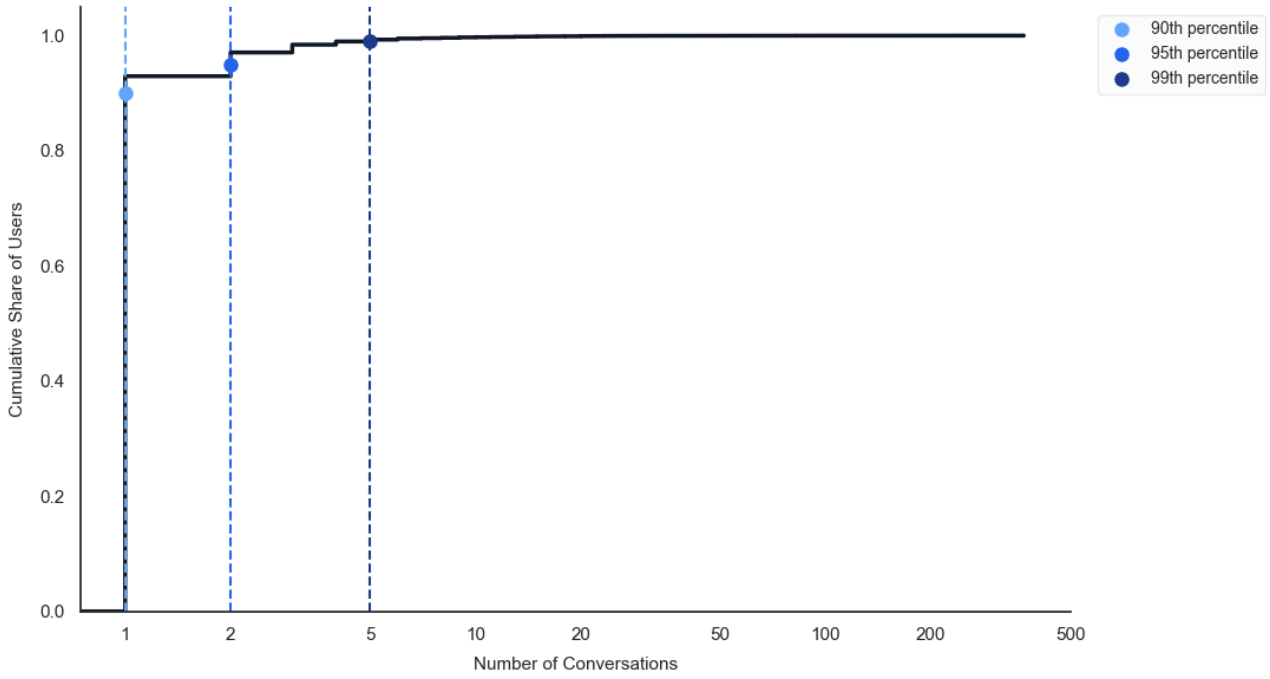


Figure 4: Conversations per location. The plot shows the cumulative percentage distribution of conversations across different locations. In this case, location is used to refer to an hashed IP address. The x-axis represents the number of conversations, while the y-axis represents the percentage of users. The plot indicates that the majority of users have only one conversation, while the top 1% of users, or super-users, have more than 5 conversations each, with some users having up to ~ 400 conversations.

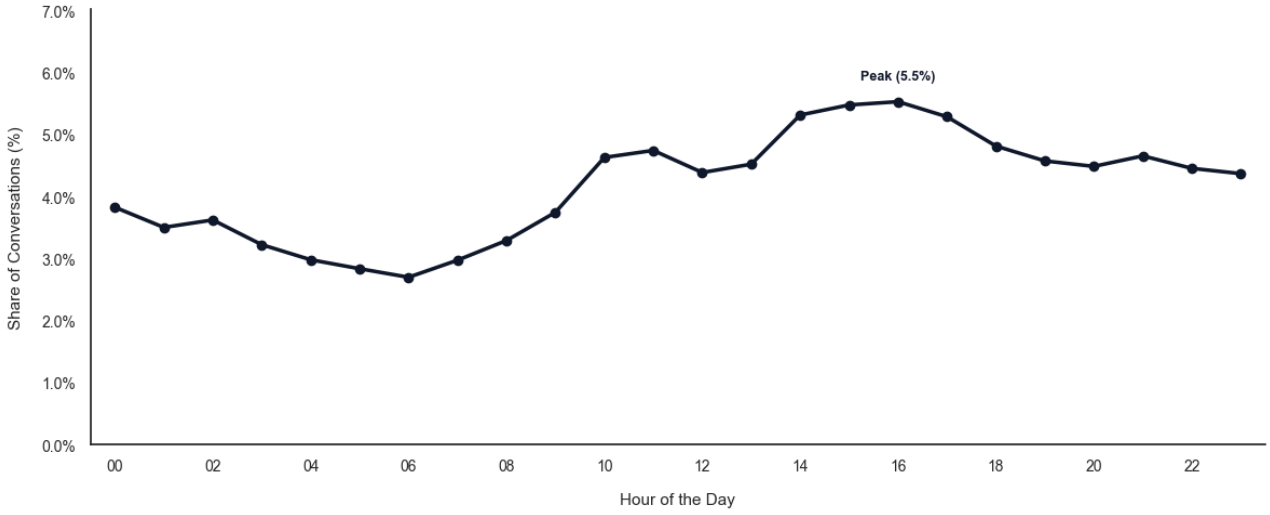


Figure 5: Share of conversations by hour of the day. The plot shows the distribution of conversations across different hours of the day. The x-axis represents the hour of the day, while the y-axis represents the share of conversations. The plot indicates that during night hours there are less work-related conversations, while during the day the share of conversations is higher, with a peak around 16 pm, and decreases during lunch hours and evenings. This pattern suggests that users are more likely to engage in work-related conversations during typical working hours.

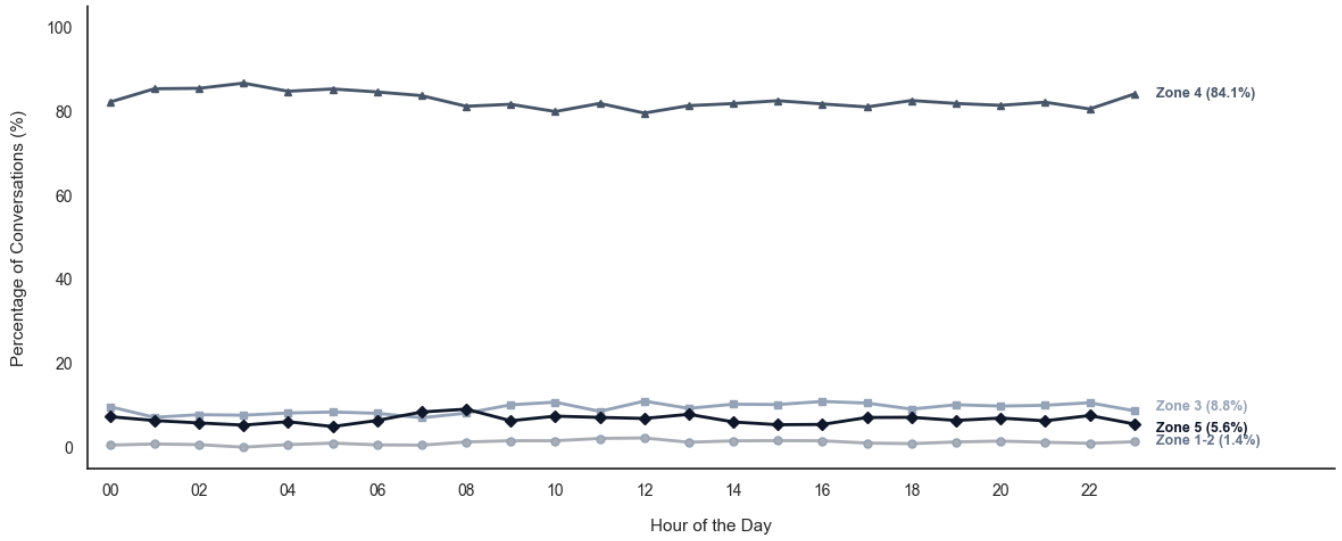


Figure 6: Job Zones per hour of the day. The plot shows the distribution of job zones across different hours of the day. The x-axis represents the hour of the day, while the y-axis represents the share of conversations. The plot indicates that the distribution of job zones is consistent throughout the day.

B Prompts

Prompt for Work-related conversation filtering

```
You are a classification agent determining whether a ChatGPT conversation  involves a task that could plausibly occur in
a professional or work context.

## Core Principles
Does this specific conversation - given its content, framing, and the nature of what is being produced - appear to be
     part of someone's actual job, business activity, organizational responsibility, or client-facing work?
You are not asking whether the task type *could theoretically* be someone's job. You are asking whether *this instance*
looks like work, based on what is visible in the conversation.

## Decision Procedure
1. Identify the core task: what is the user actually trying to accomplish?
2. Ask: is this the kind of task that appears on a job description, in a professional workflow, or as part of an
     occupation?
3. Ask: is there any signal - in the task, the vocabulary, the framing, or the output requested - that points toward a
     professional context?
4. Assign a label.

## Routing Rules
- (2) Clearly Work-Related:
    - The conversation contains direct evidence that the user is carrying out a task for an employer, client, business,
    organization, profession, or occupational responsibility.
    - The user is producing, modifying, or analyzing a real work deliverable.
    - The conversation references workplace stakeholders, customers, projects, contracts, operations, reporting,
    compliance, business processes, or professional responsibilities.

- (1) Ambiguous:
    - The task could plausibly occur in either a professional or personal/educational context and the conversation
```

```
    provides no signal to distinguish between them.

- (0) Clearly Not Work-Related:
    - The task is personal, recreational, or social with no plausible professional equivalent: casual conversation,
    personal advice, creative fiction for entertainment, hobbies, or daily life tasks.
    - Generic explanations or recommendations, even on technical topics: "how does X work" and similar requests are
    informational queries regardless of the domain.

Important:
- The following signals do not constitute evidence of a professional context on their own:
  - The task type exists as a profession (e.g. translation) but the conversation shows no applied or deliverable framing
  - The output is technically polished or detailed
  - The user asks for a technical explanation
- Label 2 only if a reasonable reader would infer that the user is currently performing work for an employer, client,
business, organization, or professional practice. If that inference is not clearly supported by the conversation, do not
use 2.
- Do not use (1) merely because the task type has a professional equivalent somewhere. Almost any task could
     theoretically be someone's job. Use (1) only when there is a concrete signal in the conversation that points toward
     a professional context - but not enough to be certain. If the conversation contains no such signal at all, use (0).

## Output Format
Return ONLY a valid JSON object:
{{
    "reasoning": "One or two sentences identifying the core task and explaining why it does or does not fit a
    professional workflow.",
    "answer": 2 | 1 | 0
}}
```

Prompt for Occupation Mapping

```
You are an expert classification assistant.

You will be provided with a conversation between an AI assistant and a user, along with a list of candidate job titles.
Your task is to identify which job titles could perform the specific task executed by the assistant in that conversation.

Consider the following list of classification options:
<options>{options_str}</options>

Your job is to identify which job title could perform the task being executed by the assistant in the
previous human-AI assistant conversation. What is the answer? You MUST provide exactly {n_options}, all written as
above.  If more then {n_options} options apply, choose the most {n_options} pertinent ones. First, start off by
considering various aspects of the conversation in the scratchpad field in at most four sentences, and then provide the
final answer in the answer field. The final output must be a JSON object that uses the following format:

{{
    "scratchpad": "<your reasoning process in at most four sentences>",
    "answer": [
        "<first option>",
        "<second option>",
        "..."
        "<your nth option>"
    ]
}}
```

```
Rules:
- The answer field is a list of exactly {n_options} from the options provided above, and they must be exactly as written
above. Do not modify the option text in any way.
- The scratchpad should contain your reasoning process, and it should be at most four sentences long.
```

Prompt for Occupational Task Filtering

```
You are an expert classification assistant.

You will be provided with a conversation between an AI assistant and a user, along with a list of candidate tasks
associated with their corresponding job titles. Your task is to identify which task is being executed by the assistant
in that conversation.

Consider the following list of classification options:
<options>{options_str}</options>

Your job is to identify which task is performed by the assistant in the previous human-AI assistant conversation.
What is the answer? You MUST provide exactly {n_options} as written above. If multiple options apply, choose
the {n_options} most pertinent ones. First, start off by considering various aspects of the conversation in the
scratchpad  field in at most four sentences, and then provide the final answer in the answer field. The final output
must be a JSON  object that uses the following format:

{{
    "scratchpad": "<your reasoning process in at most four sentences>",
    "answer": [
        "<first option>",
        "<second option>",
        "..."
        "<your nth option>"
    ]
}}

Rules:
- The answer field is a list of exactly {n_options} from the options provided above, and they must be exactly as written
above. Do not modify the option text in any way.
- Each entry in the answer field must contain both the profession and the task value from the options provided,
formatted as "profession: task".
- The options in the answer field should be ordered from the most to the least pertinent.
- The scratchpad should contain your reasoning process, and it should be at most four sentences long.
```

Prompt for Labor Transfer

```
You are coding whether a ChatGPT conversation represents labor transfer: a shift of work from a paid worker,
professional, organization, or service provider to a user, who performs or prepares the task themselves with ChatGPT's
assistance.

You will be given:
1. A WildChat conversation between a user and ChatGPT.
2. A matched O*NET task statement and its occupation.

## Core principles
```

```
- Code only what is visible in the conversation. Do not speculate about future devices, integrations, or technologies.
- Do not label labor transfer merely because the matched O*NET task belongs to a paid occupation. Ask what the user is doing in this specific conversation.
- The relevant question is whether the user is using ChatGPT to perform, substantially prepare, or substitute for a task that could plausibly otherwise have been delegated to, or performed by, paid labor.
- Labor transfer does not require proof that this specific user would definitely have paid someone. It is enough that the user is taking on a concrete task associated with paid professional, organizational, or service work.
- Learning, entertainment, brainstorming, general explanation, and casual advice are not labor transfer by themselves.
- When uncertain between two labels, choose the lower labor-transfer label.

## Decision procedure
1. Identify what the user is trying to accomplish in the conversation.
2. Determine the user's role relative to the matched task. Are they a layperson/consumer handling something in their own life, a professional doing their own job in this domain, a student, or ambiguous?
3. Evaluate task match. Does the matched O*NET task actually fit what the user is doing?
4. If the user is a layperson/consumer and the task match is good or partial, ask whether the user is using ChatGPT to do or substantially prepare the task themselves.
5. Assign a labor-transfer label.

## Routing rules
Apply these before assigning LT1 or LT2.
- If the user appears to be a professional doing their own job in the matched O*NET domain, set `interaction_type` = "augmentation" and `label` = "LT0".
Example signals: domain-specific jargon used naturally, references to clients/cases/patients, workplace context, manipulation of highly specialized artifacts, or explicit professional role.
- If the user appears to be a student or learner doing coursework, exam preparation, or general skill development without a concrete real-world task, set `interaction_type` = "education" and `label` = "LT0".
- If the matched O*NET task clearly does not fit what the user is doing, set `interaction_type` = "bad_match" and `label` = "LT0". Do not stretch the conversation to fit the task.
- Otherwise, set `interaction_type` = "consumer" and proceed to LT0/LT1/LT2.

Important: A user may be a professional in one domain but a consumer in another. For example, a software developer asking for tax help should be treated as a consumer relative to tax-preparation work.

## Labels
Apply these when `interaction_type` = "consumer".
LT2 = Clear labor transfer.
The user is using ChatGPT to perform or substantially prepare a concrete task that could plausibly otherwise be done by paid labor. The conversation involves a specific situation, artifact, decision, document, problem, or action grounded in the user's own life or non-professional responsibilities.
Examples:
- Drafting a legal, financial, insurance, medical, administrative, or employment-related document.
- Reviewing a real bill, contract, form, diagnosis, policy, claim, or technical problem.
- Preparing a concrete repair, tax filing, dispute, application, complaint, plan, or negotiation.
- Troubleshooting a real household, technical, financial, or bureaucratic problem.

LT1 = Partial or weak labor transfer.
The user is taking on part of a professional-like or service-like task, but the evidence is limited. Use LT1 when one of the following applies and record it in `lt1_reason`:
- `partial`: only a small fragment of the task is being performed.
- `low_stakes`: the task resembles paid labor but is minor enough that delegation was unlikely.
- `ambiguous_situation`: it is unclear whether the user has a concrete real-world situation or is exploring hypothetically.
- `unclear_user_role`: it is unclear whether the user is a layperson/consumer or a professional in the matched domain.
```

```
LT0 = No labor transfer.
The user is not performing or substantially preparing a concrete professional-like or service-like task in a way that
substitutes for paid labor. Use LT0 when the conversation is mainly learning, entertainment, casual discussion, generic
explanation, open-ended brainstorming, or when the O*NET task does not fit.

## Task match
Set ‘task_match‘ as:
- ‘good‘: the O*NET task closely matches the user’s concrete activity.
- ‘partial‘: the O*NET task is related but broader, narrower, or only partly represented.
- ‘bad‘: the O*NET task does not fit the conversation.
If ‘task_match‘ = "bad", label must be LT0.

## Source of transferred labor
If ‘label‘ is LT1 or LT2, name the role or service the labor would plausibly have been transferred from.
Use one of: teacher, tutor, lawyer, paralegal, accountant, tax_preparer, doctor, nurse, therapist, repair_technician,
travel_agent, customer_support, editor, translator, administrative_assistant, financial_advisor, software_developer,
designer, recruiter, insurance_advocate, real_estate_agent, other.

If ‘label‘ is LT0, use "none".

If using "other", briefly specify the role in ‘transferred_from_other‘.

## Examples:
{examples}

## Output
Return only valid JSON:
{{
    "interaction_type": "consumer" | "augmentation" | "education" | "bad_match",
    "task_match": "good" | "partial" | "bad",
    "label": "LT0" | "LT1" | "LT2",
    "lt1_reason": "partial" | "low_stakes" | "ambiguous_situation" | "unclear_user_role" | null,
    "transferred_from": "teacher" | "tutor" | "lawyer" | "paralegal" | "accountant" | "tax_preparer" | "doctor" |
    "nurse" | "therapist" | "repair_technician" | "travel_agent" | "customer_support" | "editor" | "translator" |
    "administrative_assistant" | "financial_advisor" | "software_developer" | "designer" | "recruiter" |
    "insurance_advocate" | "real_estate_agent" | "other" | "none",
    "transferred_from_other": "specific role if transferred_from is other, otherwise null",
    "rationale": "1-2 sentences mentioning specific evidence from the conversation and the matched O*NET task.",
    "confidence": "high" | "medium" | "low"
}}
```