



Management Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

AI-Driven Human Task Substitution: Evidence from Google Translate and ChatGPT

Erdem Dogukan Yilmaz, Ivana Naumovska, Vikas A. Aggarwal

To cite this article:

Erdem Dogukan Yilmaz, Ivana Naumovska, Vikas A. Aggarwal (2026) AI-Driven Human Task Substitution: Evidence from Google Translate and ChatGPT. Management Science

Published online in Articles in Advance 14 Oct 2026

. <https://doi.org/10.1287/mnsc.2024.06740>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. You are free to download this work and share with others for any purpose, except commercially, and you must attribute this work as “Management Science. Copyright © 2026 The Author(s). <https://doi.org/10.1287/mnsc.2024.06740>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc/4.0/>.”

Copyright © 2026 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes. For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

AI-Driven Human Task Substitution: Evidence from Google Translate and ChatGPT

Erdem Dogukan Yilmaz,^{a,*} Ivana Naumovska,^a Vikas A. Aggarwal^a

^aINSEAD, 77305 Fontainebleau, France

*Corresponding author

✉ erdemdogukan.yilmaz@insead.edu (E. D. Yilmaz), ivana.naumovska@insead.edu (I. Naumovska), vikas.aggarwal@insead.edu (V. A. Aggarwal)

🌐 <https://orcid.org/0000-0002-2727-3557> (E. D. Yilmaz), <https://orcid.org/0000-0001-8565-3649> (I. Naumovska), <https://orcid.org/0000-0002-2040-6767> (V. A. Aggarwal)

Received: July 18, 2024

Revised: August 27, 2025; March 1, 2026

Accepted: April 20, 2026

Published Online in Articles in Advance: October 14, 2026

<https://doi.org/10.1287/mnsc.2024.06740>

Copyright: © 2026 The Author(s)

Open Access Statement: This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. You are free to download this work and share with others for any purpose, except commercially, and you must attribute this work as "Management Science. Copyright © 2026 The Author(s). <https://doi.org/10.1287/mnsc.2024.06740>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc/4.0/>."

Abstract. We examine how different generations of artificial intelligence (AI) substitute for human knowledge worker tasks, leveraging two sequential natural experiments in the translation industry. We study the introduction of Google Neural Machine Translation (GNMT), a predictive, task-specific AI, as well as ChatGPT, a broad, general-purpose generative AI, in a major online labor market. Distinguishing between analytical tasks (regular translations) and intuitive tasks (transcreations, which require creative and cultural adaptation), we estimate difference-in-differences models on transaction-level data from 2016 to 2017 and 2022 to 2023. We find that GNMT reduced the demand for analytical tasks by economically meaningful magnitudes, while having no detectable effect on intuitive tasks. By contrast, ChatGPT produced substantially larger and more uniform reductions for both analytical and intuitive tasks, indicating a marked expansion in the scope and magnitude of AI substitution. By comparing these two technological shocks within the same labor market, we show that conclusions about which types of human work are resilient to automation do not straightforwardly generalize from predictive to generative AI.

History: Accepted by Joshua Gans, business strategy.

Supplemental Material: The online appendix and data files are available at <https://doi.org/10.1287/mnsc.2024.06740>.

Keywords: artificial intelligence • labor • organizational studies • strategy

1. Introduction

Artificial intelligence (AI) is increasingly capable of performing nonroutine cognitive tasks traditionally carried out by skilled knowledge workers, raising the prospect of substantial disruption across a wide range of industries (e.g., Agrawal et al. 2018, 2023; Felten et al. 2021, 2023; Fedyk et al. 2022; Brynjolfsson et al. 2025). Following the release of ChatGPT, public and academic discourse around the substitution potential of AI has intensified. Recent studies suggest that AI could plausibly substitute for a large share of knowledge work tasks, with estimates ranging from 35% to 80%, though these projections vary widely in scope and methodology (Fedyk et al. 2022, Agrawal et al. 2023, Brynjolfsson et al. 2025). Despite this growing body of work, there is limited retrospective empirical evidence on how AI's impact varies across different types of knowledge worker tasks. While a small number of recent studies have begun to explore heterogeneity at the level of job clusters (e.g., Hui et al. 2024, Demirci et al. 2025), task-level evidence on differential substitution

effects remains scarce (Fedyk et al. 2022, Agrawal et al. 2023, Brynjolfsson et al. 2025, Doshi and Hauser 2024, Jia et al. 2024).

Earlier waves of automation primarily substituted for human labor in occupations dominated by routine tasks, such as accounting, data entry, and assembly line work (Autor et al. 2003, Acemoglu and Restrepo 2020). By contrast, the nonroutine cognitive tasks that characterize much of contemporary knowledge work are more complex, often requiring judgment, interpretation, and contextual understanding (Brynjolfsson et al. 2018, Frank et al. 2019, Tong et al. 2021, Agrawal et al. 2023). Several factors help explain why retrospective evidence on AI's substitution effects in these domains remains limited. First, fine-grained data on the adoption and use of AI technologies is typically unavailable (Raj and Seamans 2018, Frank et al. 2019, Acemoglu et al. 2022). Second, within-firm task structures often require complementary investments in technology and human capital, complicating efforts to isolate the effects of AI itself (Choudhury et al. 2020,

Brynjolfsson et al. 2021). Third, the gradual and uneven diffusion of AI across firms and industries makes it difficult to cleanly identify causal impacts on workers and organizations (Brynjolfsson et al. 2021).

In this study, we address these limitations by responding to calls for more fine-grained empirical investigations into the effects of AI on human tasks (Frank et al. 2019) and for clearer distinctions between AI technologies with different capabilities (Acemoglu et al. 2022). We provide empirical evidence on how two sequential forms of AI—a predictive, task-specific system (PAI) and a more recent, broad, general-purpose generative system (GAI)—affected the demand for human knowledge work. Our analysis focuses on whether and how these effects differ between analytical tasks and tasks requiring intuitive and creative judgment.

Our study complements recent work by Frey and Llanos-Paredes (2025), who document the long-run effects of Google Translate on translator employment and the demand for foreign language skills at the aggregate labor market level. Although their findings establish that predictive AI reduced overall demand for translation-related labor, they leave open the question of which specific tasks drove this substitution, and whether conclusions drawn from the predictive AI era generalize to newer forms of AI. We address these gaps in three ways. First, by comparing predictive and generative AI shocks within the same empirical context, we identify how the magnitude and boundaries of AI substitution have shifted over time, rather than simply documenting differences in performance across tools. Second, we disaggregate effects at the task level, distinguishing between analytical (regular translation) and intuitive (transcreation) tasks, revealing heterogeneity that occupation-level analyses necessarily obscure. Third, by using transaction-level data from an online labor market (OLM), we provide granular evidence on the immediate substitution effects of successive AI technologies.

The empirical evidence we offer is anchored in two sequential natural experiments in the context of language translation, a domain populated by skilled knowledge workers whose tasks involve the generation, processing, and transformation of information. The first natural experiment is the deployment of Google Neural Machine Translation (GNMT) as a backend upgrade to the existing Google Translate product for specific language pairs beginning in November 2016. Although Google Translate had been publicly available and widely used for years, GNMT replaced its earlier phrase-based engine with a neural architecture, constituting a discrete and substantial improvement in translation quality. GNMT is a sophisticated neural machine translation system and exemplifies a predictive AI designed for a narrowly defined, sequence-to-sequence task. Its staggered rollout constituted an

exogenous shock that suddenly and unexpectedly made high-quality machine translation widely accessible at zero marginal cost.

The second natural experiment is the release of ChatGPT by OpenAI in November 2022. ChatGPT represents a qualitatively different form of AI: a broad, general-purpose generative system capable of performing translation alongside a wide range of other cognitive tasks, such as ideation, synthesis, and analysis (Korinek 2023). Taken together, these two sequential shocks allow us to examine not only whether AI substitutes for human translation work, but also how the scope of substitution changes as AI architectures evolve. This design provides a retrospective perspective that complements forward-looking assessments of AI's labor-market implications (e.g., Raj and Seamans 2018, Frank et al. 2019) and enables a task-level analysis of substitution that is difficult to obtain in studies focused on a single technology or period (e.g., Brynjolfsson et al. 2025).

With the release of ChatGPT, questions about whether generative AI systems differ meaningfully from earlier predictive AI systems in their labor market effects became more salient in public discourse. In particular, observers began to ask whether advances in AI architecture expand the set of tasks for which human labor is substitutable, especially among nonroutine cognitive tasks that require interpretation, judgment, or creative adaptation. While both predictive and generative AI systems can perform analytical tasks, generative AI's training and architecture may allow it to perform tasks that rely more heavily on contextual understanding and meaning. Despite widespread discussion of these possibilities, retrospective empirical evidence comparing how different generations of AI affect distinct task types remains limited.

To structure our investigation, we ask three related questions. First, how do AI technologies affect overall labor demand for knowledge work? Second, does the magnitude of this effect vary across task types, specifically between analytical tasks and tasks requiring intuitive or creative judgment? Third, do different AI architectures (predictive versus generative) produce distinct substitution patterns across these tasks? Together, these questions allow us to assess not only whether AI substitutes for human labor but whether conclusions drawn from earlier AI waves generalize to newer forms of AI. A central contribution of our analysis is not only documenting whether substitution occurs but quantifying how substitution magnitudes differ across task types and AI generations.

To answer these questions, we examine the effects of GNMT and ChatGPT on the demand for human language translation in a major OLM, which provides a standardized and complete archive of transactions (Agrawal et al. 2016, Horton and Tambe 2025). We distinguish between two types of translation tasks:

(i) regular translation, which involves converting text from one language to another, and (ii) transcreation, which requires creative and cultural adaptation to preserve meaning, tone, and style for a new audience. Transcreation is common in settings such as marketing, poetry, and literary translation and entails stylistic reinvention and cultural sensitivity (Holland and Quinn 1987, Glăveanu et al. 2013).

Employing a difference-in-differences design, we find that GNMT reduced demand for regular translation services by approximately 9.1% within seven months of launch, having no statistically significant effect on transcreation tasks. ChatGPT, by contrast, exhibits a markedly different pattern: demand declines substantially for both regular translations (21.1%) and transcreations (29.6%), indicating comparable substitution effects across analytical and intuitive tasks. Supplemental analyses reveal an additional distinction across AI generations. Whereas GNMT's impact does not extend to tasks requiring deep domain-specific knowledge, ChatGPT is associated with significant reductions in demand for specialized technical translations as well.

Taken together, these findings suggest that, whereas both predictive and generative AI substitute for non-routine cognitive work, they do so in qualitatively different ways. In the predictive AI era, substitution was concentrated in analytical tasks, leaving intuitive and creative tasks relatively insulated. In the generative AI era, this boundary no longer holds. As a result, conclusions about which types of knowledge work are resilient to automation (drawn from earlier AI waves) do not straightforwardly generalize to newer forms of AI.

These findings are particularly relevant given the accelerating pace of AI development and our task-level focus within an occupation. Occupations, firms, and industries can be understood as bundles of tasks: that is, distinct units of work that vary in their susceptibility to automation (Autor et al. 2003, Acemoglu and Autor 2011, Garicano and Wu 2012). AI technologies are, therefore, unlikely to substitute for entire occupations at once, instead affecting specific tasks depending on the alignment between a technology's capabilities and task demands. As a result, even within the same occupation or firm, workers may face very different substitution risks depending on the bundle of tasks they perform.

Our results show that this task-level heterogeneity is shaped not only by whether tasks are nonroutine but also by whether they are primarily analytical or intuitive, and by the generation of AI technology being deployed. In particular, task allocations that appeared relatively insulated from automation in the predictive AI era become exposed under generative AI. This task-level lens is, therefore, essential for understanding how advances in AI reshape the boundaries of substitution over time, expanding the range of tasks that can be

automated, and limiting the extent to which inferences drawn from earlier automation waves generalize to newer forms of AI.

2. Background Literature

2.1. The Substitution Effect of AI on Human Labor

Developments in AI have the potential to substantially affect firms, industries, and labor markets. Within management research, scholarly discourse has focused on how quickly and to what extent AI technologies may reshape competition, employment structures, and firms' human capital (Choudhury et al. 2020, Tschang and Almirall 2021, Krakowski et al. 2023). Much of the early work on AI's labor market implications has been forward looking, relying on theoretical models or expert assessments to forecast the potential scope of AI-driven substitution (e.g., Pajarinen and Rouvinen 2014, Brzeski and Burk 2015, Frey and Osborne 2017, Webb 2019, Felten et al. 2023, Eloundou et al. 2024).

More recent studies have begun to provide retrospective evidence by examining hiring outcomes. For example, Acemoglu et al. (2022) study firm-level online vacancy data and show that firms adopting AI reduce hiring for certain non-AI positions, altering skill requirements in remaining postings. Similarly, Fedyk et al. (2022) analyze resume data from large audit firms and find that increased AI investment is associated with a decline in accounting employment over time. Although these studies offer important insights into how AI adoption correlates with changes in labor demand, they focus primarily on occupation-level hiring outcomes and rely on observational variation that may be influenced by multiple confounding factors.

In a related study, Frey and Llanos-Paredes (2025) document a decline in aggregate demand for translators following the diffusion of Google Translate, providing compelling long-run evidence of labor displacement associated with predictive AI, estimating that Google Translate adoption reduced translator employment growth by 0.84 percentage points (instrumental variable estimate) and resulted in approximately 28,000 fewer positions over the 2010–2023 period. However, their occupation-level approach leaves open the question of which specific tasks within translation drove this substitution and whether similar patterns extend to newer generations of AI. Our approach complements this work by focusing on tasks within an occupation and by leveraging two natural experiments to isolate substitution effects more directly, thereby offering a more granular perspective on how AI technologies reshape labor demand at the task level (Frank et al. 2019, Acemoglu et al. 2022).

Distinguishing between different classes of AI systems is particularly important for understanding these

dynamics. Although both PAI and GAI fall under the broader umbrella of AI, they differ in architectural scope and the types of tasks for which they are optimized. PAI systems are typically designed for convergent, sequence-to-sequence problems. The GNMT system we study is a highly sophisticated example of such an approach, optimized to predict accurate translations within a well-defined output space. Although neural machine translation models are technically generative in that they produce new text, their architecture is geared toward convergent objectives, and this has historically limited their effectiveness in tasks requiring divergent thinking, creative judgment, or deep contextual interpretation (Argote et al. 2021, He et al. 2024).

Generative AI systems, by contrast, are built on large language models trained on diverse corpora and designed to generate a wide range of outputs across contexts. Tools such as ChatGPT or DALL-E can produce text or images that vary in tone, style, and form, and this has raised questions about whether such systems may affect a broader set of nonroutine cognitive tasks. Whereas these capabilities suggest potential differences in labor market impact relative to earlier AI systems, empirical evidence comparing how distinct AI architectures translate into differential substitution effects remains limited.

Understanding AI-driven substitution, therefore, requires moving beyond occupation-level analyses to examine the tasks that constitute human work. Occupations comprise heterogeneous bundles of tasks that vary in their susceptibility to automation, and task characteristics play a central role in shaping the substitution risk. By focusing on tasks rather than occupations, we follow established frameworks that emphasize how technological change reallocates work within jobs rather than eliminating jobs wholesale. In the next section, we build on this task-based perspective to categorize knowledge worker tasks and to examine how predictive and generative AI systems differentially affect them.

2.2. Knowledge Worker Tasks and the Implications of AI

Knowledge workers typically perform a range of tasks that differ in their susceptibility to automation and substitution. Prior research commonly distinguishes tasks along two dimensions: (i) routine versus nonroutine, capturing whether tasks involve explicitly repetitive procedures or require adaptation and judgment, and (ii) manual versus cognitive, reflecting whether tasks rely primarily on physical or intellectual work (Autor et al. 2003, Acemoglu and Autor 2011).

Routine tasks, both manual and cognitive, follow explicit rules that can be readily specified in code and have long been subject to automation through pre-AI

technologies (Autor 2014). Routine manual tasks such as product assembly, painting, and welding have been widely automated using industrial machinery (Acemoglu and Restrepo 2020, Dixon et al. 2021). Routine cognitive tasks, including billing, accounting, and data entry, have similarly been automated through software systems that execute preprogrammed instructions (Autor et al. 2003). Nonroutine tasks, by contrast, involve greater complexity and uncertainty and often require problem solving, judgment, and contextual interpretation. Examples of nonroutine manual tasks include driving in traffic or preparing a meal, whereas nonroutine cognitive tasks include medical diagnosis, legal reasoning, marketing campaign creation, scientific research, and translation. Until recently, such nonroutine tasks were widely viewed as relatively insulated from automation (Brynjolfsson et al. 2018).

Recent scholarship suggests that AI technologies may increasingly substitute for some nonroutine tasks (Brynjolfsson et al. 2018, Frank et al. 2019, Acemoglu et al. 2022, Fedyk et al. 2022), including medical diagnosis (Wang et al. 2019), wind turbine maintenance (Franko et al. 2020), and auditing (Fedyk et al. 2022). However, the substitution potential of AI is argued to be more limited when nonroutine cognitive tasks rely heavily on emotional intelligence, cultural interpretation, or other forms of context-dependent judgment (Faraj et al. 2018). This limitation reflects the fact that humans process not only codifiable inputs but also intangible and context-dependent cues such as emotions, cultural norms, and social subtleties that are difficult to formalize and encode (Marcus and Davis 2019, Wu et al. 2020). As a result, some nonroutine cognitive tasks have been viewed as more resistant to automation at least in the near term (Agrawal et al. 2018, Autor and Salomons 2018, Acemoglu et al. 2022).

Beyond the analytical–intuitive distinction, a task’s susceptibility to automation may also depend on its reliance on deep, domain-specific knowledge. Fields such as law, medicine, and engineering operate with specialized vocabularies and high-stakes accuracy requirements, which early AI systems that were often trained on broad, general-purpose corpora were thought to handle imperfectly. This has given rise to a perceived knowledge barrier to substitution. Whether the transition from narrow, task-specific AI to broad, general-purpose generative AI has altered the permeability of this barrier remains an open empirical question.

A growing body of research examines the performance of generative AI systems in tasks involving creativity, interpretation, and judgment. Laboratory and field experiments show that generative AI can enhance output quality in tasks such as business ideation and consulting (Girotra et al. 2023, Boussioux et al. 2024, Doshi and Hauser 2024, Dell’Acqua et al. 2026), whereas quasi-experimental evidence from creative

industries suggests effects on professional behavior and strategy (Peukert et al. 2024). Studies in professional services, including law, document substantial productivity gains associated with AI use (Schwarcz et al. 2026). At the same time, the extent to which these capabilities translate into substitution in real-world labor markets remains underexamined. Indeed, recent work highlights that large language models underlying generative AI systems are better at processing literal text than interpreting broader social or cultural context, and this may limit their ability to interpret meaning in culturally specific settings (NLLB Team 2024). For example, linguistic conventions such as the use of politeness to convey irony may be difficult for such systems to capture reliably.

Taken together, these perspectives suggest that different classes of AI may exhibit distinct substitution patterns across tasks with important implications for occupations, firms, and industries. Because occupations are composed of heterogeneous bundles of tasks (Autor et al. 2003, Acemoglu and Autor 2011, Garicano and Wu 2012), workers performing different task mixes may face very different automation risks even within the same job category.

In this paper, we examine how predictive and generative AI differentially affect nonroutine cognitive tasks. We disaggregate such tasks into those that are primarily analytical and those that rely more heavily on intuitive or creative judgment (Autor et al. 2003, Spitz-Oener 2006). Whereas intuitive tasks also involve substantial cognitive effort, they differ from analytical tasks in their objectives and structure. Analytical tasks are typically convergent and rule-based, requiring the application of logic and learned patterns to arrive at an optimal or probabilistically correct solution. Intuitive tasks, by contrast, are often divergent and goal-oriented, requiring the generation of novel outputs that achieve emotional, cultural, or stylistic objectives in a new context.

This distinction maps onto the architectural differences between the AI systems we study. Predictive AI systems such as GNMT are optimized for convergent, sequence-to-sequence problems and perform well within well-defined output spaces. Generative AI systems such as ChatGPT are designed to produce a broader range of outputs and may, therefore, affect a wider set of tasks. Whether and how these architectural differences translate into differential substitution effects is ultimately an empirical question, one that we examine in the context of translation, a domain involving nonroutine cognitive work and previously assessed as facing moderate automation risk (Frey and Osborne 2017, Felten et al. 2021).

2.3. Research Questions

In summary, we address three interlinked research questions designed to clarify how the scope of

AI-driven task substitution has evolved. First, what is the substitution effect of AI on nonroutine cognitive tasks? We address this by examining whether and to what extent AI technologies substitute for human labor within the domain of language translation. Second, does this substitution effect differ across generations of AI technologies? We examine this by comparing the impacts of predictive AI (GNMT) and generative AI (ChatGPT) within the same empirical setting. Third, does the effect vary systematically across task types, specifically between nonroutine cognitive tasks that are primarily analytical and those that rely more heavily on intuitive or creative judgment? We examine this by distinguishing between regular translations and transcreations.

Taken together, retrospective empirical evidence on these questions allows us to assess not only whether AI substitutes for human knowledge work but also whether conclusions drawn from earlier AI waves generalize to newer forms of AI. Addressing these questions jointly provides a task-level perspective on how advances in AI architecture reshape the boundaries of substitution for knowledge workers.

3. Empirical Contexts

3.1. Comparing GNMT and ChatGPT

We compare GNMT, as an exemplar of a highly specialized PAI system with ChatGPT, as an exemplar of a broad, general-purpose GAI system. GNMT is designed to perform a narrowly defined, convergent task: producing accurate translations by predicting the most likely output sequence for a given input. ChatGPT, by contrast, is based on a large language model trained to generate text across a wide range of contexts and prompts. As a result, it can perform translation alongside other language-related tasks, such as summarization, synthesis, and text generation.

These architectural differences prompt speculation that generative AI systems may affect a broader set of tasks than earlier predictive systems, particularly tasks involving interpretation, tone, or contextual adaptation. At the same time, such differences do not mechanically imply broader labor substitution as demonstrated performance does not necessarily translate into displacement in market settings. Whether and how these distinct architectures lead to different substitution patterns, therefore, remains an empirical question.¹

Following the release of ChatGPT, attention to the labor market implications of AI intensified relative to earlier periods dominated by predictive AI technologies (Kochhar 2023, World Economic Forum 2023). Popular and practitioner-oriented accounts often emphasize the potentially wide-ranging impact of generative AI on work (e.g., Accenture 2023). Our study complements this discourse by providing retrospective, task-level

evidence that allows us to assess how differences in AI architecture translate into observable substitution effects in a real labor market.

3.2. GNMT and ChatGPT as Shocks to Translation Tasks

Google Translate has been publicly available since 2006 and achieved widespread consumer adoption well before 2016, serving as one of the most widely used translation tools globally (Yamada 2019). However, early versions of Google Translate (as did other machine translation technologies of that era) relied on rule- and phrase-based statistical approaches that produced outputs substantially inferior to human translation (Wu et al. 2016). The emergence of neural machine translation (NMT), which incorporates broader grammatical and semantic context, led to marked improvements in translation quality and fluency (Luong and Manning 2015, Wu et al. 2016, Jia et al. 2019, Yamada 2019). Prior to Google's upgrade, however, access to high-quality NMT systems was limited. In late 2016, Google deployed GNMT as a backend architectural upgrade to its existing Google Translate product, replacing the earlier phrase-based system with a neural machine translation engine.² Our treatment, therefore, corresponds to the deployment of GNMT within Google Translate, a discrete shift from phrase-based to neural machine translation rather than to the introduction of Google Translate per se. This upgrade substantially improved translation quality and represented a sudden and unexpected increase in the availability of accurate, low-cost machine translation (Wu et al. 2016, Jia et al. 2019).³ Given Google Translate's already widespread adoption at the time of the GNMT upgrade, the sudden improvement in translation quality raised immediate concerns about the potential impact of machine translation on professional translators (Lewis-Kraus 2016, Yoon 2017, Massey and Kiraly 2019).

Although task-specific AI systems such as GNMT perform well in regular translation contexts such as news articles, essays, and technical documents, their applicability is argued to be more limited for tasks requiring adaptation to cultural, stylistic, or emotional nuance (Pedersen 2014, Vieira 2020). Linguists use the term "transcreation" to describe translation tasks that require intuitive and creative reinterpretation rather than literal conversion, such as advertising copy or poetry, for which understanding cultural and emotional context is central (Pedersen 2014).

The release of ChatGPT in late 2022 constituted a second, qualitatively different shock. ChatGPT is a general-purpose conversational AI system built on a large language model and capable of performing translation alongside a wide range of other language-based tasks. This broader applicability is characteristic of technologies that exhibit general-purpose properties

with implications for diffusion and labor market impact (Goldfarb et al. 2023). Its rapid diffusion brought generative AI capabilities to a broad user base in a short period of time. At the same time, contemporaneous assessments raised questions about its reliability, accuracy, and ability to consistently handle culturally specific or high-stakes content (Glorioso 2023, Hunter 2023), concerns that subsequent systematic evaluations have substantiated (Dahl et al. 2024, Yao et al. 2024). As with GNMT, the introduction of ChatGPT, therefore, altered the technological environment faced by translators in ways that were salient, rapid, and largely outside the control of individual workers.

4. Data and Methodology

4.1. Data Sources

To examine the impact of GNMT, we use data from one of the largest OLMs. OLMs have become an increasingly important venue for sourcing human labor in activities such as translation, proofreading, and software development (Chen and Horton 2016). Importantly for our purposes, OLMs provide a complete record of posted tasks and completed transactions, which makes them well suited for assessing how technological shocks affect labor demand (Agrawal et al. 2016, Chen and Horton 2016).

We analyze two distinct time periods corresponding to our natural experiments. To study the effects of GNMT, we focus on January 2016 through May 2017, ending the sample prior to GNMT's expansion to most language pairs offered by Google Translate in June 2017. This data set comprises 38,573 transactions across treated and control language pairs for which the language pair can be identified. To study the effects of ChatGPT, we analyze the period from January 2022 through May 2023, chosen to provide a comparable window length. This data set comprises 82,277 transactions with identifiable language pairs. Aggregating to the appropriate panel unit—language pair $\times$ month for GNMT and language pair $\times$ task type $\times$ month for ChatGPT—yields 2,482 observations for the GNMT analysis and 4,352 for the ChatGPT transaction regressions (4,301 for ChatGPT revenue regressions, for which a small number of cells drop because no price is observed).

As in most professions, translation involves heterogeneous tasks. To distinguish between different types of translation work, we classify tasks into analytical and intuitive categories corresponding to two commonly recognized forms of translation.⁴ The first category consists of regular translations, which are primarily analytical in nature and involve the literal conversion of text from one language to another. The second category consists of transcreation tasks, which require adaptive, creative, and culturally sensitive judgment, such as the translation of advertising content or poetry (Massey

and Wieder 2019). We classify transcreation as intuitive because task success is evaluated based on achieving a desired cultural, emotional, or stylistic effect rather than on literal fidelity. Each job posting in the translation sample is assigned to exactly one task category with no residual or ambiguous classification (see Online Appendix A2 for the full classification protocol).

In supplemental analyses, we further refine this categorization by distinguishing between regular translations that rely primarily on general knowledge and those that require specialized technical knowledge, such as legal, medical, or engineering content.⁵ This yields three task categories: (i) regular translation tasks involving general knowledge, (ii) regular translation tasks involving technical knowledge, and (iii) transcreation tasks. We formally test for differences in substitution effects across these task categories as reported in Section 5.3.

4.2. Identification Strategy

The introduction of GNMT occurred at different points in time across language pairs and constituted a sudden increase in the quality of machine translation.⁶ In November 2016, GNMT became available for eight language pairs, followed by the addition of 33 further language pairs between March and April 2017. Our identification strategy exploits this staggered rollout. The key identifying assumption is that the timing of GNMT's introduction for specific language pairs was not anticipated by market participants. This assumption is plausible given that neither translators nor clients could foresee the precise sequence or timing of language pair coverage. Accordingly, we treat the GNMT rollout as a quasi-experimental shock that generates plausibly exogenous variation in technology availability across language pairs rather than as a randomized experiment. Because our design is quasi-experimental rather than randomized, identification rests on parallel trends and fixed effects assumptions rather than covariate balance. We assess validity through event study analyses and placebo tests rather than balance tables (see Section 5.1 and Online Appendix A12). To isolate the effect of GNMT, we restrict the sample to January 2016 through May 2017, ending the posttreatment window before GNMT expanded to most language pairs in June 2017.

We conduct the GNMT analysis at the market level, examining how the volume of translation transactions evolves following the introduction of GNMT. Because a substantial share of observations take the value zero (47.3% in the GNMT sample), we estimate fixed effects Poisson pseudo-maximum-likelihood (PPML) models. PPML accommodates zero outcomes naturally and avoids the need for ad hoc transformations required by log-linear ordinary least squares (OLS) specifications (Chen and Roth 2024). Prior work shows that PPML yields consistent estimates under a wide range

of heteroskedasticity and data-generating processes, whereas log-linear OLS and related alternatives can be biased or inefficient when variance is misspecified or zeros are prevalent (Gourieroux et al. 1984, Santos Silva and Tenreiro 2006, Cameron and Trivedi 2013, Powell and Seabury 2018). Moreover, unlike fixed effects negative binomial models, PPML permits the inclusion of high-dimensional fixed effects without introducing incidental parameter bias (Allison and Waterman 2002, Santos Silva and Tenreiro 2011, Martin and Pham 2020).

Our baseline specification uses a language pair panel and is given by

$$Y_{it} = \alpha + \beta_1 \text{PostTreatment}_{it} + \gamma_i + \delta_t + \varepsilon_{it},$$

where Y_{it} denotes the total number of transactions for language pair i in month t . The variable $\text{PostTreatment}_{it}$ equals one for observations corresponding to language pairs after the introduction of GNMT. Language pair fixed effects (γ_i) control for time-invariant characteristics such as linguistic complexity or baseline demand, whereas time fixed effects (δ_t) capture market-wide trends common to all language pairs. Standard errors are clustered at the language pair level to account for serial correlation. Following Angrist and Pischke (2009), we do not include control variables that may themselves be affected by the treatment (e.g., prices) as doing so could bias the estimated treatment effect. Event study analyses confirm that treated and control language pairs exhibit parallel pretreatment trends with divergence emerging only after GNMT's introduction (see Section 5.1 and Online Appendix A12).⁷

The ChatGPT analysis requires a different empirical strategy because ChatGPT affected all language pairs simultaneously. We, therefore, define treatment and control groups based on task composition rather than timing. The treatment group consists of pure play translation jobs—projects focused exclusively on translation—whereas the control group comprises bundled translation tasks in which translation is only one component of a broader assignment (e.g., translation and voiceover, translation and captioning, or research and translation). This comparison allows us to examine whether tasks more directly exposed to generative AI experienced differential impacts relative to tasks for which translation is embedded within a multiskill workflow.

This control group provides a meaningful counterfactual for two reasons. First, whereas generative AI can substitute for the translation component of a task, it cannot fully perform the additional skills required in bundled projects, at least initially, reducing substitution pressure on the overall task. Second, because bundled tasks still include translation, they are subject to similar market-wide trends, allowing us to retain language pairs as the core unit of analysis.⁸ Nevertheless, we recognize that some bundled tasks (e.g., research

and translation) include components that may themselves be affected by generative AI. If so, the control group would be partially treated, biasing estimates toward zero. We, therefore, interpret our estimates as conservative.

To further address this concern, we also implement a within-unit panel difference-in-differences design. This approach compares changes in outcomes for a given language pair around the introduction of ChatGPT (e.g., 2022–2023) with changes for the same language pair during a pretreatment placebo period (e.g., 2018–2019). This self-controlled design absorbs all time-invariant heterogeneity and relies on the more plausible assumption of stable year-over-year trends in the absence of treatment.

As in the GNMT analysis, we estimate ChatGPT's effects using fixed effects PPML models. Because treatment status varies by task type as well as by language pair, the unit of observation is language pair $\times$ task type $\times$ month. We include language pair $\times$ task type fixed effects (γ_{ij}), which control for time-invariant differences in demand across language–task combinations, and month fixed effects (δ_t), which capture market-wide shocks and seasonality common to all observations.

5. Empirical Results

5.1. Baseline Results: AI Technologies' Impact on Translation Demand

We report summary statistics in the Online Appendix (Online Tables A3-1 through A3-11). Figure 1(a) plots the number of translation transactions before and after the introduction of GNMT, comparing the first cohort of treated language pairs with control language pairs. Prior to GNMT's introduction, treated and control language pairs exhibit similar demand patterns. Following GNMT's rollout in November 2016, the two groups begin to diverge: transaction volumes for control language pairs grow more rapidly than those for treated language pairs in the posttreatment period. This divergence, in which the control group grows faster than the treated group (which is growing but at a lower rate), is not a threat to identification but rather the variation that identifies the treatment effect. Because GNMT improved translation quality only for treated language pairs, it is the relative slowdown in treated pair growth, not the absolute level of control group activity, that captures the effect of the technology. During our study period, the online labor market was expanding overall, and our difference-in-differences design is specifically intended to net out this common platform-level growth by comparing relative trajectories between treated and control language pairs.⁹

Figure 1(b) presents analogous evidence for the ChatGPT analysis, comparing transaction patterns

between pure play translation tasks and bundled translation tasks. The divergence begins in December 2022, shortly after ChatGPT's release in November 2022, and becomes more pronounced by May 2023, following the release of GPT-4 in mid-March 2023.

Table 1 reports our main estimates of the effects of GNMT and ChatGPT on translation demand. Both technologies are associated with reductions in translation activity though the magnitude and precision of the estimates differ. GNMT reduces transaction volume by 8.2% ($(e^{-0.0853} - 1) \times 100$, significant at the 5% level. ChatGPT exhibits a larger effect with a 22.2% reduction in transaction volume ($\beta = -0.251$, $p < 0.001$).

For revenue, GNMT yields a negative but imprecisely estimated coefficient, corresponding to a 20.6% reduction ($\beta = -0.231$, $p = 0.120$). We interpret this estimate cautiously. Revenue in online labor markets is inherently more volatile than transaction counts as individual high-value projects can generate substantial month-to-month variation without corresponding changes in transaction volume. By contrast, ChatGPT's revenue effect is both economically and statistically significant, corresponding to a 26.1% reduction ($\beta = -0.302$, $p < 0.05$).

Figures 2 and 3 present event study estimates for GNMT and ChatGPT, respectively. For transaction volumes, pretreatment coefficients remain close to zero with no systematic trend, supporting the parallel trends assumption. The revenue specifications display greater pretreatment variability, particularly in the GNMT case, reflecting the higher inherent volatility of revenue measures relative to transaction counts. To mitigate this noise, we also implement coarsened exact matching (CEM) between treated and control observations as described below and obtain consistent estimates.

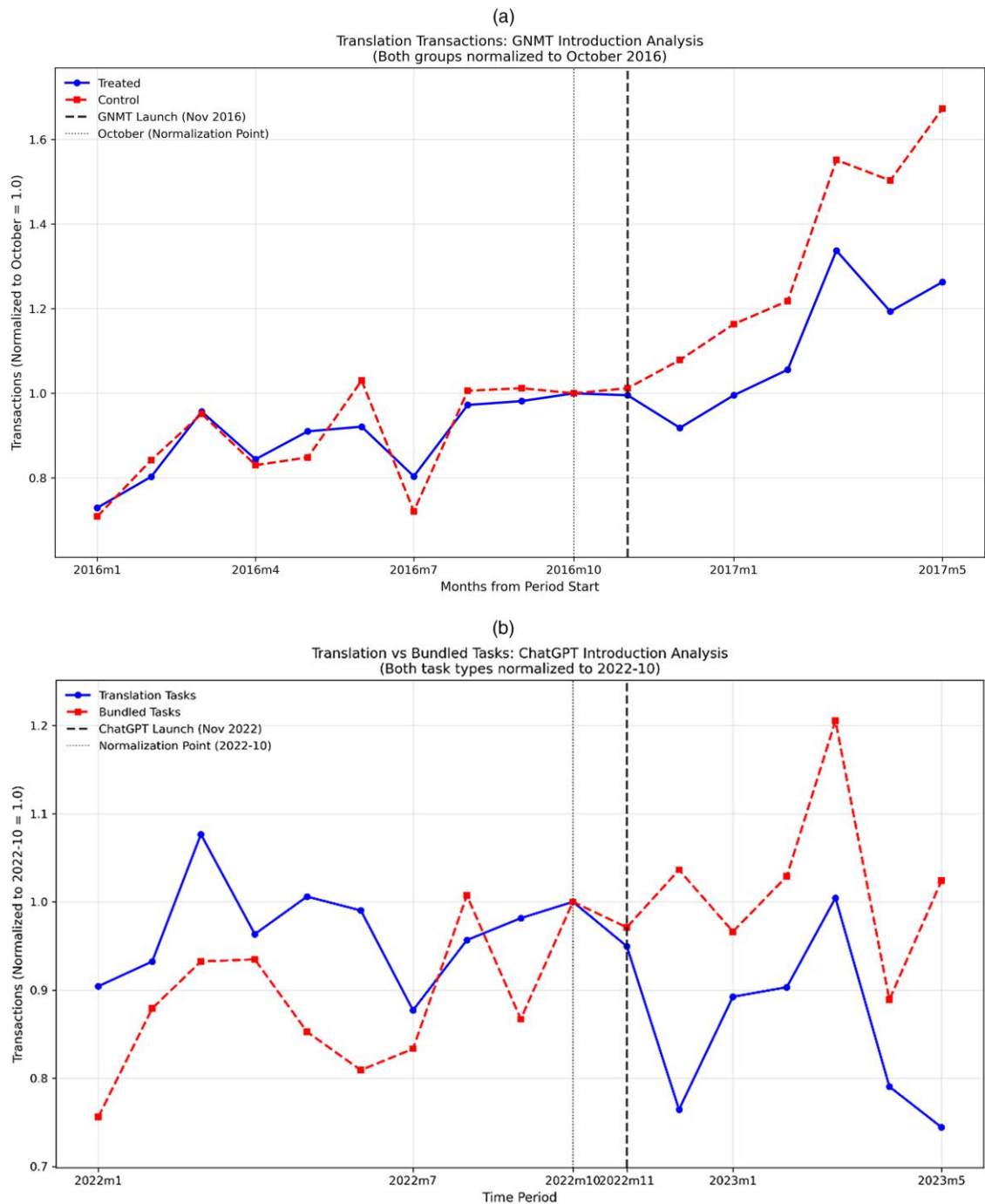
In both settings, treatment effects emerge shortly after technology introduction and persist throughout the observation window. Whereas the magnitude of the effects varies over time, the estimates generally increase toward the end of the sample period.

5.2. Heterogeneous Effects by Translation Type

Table 2 reports estimates of the heterogeneous effects of AI technologies across analytical and intuitive translation tasks. To facilitate a direct comparison between GNMT and ChatGPT, the analysis is restricted to the subset of language pairs that are active in both observation periods. We distinguish between regular translations, which involve primarily analytical language conversion, and transcreations, which require creative judgment and cultural adaptation.

For GNMT, the estimates indicate a reduction in regular translation activity with transaction volume declining by 9.1% ($\beta = -0.0958$, $p < 0.05$) and revenue declining by 25.6% ($\beta = -0.296$, $p < 0.10$). By contrast, ChatGPT is associated with larger and more precisely

Figure 1. (Color online) Number of Transactions Before and After GNMT and ChatGPT Introduction



Notes. (a) Number of transactions before and after GNMT introduction (first treatment cohort). Transactions are measured in quantity (i.e., volume) and are normalized to the level in October 2016 ($t = -1$). (b) Number of transactions before and after ChatGPT introduction. Transactions are measured in quantity (i.e., volume) and are normalized to the level in October 2022 ($t = -1$) for both treatment and control series.

estimated effects on regular translations, reducing transaction volume by 21.1% ($\beta = -0.237, p < 0.001$) and revenue by 26.2% ($\beta = -0.304, p < 0.05$).

The patterns differ more sharply for transcreation tasks. For GNMT, we find no statistically significant effect on either transcreation transactions ($\beta = 0.152, p = 0.155$) or revenue ($\beta = 0.290, p = 0.526$). Whereas the

point estimates are positive, they are not statistically distinguishable from zero. By contrast, ChatGPT is associated with a substantial reduction in transcreation transactions of 29.6% ($\beta = -0.351, p < 0.001$). The corresponding revenue effect is negative but not statistically significant at conventional levels ($\beta = -0.247, p = 0.169$). This pattern is consistent with a decline in

Table 1. Baseline Estimates: GNMT vs. ChatGPT

	(1) Total translation transactions (GNMT)	(2) Total revenue (GNMT)	(3) Total translation transactions (ChatGPT)	(4) Total revenue (ChatGPT)
<i>Posttreatment</i>	−0.0853** (0.0396)	−0.231 (0.149)	−0.251*** (0.0521)	−0.302** (0.130)
<i>Constant</i>	4.374*** (0.0131)	9.981*** (0.0453)	4.816*** (0.0184)	10.98*** (0.0416)
Language pair fixed effects	Yes	Yes	No	No
Month-year fixed effects	Yes	Yes	Yes	Yes
Language pair–task fixed effects	No	No	Yes	Yes
Pseudo R^2	0.939	0.870	0.948	0.939
Log-likelihood	−3,690.7	−2,226,165.4	−7,882.1	−3,853,629.6
Observations	2,482	2,482	4,352	4,301

Notes. Standard errors are robust to heteroskedasticity and clustered at the language pair level for GNMT and at the language pair–task level for ChatGPT. Observation counts may differ across columns because the PPML estimator drops singleton observations and observations that are separated by a fixed effect and because some language pair–month cells have no observable price for the revenue regressions. GNMT reduced total translation transactions by 8.2% ($\beta = -0.085$, $p < 0.05$) with no significant effect on revenue. ChatGPT reduced total translation transactions by 22.2% ($\beta = -0.251$, $p < 0.001$) and revenue by 26.1% ($\beta = -0.302$, $p < 0.05$).

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

the number of transcreation projects alongside greater variability in project size or scope among remaining tasks.

To assess whether the effects differ statistically across task types, we conduct coefficient comparison tests. For GNMT, the difference between the effects on regular translation and transcreation transactions is statistically significant ($\chi^2 = 4.44$, $p < 0.05$), indicating heterogeneous effects across task types. For ChatGPT, although both task categories experience negative effects, the difference between them is not statistically significant ($\chi^2 = 2.42$, $p = 0.12$), suggesting more similar impacts across analytical and intuitive tasks.

Figures 4 and 5 present event study estimates that illustrate the dynamic evolution of these effects. For GNMT, the reduction in regular translation activity emerges gradually over time, whereas transcreation exhibits no discernible posttreatment shift. For ChatGPT, the decline in regular translation activity occurs immediately following introduction. Transcreation tasks display a change from a nonnegative pretreatment trend to a predominantly negative posttreatment trajectory, albeit with greater month-to-month variability.

Taken together, these results indicate that GNMT's substitution effects are concentrated in analytical translation tasks, whereas ChatGPT is associated with reductions across both analytical and intuitive tasks with no statistically significant difference in magnitude between them.

5.3. Distinguishing Between General and Technical Translation

We further distinguish between general translations and technical translations, defined as translations of medical, legal, and engineering texts, within the regular

translation sample. Table 3 reports the corresponding estimates.

For GNMT, we find no statistically significant effects on technical translation activity. Transaction volume shows no detectable change ($\beta = -0.0418$, $p = 0.681$), and the revenue estimate is likewise small and imprecise ($\beta = -0.144$, $p = 0.689$). By contrast, ChatGPT is associated with substantial reductions in technical translation demand with transactions declining by 25.9% ($\beta = -0.299$, $p < 0.001$) and revenue declining by 30.7% ($\beta = -0.366$, $p = 0.02$).

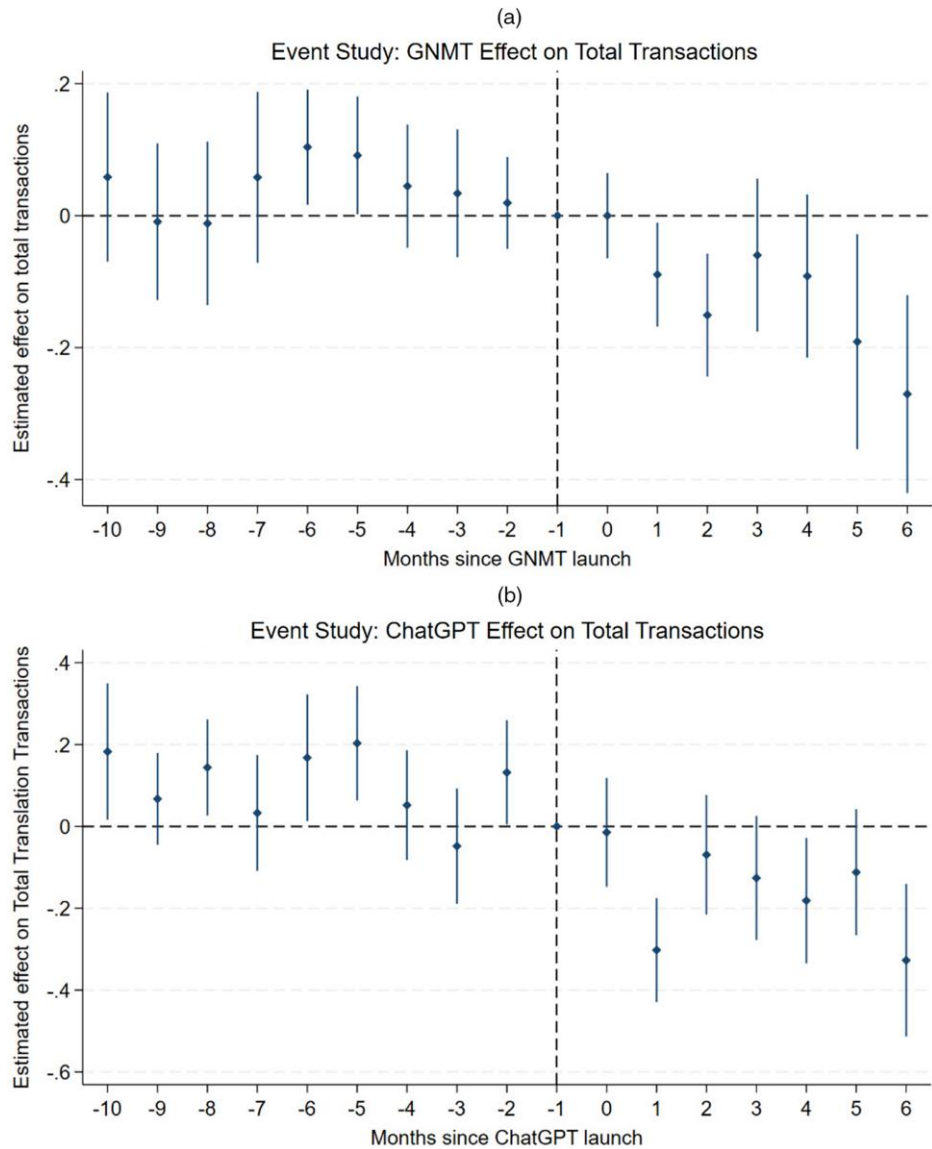
As in the previous section, we formally test whether technical and general translations respond differently to each technology. For GNMT, the difference in effects between technical and general translations is not statistically significant ($\chi^2 = 0.19$, $p = 0.662$), indicating similarly weak effects across these analytical task subcategories. For ChatGPT, both general and technical translations are significantly affected individually, but the difference between them is not statistically significant ($\chi^2 = 1.35$, $p = 0.246$). We, therefore, read the pattern as broadly similar magnitudes of substitution across the two domains rather than evidence of sharp differentiation between them.

5.4. Robustness Checks

5.4.1. CEM Analyses. Although there is no ex ante reason to expect that the timing of the introduction of GNMT or ChatGPT was anticipated by market participants, some pretreatment coefficients in the event study specifications exhibit statistical noise. To further strengthen our identification strategy and reduce concerns about residual pretreatment imbalance, we implement CEM.

CEM constructs matched samples in which treated and control observations are identical on coarsened

Figure 2. (Color online) Coefficient Estimates on the Number of Transactions Before and After GNMT and ChatGPT Introduction



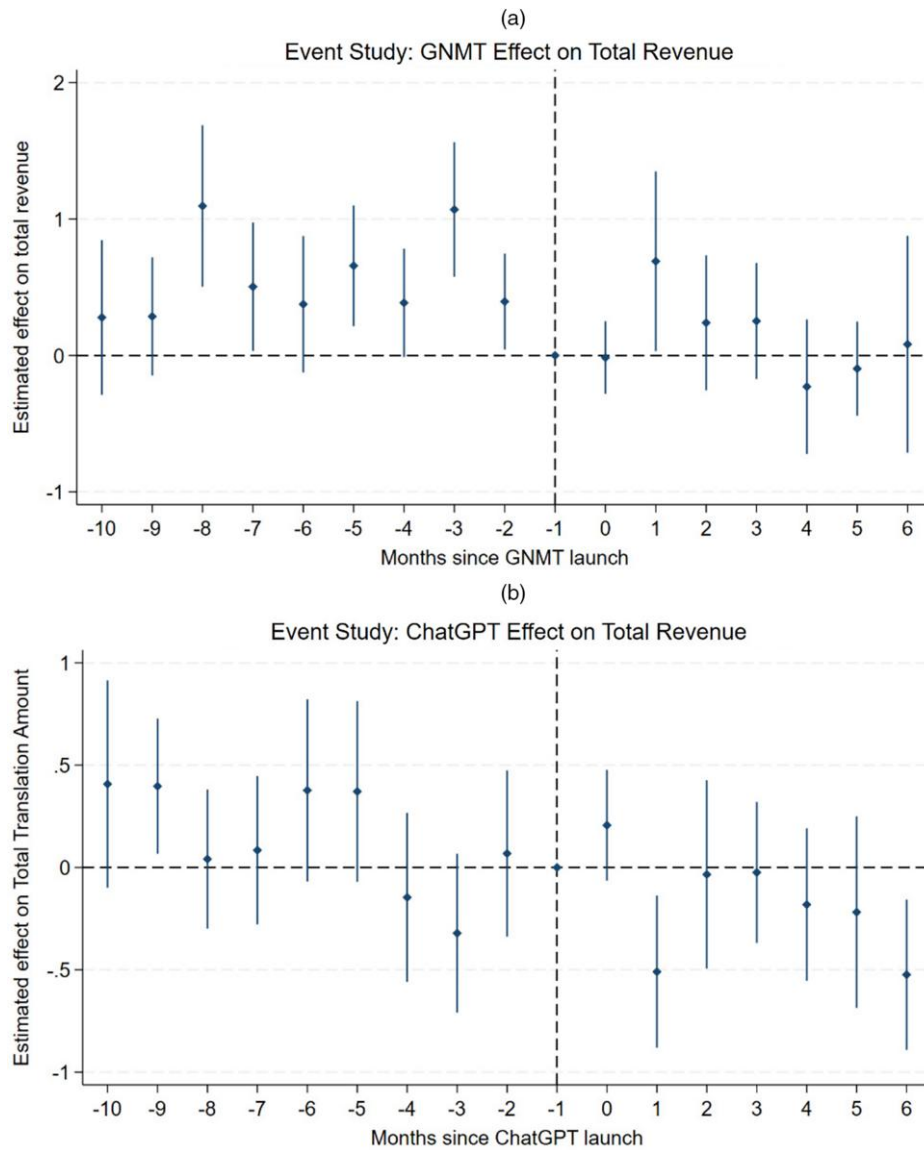
Notes. (a) Coefficient estimates on the number of transactions before and after GNMT introduction. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $Number\ of\ Transactions = \alpha + \sum z \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i is language pair fixed effects, δ_t is month-year fixed effects, and ε_{it} is the error term. The index z represents the months relative to a reference month, which marks the month when GNMT first becomes available for a given language pair. (b) Coefficient estimates on the number of transactions before and after ChatGPT introduction. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $Number\ of\ Transactions = \alpha + \sum z \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i is language pair-task fixed effects, δ_t is month-year fixed effects, and ε_{it} is the error term. The index z represents the months relative to a zero month, which marks the month when ChatGPT first becomes publicly available (November 2022). The dashed vertical line indicates the treatment month (ChatGPT launch).

versions of selected covariates, thereby improving covariate balance and reducing model dependence. For both the GNMT and ChatGPT analyses, we match on rolling averages of transaction volumes, market size (measured one, three, and six months prior to treatment), pricing patterns, and the exact month of observation.

The estimates obtained from the matched samples are consistent with our baseline results as reported

in Tables 4 and 5, reinforcing the robustness of our findings.

5.4.2. Individual-Level Analysis. We next examine changes in translation activity at the individual freelancer level following the introduction of GNMT and ChatGPT. The unit of analysis is the freelancer-month. For the GNMT analysis, we compare changes in

Figure 3. (Color online) Coefficient Estimates on the Total Revenues Before and After GNMT and ChatGPT Introduction

Notes. (a) Coefficient estimates on the total revenues before and after GNMT introduction. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $Total\ Dollar\ Amount = \alpha + \sum z \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i is language pair fixed effects, δ_t is month-year fixed effects, and ε_{it} is the error term. The index z represents the months relative to a reference month, which marks the month when GNMT first becomes available for a given language pair. (b) Coefficient estimates on the total revenues before and after ChatGPT introduction. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $Total\ Dollar\ Amount = \alpha + \sum z \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i is language pair–task fixed effects, δ_t is month-year fixed effects, and ε_{it} is the error term. The index z represents the months relative to a zero month, which marks the month when ChatGPT first becomes publicly available (November 2022). The dashed vertical line indicates the treatment month (ChatGPT launch).

transactions for freelancers active in treated language pairs prior to GNMT's introduction with a counterfactual group of freelancers who did not specialize in those language pairs. We restrict the sample to freelancers who registered on the platform at least three months before the introduction of the first GNMT cohort. This restriction yields a conservative sample as freelancers with prior transaction history are more likely to secure

work in a given period than newly registered sellers (Kokkodis and Ipeirotis 2016, Benson et al. 2020).

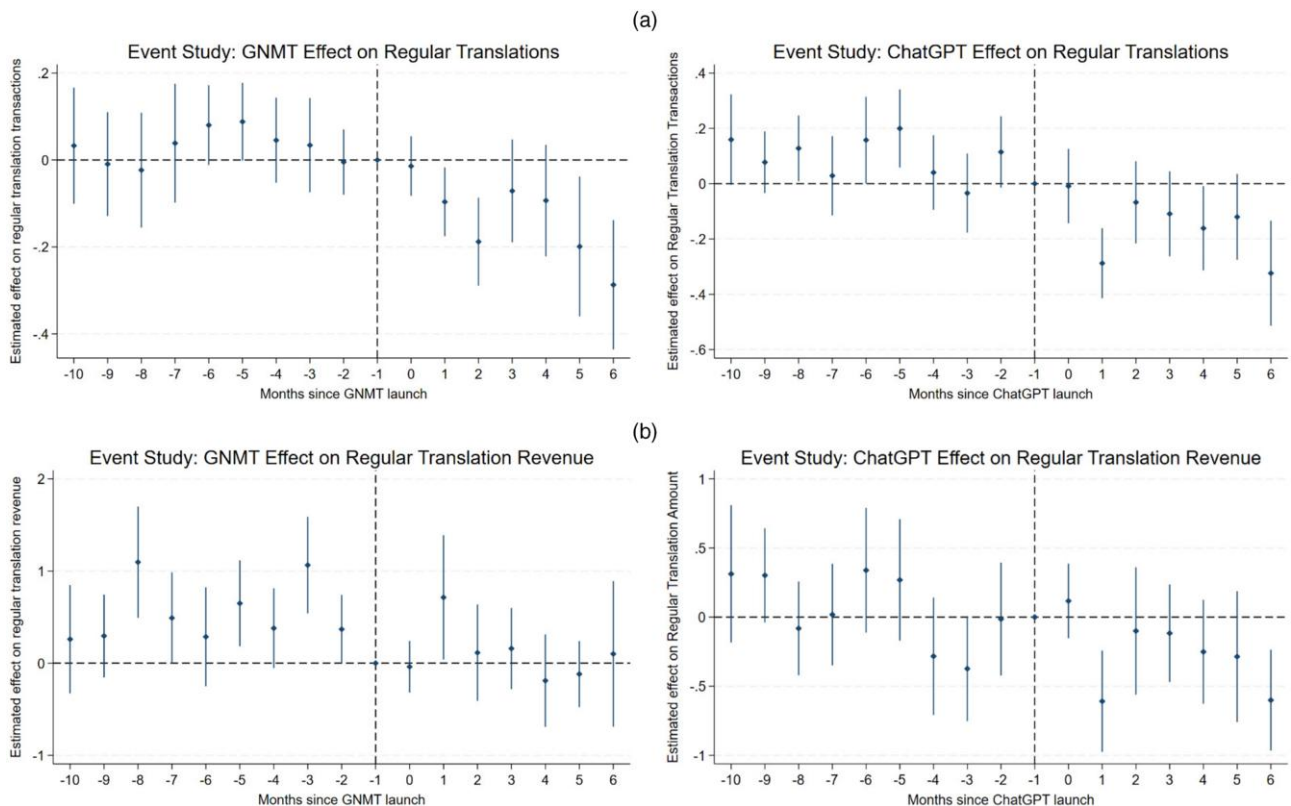
For the ChatGPT analysis, we impose the same three-month pretreatment registration requirement but adopt a balanced panel design. Specifically, we include only freelancers who completed at least one pure play translation task and at least one bundled translation task during the study period. This design

Table 2. Heterogeneous Effect of GNMT and ChatGPT on Different Translation Tasks: Regular Translations vs. Transcreations

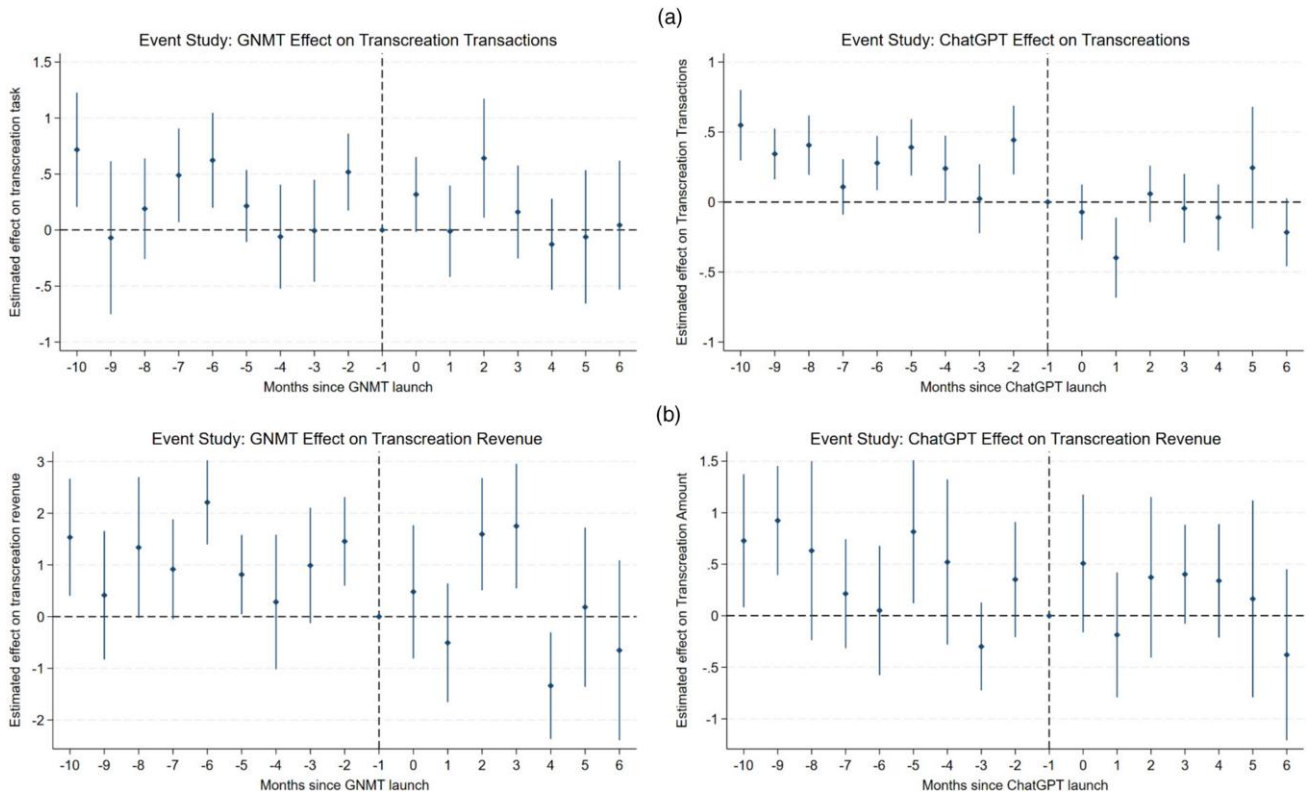
	(1) Reg Txn (GNMT)	(2) Transc Txn (GNMT)	(3) Reg Rev (GNMT)	(4) Transc Rev (GNMT)	(5) Reg Txn (ChatGPT)	(6) Transc Txn (ChatGPT)	(7) Reg Rev (ChatGPT)	(8) Transc Rev (ChatGPT)
<i>Posttreatment</i>	−0.0958** (0.0403)	0.152 (0.107)	−0.296* (0.155)	0.290 (0.458)	−0.237*** (0.0514)	−0.351*** (0.0862)	−0.304** (0.133)	−0.247 (0.180)
<i>Constant</i>	4.327*** (0.0133)	1.368*** (0.0381)	10.05*** (0.0480)	7.452*** (0.186)	4.808*** (0.0181)	2.455*** (0.0107)	10.85*** (0.0418)	9.362*** (0.0330)
Language pair fixed effects	Yes	Yes	Yes	Yes	No	No	No	No
Month-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language pair–task fixed effects	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R^2	0.937	0.624	0.820	0.679	0.943	0.740	0.926	0.798
Log-likelihood	−3,659.2	−963.4	−1,790,115.3	−313,694.4	−6,684.8	−4,101.1	−3,419,994.0	−2,497,675.0
Observations	2,482	1,054	1,054	1,037	2,890	2,890	2,890	2,890

Notes. Standard errors are robust, clustered at the language pair level for GNMT and at the language pair–treatment group level for ChatGPT. Columns (1) and (5): regular translation (general and technical) transactions; columns (2) and (6): transcreation transactions; columns (3) and (7): revenue from regular translation; columns (4) and (8): revenue from transcreation. Observation counts may differ across columns because the PPML estimator drops singleton observations and observations that are separated by a fixed effect and because some language pair–month cells have no observable price for the revenue regressions. Significant coefficients indicate that GNMT reduced regular translation transactions by 9.1% ($\beta = -0.0958$, $p < 0.05$) and revenue by 25.6% ($\beta = -0.296$, $p < 0.10$); ChatGPT reduced regular translation transactions by 21.1% ($\beta = -0.237$, $p < 0.001$), revenue by 26.2% ($\beta = -0.304$, $p < 0.05$), and transcreation transactions by 29.6% ($\beta = -0.351$, $p < 0.001$). Coefficient comparison tests: For GNMT, regular versus transcreation transaction effects differ significantly ($\chi^2 = 4.44$, $p < 0.05$). For ChatGPT, regular versus transcreation transaction effects do not differ significantly ($\chi^2 = 2.42$, $p = 0.12$). The chi-square tests are estimated on the appended common sample of regular and transcreation observations using a model with task $\times$ time and task $\times$ language pair fixed effects; the coefficients displayed in columns (1)–(8) are from the task-specific subsamples.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Figure 4. (Color online) The Effect of GNMT and ChatGPT on Regular Translation Tasks and Revenue

Notes. (a) The effect of GNMT and ChatGPT on regular translation tasks. (b) The effect of GNMT and ChatGPT on regular translation revenue. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $\gamma = \alpha + \sum z \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ is Total Transactions in panel (a) and Total Dollar Amount in panel (b), γ_i is language pair fixed effects for GNMT and language pair–task fixed effects for ChatGPT, δ_t is month-year fixed effects, and ε_{it} is the error term. The index z represents the months relative to a reference month, which marks the month when each technology first becomes available (November 2016 for GNMT cohort 1, March 2017 for GNMT cohort 2, April 2017 for GNMT cohort 3, November 2022 for ChatGPT). The dashed vertical line indicates the treatment month.

Figure 5. (Color online) The Effect of GNMT and ChatGPT on Transcreations and Transcreations Revenue

Notes. (a) The effect of GNMT and ChatGPT on transcreations. (b) The effect of GNMT and ChatGPT on transcreations revenue. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $\gamma = \alpha + \sum z \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ is Total Transactions in panel (a) and Total Dollar Amount in panel (b), γ_i is language pair fixed effects for GNMT and language pair–task fixed effects for ChatGPT, δ_t is month–year fixed effects, and ε_{it} is the error term. The index z represents the months relative to a reference month, which marks the month when each technology first becomes available (November 2016 for GNMT cohort 1, March 2017 for GNMT cohort 2, April 2017 for GNMT cohort 3, November 2022 for ChatGPT). The dashed vertical line indicates the treatment month.

Table 3. Heterogeneous Effect of GNMT and ChatGPT on Different Translation Tasks: General vs. Technical Translation

	(1) Gen Txn (GNMT)	(2) Tech. Txn (GNMT)	(3) Gen Rev (GNMT)	(4) Tech. Rev (GNMT)	(5) Gen Txn (ChatGPT)	(6) Tech. Txn (ChatGPT)	(7) Gen Rev (ChatGPT)	(8) Tech. Rev (ChatGPT)
Posttreatment	−0.0844** (0.0400)	−0.0418 (0.102)	−0.294** (0.149)	−0.144 (0.360)	−0.238*** (0.0515)	−0.299*** (0.0702)	−0.297** (0.132)	−0.366** (0.157)
Constant	4.245*** (0.0132)	2.272*** (0.0348)	9.796*** (0.0465)	8.546*** (0.0903)	4.666*** (0.0181)	2.618*** (0.0121)	10.69*** (0.0412)	9.062*** (0.0253)
Language pair fixed effects	Yes	Yes	Yes	Yes	No	No	No	No
Month–year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language pair–treatment fixed effects	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R^2	0.920	0.760	0.826	0.719	0.938	0.770	0.919	0.790
Log-likelihood	−3,016.4	−1,438.7	−1,761,759.4	−826,701.6	−7,187.9	−4,658.9	−3,598,151.9	−2,322,953.7
Observations	1,428	1,428	1,428	1,428	3,366	3,366	3,332	3,332

Notes. Standard errors are robust, clustered at the language pair level for GNMT and at the language–pair–treatment group level for ChatGPT. Columns (1) and (5): general translation transactions; columns (2) and (6): technical transactions; columns (3) and (7): revenue from general translation; columns (4) and (8): revenue from technical translation. Observation counts may differ across columns because the PPML estimator drops singleton observations and observations that are separated by a fixed effect and because some language pair–month cells have no observable price for the revenue regressions. Significant coefficients indicate that GNMT reduced general translation transactions by 8.1% ($\beta = -0.084$, $p < 0.05$) and revenue by 25.5% ($\beta = -0.294$, $p < 0.05$). ChatGPT reduced general translation transactions by 21.2% ($\beta = -0.238$, $p < 0.001$), revenue by 25.7% ($\beta = -0.297$, $p < 0.05$), technical translation transactions by 25.9% ($\beta = -0.299$, $p < 0.001$), and revenue from technical translations by 30.7% ($\beta = -0.366$, $p < 0.05$). Coefficient comparison tests: For GNMT, general versus technical translation effects do not differ significantly ($\chi^2 = 0.19$, $p = 0.662$). For ChatGPT, general versus technical translation effects do not differ significantly ($\chi^2 = 1.35$, $p = 0.246$). The chi-square tests are estimated on the appended common sample of general and technical translation observations using a model with task $\times$ time and task $\times$ language pair fixed effects; the coefficients displayed in columns (1)–(8) are from the task-specific subsamples.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 4. CEM Sample Baseline Effects of GNMT and ChatGPT on Translation Markets

	(1) Total translation transactions (GNMT)	(2) Total revenue (GNMT)	(3) Total translation transactions (ChatGPT)	(4) Total revenue (ChatGPT)
<i>Posttreatment</i>	−0.144*** (0.0465)	−0.122 (0.173)	−0.237*** (0.0524)	−0.306** (0.133)
<i>Constant</i>	3.850*** (0.0116)	9.611*** (0.0435)	4.218*** (0.0179)	10.21*** (0.0399)
Language pair fixed effects	Yes	Yes	No	No
Month-year fixed effects	Yes	Yes	Yes	Yes
Language pair–task fixed effects	No	No	Yes	Yes
Pseudo R^2	0.899	0.800	0.920	0.898
Log-likelihood	−1,942.0	−1,343,422.4	−7,599.6	−3,633,591.4
Observations	1,037	1,037	4,250	4,199

Notes. Standard errors are robust to heteroskedasticity and clustered at the language pair level for GNMT and at the language pair–task level for ChatGPT. Sample includes CEM-matched observations only. Observation counts may differ across columns because the PPML estimator drops singleton observations and observations that are separated by a fixed effect and because some language pair–month cells have no observable price for the revenue regressions. Significant coefficients indicate that GNMT reduced total translation transactions by 13.5% ($\beta = -0.144$, $p < 0.01$) with no significant effect on revenue ($\beta = -0.122$, $p = 0.483$); ChatGPT reduced total translation transactions by 21.1% ($\beta = -0.237$, $p < 0.001$) and revenue by 26.3% ($\beta = -0.306$, $p < 0.05$).
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

allows each freelancer to serve as the freelancer’s own control when assessing the impact of generative AI.

Table 6 reports the individual-level results. For GNMT, we observe an 8.6% decline in freelancer transactions associated with treated language pairs following GNMT’s introduction (coefficient = -0.0903 , $p < 0.05$). This pattern is consistent with the market-level estimates, indicating that aggregate reductions in translation activity are reflected in decreased demand for individual freelancers. For ChatGPT, we find a 9.1% reduction in

translation transactions for pure play tasks relative to bundled translation tasks at the individual level (coefficient = -0.0949 , $p < 0.05$).

The smaller magnitudes observed in the individual-level analysis relative to the market-level effects likely reflect the conservative nature of the sample construction. By focusing on established freelancers with prior platform experience, the analysis emphasizes workers who may be less exposed to substitution pressures than marginal or new entrants. Such freelancers may

Table 5. CEM Sample Heterogeneous Effect of GNMT and ChatGPT on Different Translation Tasks: Regular Translations vs. Transcreations

	(1) Reg Txn (GNMT)	(2) Transc Txn (GNMT)	(3) Reg Rev (GNMT)	(4) Transc Rev (GNMT)	(5) Reg Txn (ChatGPT)	(6) Transc Txn (ChatGPT)	(7) Reg Rev (ChatGPT)	(8) Transc Rev (ChatGPT)
<i>Posttreatment</i>	−0.0989*** (0.0375)	0.253 (0.176)	−0.116 (0.136)	0.0863 (0.294)	−0.226*** (0.0518)	−0.392*** (0.0927)	−0.311** (0.135)	−0.120 (0.235)
<i>Constant</i>	3.619*** (0.00780)	0.435*** (0.0381)	9.423*** (0.0280)	6.499*** (0.0772)	4.229*** (0.0177)	2.183*** (0.0118)	10.17*** (0.0395)	8.741*** (0.0445)
Language pair fixed effects	Yes	Yes	Yes	Yes	No	No	No	No
Month-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language pair–task fixed effects	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R^2	0.843	0.411	0.715	0.571	0.909	0.690	0.878	0.722
Log-likelihood	−1,929.2	−663.0	−1,412,882.4	−141,209.2	−6,390.2	−3,968.0	−3,236,350.1	−2,273,340.1
Observations	799	799	799	782	2,788	2,856	2,788	2,856

Notes. Standard errors are robust, clustered at the language pair level for GNMT and at the language pair–treatment group level for ChatGPT. Sample includes CEM-matched observations only. Columns (1) and (5): regular translation (general and technical) transactions; columns (2) and (6): transcreation transactions; columns (3) and (7): revenue from regular translation; columns (4) and (8): revenue from transcreation. Observation counts may differ across columns because the PPML estimator drops singleton observations and observations that are separated by a fixed effect and because some language pair–month cells have no observable price for the revenue regressions. Significant coefficients indicate that GNMT reduced regular translation transactions by 9.4% ($\beta = -0.099$, $p < 0.01$) with no other significant effects. ChatGPT reduced regular translation transactions by 20.2% ($\beta = -0.226$, $p < 0.001$), transcreation transactions by 32.4% ($\beta = -0.392$, $p < 0.001$), and regular translation revenue by 26.7% ($\beta = -0.311$, $p < 0.05$).
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 6. Individual-Level Estimates: GNMT vs. ChatGPT

	(1) Transactions (GNMT)	(2) Transactions (ChatGPT)
<i>Posttreatment</i>	−0.0903** (0.0425)	−0.0949** (0.0429)
<i>Constant</i>	0.307*** (0.0131)	0.169*** (0.0141)
Individual fixed effects	Yes	No
Month-year fixed effects	Yes	Yes
Individual-task fixed effects	No	Yes
Pseudo R^2	0.351	0.378
Log-likelihood	−34,325.8	−52,353.8
Observations	40,086	75,888

Notes. Standard errors are robust, clustered at the individual level for GNMT and at the individual-task group level for ChatGPT. Individual-level results show GNMT reduced freelancer transactions in treated languages by 8.6% ($\beta = -0.0903$, $p < 0.05$) and ChatGPT reduced freelancer translation transactions relative to bundled translation tasks by 9.1% ($\beta = -0.0949$, $p < 0.05$).

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

benefit from existing client relationships, accumulated reputation, or greater capacity to adjust their work practices in response to new technologies.

Event study estimates shown in Figure 6 illustrate the dynamic evolution of these effects. In both the GNMT and ChatGPT settings, treatment effects emerge shortly after technology introduction and persist throughout the observation window.

5.4.3. Other Robustness Checks. We conduct a series of additional robustness checks to assess the sensitivity of our findings to alternative specifications, identification strategies, and potential confounds.

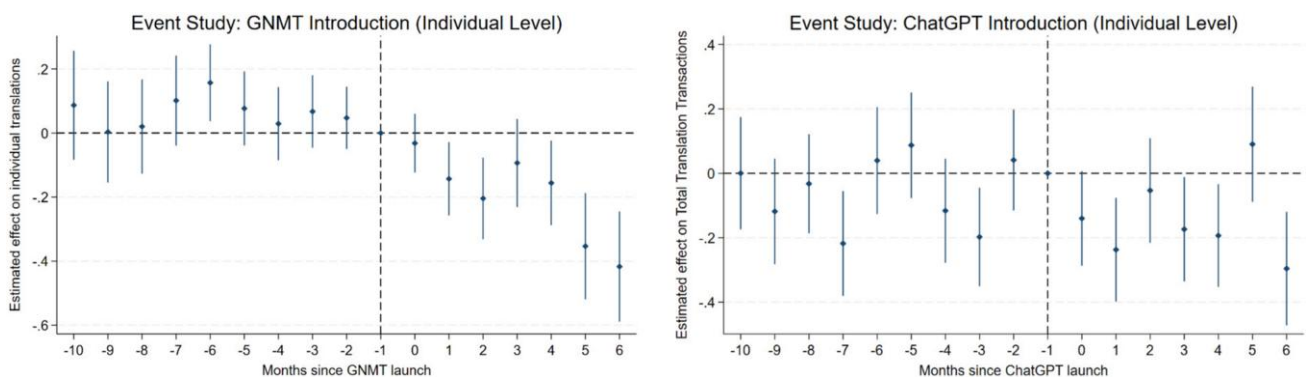
First, we test an alternative treatment timing for the Chinese–English language pair. Public announcements

from Google’s Brain Team indicate that the first deployment of GNMT occurred for Chinese–English translation in September 2016 prior to the broader rollout. Reestimating our models using this earlier treatment date yields results consistent with the baseline estimates (Online Tables A4-1 through A4-3).

Second, because the GNMT rollout involved staggered treatment timing across language pairs, we address recent concerns about bias in two-way fixed effects estimators under treatment effect heterogeneity (Goodman-Bacon 2021, Baker et al. 2022, de Chaisemartin and D’Haultfoeuille 2026). We reestimate the GNMT effects using Wooldridge’s (2023) extended difference-in-differences estimator, which treats each treatment cohort separately and avoids problematic comparisons between early and late treated units. This approach is asymptotically equivalent to the group time average treatment effects framework of Callaway and Sant’Anna (2021) and accommodates the Poisson pseudo-maximum-likelihood specification required by the prevalence of zero outcomes in our data. The resulting estimates are qualitatively similar to those from our baseline models (Online Tables A5-1 through A5-3).

Third, in addition to transaction volume and revenue, we examine the effects of GNMT and ChatGPT on average prices. As reported in Online Table A6-1, we do not detect statistically significant changes in average prices for either technology. This pattern is consistent with prior evidence that freelancers in online labor markets tend to adjust quantities before prices in response to demand shocks (Cullen and Faronato 2021).

Fourth, we estimate an alternative panel differences specification for the ChatGPT analysis, comparing outcomes in the post-ChatGPT period (2022–2023) with a

Figure 6. (Color online) Coefficient Estimates at the Individual Level Before and After GNMT and ChatGPT Introductions

Notes. The vertical axis shows Poisson coefficients (and 90% confidence intervals). The horizontal axis shows time relative to treatment. The coefficients are obtained from a regression of $Total\ Transactions = \alpha + \sum \beta_z \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i is individual fixed effects for GNMT and individual-task fixed effects for ChatGPT, δ_t is month-year fixed effects, and ε_{it} is the error term. The index z represents the lag or the months relative to a reference month, which marks the month when each technology first becomes available (November 2016 for GNMT cohort 1, March 2017 for GNMT cohort 2, April 2017 for GNMT cohort 3, November 2022 for ChatGPT). The dashed vertical line indicates the treatment month. GNMT analysis includes freelancers active in treated languages, whereas ChatGPT analysis compares translation tasks to bundled translation tasks for the same freelancers.

pre-COVID baseline period (2018–2019). This approach follows recent work using long-difference designs in related contexts (Eichenbaum et al. 2024, Goldberg et al. 2024). The estimates remain consistent with our main findings (Online Tables A7-1 through A7-3; Online Figure A7-1, (a) and (b)).

Fifth, we conduct placebo tests by assigning a false treatment date six months prior to the actual introduction of GNMT and ChatGPT and reestimating the models using only pretreatment data. These placebo analyses yield no statistically significant effects, supporting the validity of our identification strategy (Online Table A8-1).

Sixth, because monthly transaction data are inherently noisy, particularly for smaller language pairs with infrequent but high-variance demand, we aggregate the data to the quarterly level and reestimate the event study specifications. This substantially reduces measurement noise, and the resulting estimates are consistent with the monthly results (Online Tables A12-1 and A12-2, Online Figures A12-1 and A12-2).

Seventh, to address the large number of zero observations in the revenue data, we reestimate revenue effects using a Cragg's hurdle ("churdle") model. The first stage models the probability of observing positive revenue, and the second stage models revenue conditional on being positive using an inverse hyperbolic sine transformation. The results confirm our baseline findings: ChatGPT significantly reduces revenue for both regular and technical translations, whereas GNMT's effects are smaller and less precisely estimated (Online Tables A9-1 and A9-2).

Eighth, to assess sensitivity to task classification, we reestimate our main specifications using an independent classifier based on bidirectional encoder representations from transformers (BERT).¹⁰ The results are qualitatively identical to the baseline classification: GNMT significantly reduces demand for regular translations but has no significant effect on transcreation (Online Table A15-1).

Ninth, to ensure comparability across our two research designs, we replicate the bundled task identification strategy used in the ChatGPT analysis on the GNMT sample. Specifically, we stack translation tasks (treatment) and performance tasks (control) within the same regression with language pair $\times$ task type fixed effects, mirroring the ChatGPT specification. The results are consistent with our primary GNMT estimates (Online Tables A13-1 and A13-2, Online Figure A13-1).

Tenth, we address the potential confound that Microsoft introduced neural machine translation for several major language pairs in November 2016. Excluding language pairs overlapping with Microsoft's rollout yields no significant GNMT effect on total or regular translation. A triple-difference specification that explicitly interacts GNMT exposure with Microsoft overlap

shows that the substitution effect is concentrated in overlapping language pairs (Online Table A14-1), indicating that increased availability of high-quality neural machine translation in major commercial markets amplified substitution effects.

Finally, to assess representativeness with respect to translation demand more broadly, we replicate our analysis using Google Trends data. For each treated and control language pair, we collect search query volumes for language pair-specific Google Translate terms over the January 2016–May 2017 period. Consistent with increased adoption of machine translation following GNMT's introduction, search intensity rises for treated language pairs after treatment (Online Table A10-1, Online Figure A10-1).

6. Discussion and Conclusion

This study advances our understanding of AI's substitution effects on human labor by providing task-level empirical evidence from two natural experiments in the domain of language translation. Exploiting the staggered introduction of a predictive AI system (GNMT) and the later introduction of a generative AI system (ChatGPT), we show that both technologies substitute for nonroutine cognitive work but in systematically different ways. Whereas GNMT primarily reduced demand for analytical tasks (i.e., regular translations), ChatGPT reduced demand for both analytical and intuitive tasks (i.e., transcreations) at comparable magnitudes. These findings highlight that the substitution effects of AI are heterogeneous across tasks and depend critically on the underlying architecture of the technology being deployed (Autor 2014, Frank et al. 2019, Choudhury et al. 2020, Acemoglu et al. 2022).

Importantly, the magnitude of these effects is economically meaningful. GNMT reduced demand for regular translation by approximately 9% within seven months of introduction, having no detectable effect on transcreation. By contrast, ChatGPT is associated with declines of more than 20% for regular translation and nearly 30% for transcreation over a comparable horizon. This shift is not merely qualitative (i.e., affecting more task types) but quantitative in terms of both effect size and in the compression of task-level heterogeneity. Under predictive AI, substitution effects were sharply differentiated across tasks; under generative AI, declines are large and broadly similar across analytical and intuitive work. This convergence in magnitudes suggests a fundamental change in the structure of task-level substitution rather than an incremental extension of earlier automation trends.

A central insight of our analysis is that the impact of ChatGPT should be interpreted as the marginal effect of generative AI in a labor market already shaped by six years of exposure to predictive AI. By 2022,

predictive AI had already captured much of the analytical component of translation work. The additional reductions observed following ChatGPT's introduction, particularly in transcreation tasks, therefore reflect an expansion of the scope of AI substitution rather than a reallocation within previously automated tasks. This sequential comparison allows us to isolate what is genuinely new about generative AI relative to earlier automation waves. Our goal is, therefore, not to compare the absolute magnitudes of GNMT and ChatGPT effects as if they were independent shocks to an unaltered market but to use the sequential structure to identify how the boundaries of AI substitution evolved as the underlying technology shifted from predictive to generative architectures.

Our findings lend empirical support to the distinction between analytical convergent and intuitive divergent tasks. GNMT, which is optimized for convergent sequence-to-sequence prediction, substitutes effectively for analytical translation but leaves intuitive transcreation largely unaffected. By contrast, ChatGPT's generative architecture enables substitution across both task types. Importantly, this does not imply that generative AI perfectly replicates human intuition or cultural understanding; rather, it indicates that the level of performance required for substitution in market settings may be lower than the level required for parity on controlled benchmarks. As recent work on multilingual language models also suggests, generative systems can still falter when meaning hinges on subtle cultural or emotional cues (NLLB Team 2024).

This distinction helps reconcile our findings with evidence from benchmark-based and experimental studies. Whereas specialized models may outperform generative models on narrowly defined analytical benchmarks, our results show that a general purpose system such as ChatGPT can generate economically meaningful substitution effects for intuitive work in real labor markets. This suggests that task substitution in applied, interactive, and client-facing contexts may operate differently than performance in controlled evaluations.

Our results complement recent aggregate labor market evidence. Frey and Llanos-Paredes (2025), for example, document reduced translator employment growth following the diffusion of Google Translate with an instrumental variable estimate indicating a 0.84 percentage point reduction in employment growth per unit of adoption and an estimated 28,000 fewer positions created over the 2010–2023 period. Our task-level evidence reveals the mechanism underlying this aggregate pattern: predictive AI substitutes for analytical translation tasks, leaving intuitive work largely intact. This distinction is not observable in occupation-level data but becomes visible when labor demand is examined at the task level.

6.1. Managerial Implications

Beyond their theoretical contribution, our findings have direct implications for managerial decision making. For much of the past decade, a dominant managerial heuristic, rooted in experience with predictive AI, held that automation primarily threatens analytical work, whereas intuitive or creative tasks provide a safe harbor for human labor (Brynjolfsson and McAfee 2014, Frey and Osborne 2017). Our results show that this heuristic no longer generalizes in the era of generative AI. In this sense, our comparative design serves a belief-updating function: it provides empirical evidence that the task boundaries inferred from the predictive AI era do not hold under generative AI rather than comparing alternative software tools. Importantly, this belief-updating insight is not merely retrospective but forward looking. Specifically, if the empirical substitution boundary shifted once—from analytical tasks under predictive AI to both analytical and intuitive tasks under generative AI—managers should anticipate that it may shift again as AI systems continue to evolve. Therefore, processes such as workforce planning, reskilling, and task design frameworks built on a single generation of AI capabilities risk being overtaken by the next. Given this, the longitudinal lens we adopt here is best read as a template for continuous recalibration rather than a one-off diagnosis.

For managers, the key implication is that automation risk can no longer be assessed solely based on whether a task is creative or analytical. Instead, managers must evaluate the specific capabilities of the AI systems entering their domain and how those capabilities map onto the tasks embedded within jobs. Strategies that rely on reallocating workers from analytical tasks into creative roles as a hedge against automation may be effective under predictive AI but substantially less reliable under generative AI.

This insight has practical consequences for workforce planning, task redesign, and reskilling initiatives. Rather than organizing work around a simple human-versus-machine division, organizations may need to adopt more granular task-level strategies that identify where humans add value relative to rapidly evolving AI capabilities. In the translation context, for example, the growth of postediting services for both regular translation and transcreation suggests a shift toward hybrid roles that combine domain expertise with AI supervision and quality control (International Organization for Standardization 2017).

More broadly, our findings suggest that managerial responses to AI should emphasize organizational flexibility rather than task insulation. As AI systems evolve, previously reliable boundaries between human and machine work may erode. Managers who anchor their strategies to fixed notions of human-only tasks risk being repeatedly surprised as AI capabilities expand.

For individual workers, our results imply that specialization in tasks previously considered automation-resistant may no longer provide durable protection. Competitive advantage may instead arise from complementary skills such as domain knowledge, judgment, oversight, or integration that allow workers to work alongside AI systems rather than compete directly with them (Argyres and Zenger 2012, Teodoridis 2018).

At a broader level, our study contributes to ongoing debates about whether AI primarily substitutes for or augments human labor. By showing that generative AI can substitute for both analytical and intuitive tasks in real labor markets, we demonstrate that augmentation and substitution are not mutually exclusive outcomes but may coexist and evolve over time as technologies mature. More broadly, the task-level substitution effects we document in written translation may represent an early stage of a wider transformation in human communication. As AI-powered translation technologies extend from text-based platforms to real-time speech interfaces, such as AI translation-powered wearable devices capable of instantaneous verbal translation, similar substitution dynamics may emerge in domains well beyond formal labor markets. This trajectory raises important open questions for future research: whether widespread AI-mediated translation reduces incentives to learn foreign languages, whether it reshapes patterns of linguistic convergence or divergence, and whether the current global trend toward English as a dominant lingua franca may weaken if any language can be instantaneously translated using AI. Whereas these questions lie outside the scope of our data, our findings provide a task-level empirical foundation for understanding how advances in AI translation may reshape not only professional labor markets but also broader patterns of cross-cultural communication.

In sum, by comparing two generations of AI within the same labor market, our study provides a longitudinal perspective on how the boundaries of machine substitution shift as AI architectures evolve. This perspective clarifies why conclusions drawn from the predictive AI era do not straightforwardly generalize to the generative AI era and why managers, workers, and scholars alike must continually update their assumptions about the future of knowledge work.

6.2. Limitations

Our study has several limitations that point to opportunities for future research. First, our empirical context—language translation on a major OLM—has characteristics that may shape the observed effects. Freelance knowledge workers on OLMs differ from traditional employees in their contractual arrangements, pricing mechanisms, and exposure to market competition. Likewise, translation is particularly well suited to our

identification strategy because of its clear task demarcations and discrete technological shocks. Whereas this context enables precise task-level analysis, the magnitude and timing of substitution effects may differ in other forms of knowledge work.

Second, our analysis focuses on language-based AI technologies (GNMT and ChatGPT). Although language-related AI has advanced rapidly, other domains such as speech recognition, computer vision, or robotics may follow different substitution trajectories. Extending a comparative, task-level approach to these domains is an important direction for future research.

Third, our estimates capture short- to medium-run effects following each technological introduction. Our decision to focus on a relatively contained posttreatment window was driven primarily by concerns about internal validity: extending the analysis horizon substantially increases exposure to confounding forces that are difficult to isolate from the treatment effect. Our event study specifications show no evidence of attenuation within the observation window, suggesting persistent rather than transient effects. Nevertheless, as AI models continue to evolve, training data expand, and complementary organizational practices adapt, longer run substitution and adjustment dynamics may differ from those documented here. Our contribution is, therefore, to establish clean task-level baselines for the predictive and generative AI eras against which longer run analyses can be compared.

Fourth, whereas we distinguish between regular (analytical) translation and transcreation (intuitive) tasks, this boundary is not always sharp in practice. Some translation projects combine literal conversion with creative adaptation. However, any resulting misclassification is unlikely to vary systematically across treatment and control groups or over time. Moreover, classification error in the dependent variable induces attenuation bias, shrinking estimated effects toward zero rather than exaggerating them. Correcting for symmetric misclassification rates of 7%–13% yields larger implied effects than those reported (Online Appendix A11). In addition, misclassification is asymmetric—transcreation tasks are more likely to be misclassified as regular than vice versa—biasing estimates toward finding effects on transcreation. Our null results for GNMT on transcreation are, therefore, conservative.

Fifth, translators often work across multiple language pairs. If freelancers use machine translation tools for supported language pairs as an intermediate input when completing work in unsupported pairs, spillovers to the control group could arise. Consistent with this view, approximately 14.3% of freelancers transact in both treated and control language pairs during the GNMT window. We acknowledge this as a limitation but note that any such spillover mechanism would operate similarly across task types within the

same language pairs and, therefore, cannot explain the differential effects we observe between analytical and intuitive tasks.

Sixth, ChatGPT received substantially greater public attention than GNMT, which was introduced as a backend improvement to an existing product. Differences in awareness and adoption speed may, therefore, contribute to the larger observed effects for ChatGPT. From a managerial perspective, however, adoption dynamics are an integral component of a technology's market impact. Our event study evidence indicates that effects emerge following actual availability rather than prelaunch media coverage. Our objective is to estimate the realized impact of AI technologies as they enter and diffuse into labor markets, which is precisely the quantity relevant for managerial decision making.

Finally, because the GNMT (2016–2017) and ChatGPT (2022–2023) shocks occurred in different macroeconomic environments, cross-technology comparisons should be interpreted as suggestive rather than definitive evidence of architectural differences. Relatedly, because Microsoft introduced neural machine translation for several overlapping language pairs in November 2016, the estimated GNMT effects for those pairs capture the joint impact of increased availability of high-quality neural machine translation rather than the effect of Google's system alone (see Online Appendix A14).

Taken together, these limitations do not undermine our core contribution. Rather, they clarify the scope of our claims and highlight the value of task-level, longitudinal analyses for understanding how advances in AI reshape the boundaries of substitution for knowledge work.

Acknowledgments

For helpful comments during the development of this manuscript, the authors thank Department Editor Joshua Gans and the editorial review team, Victor Bennett, Raj Choudhury, Chiara Farronato, Alfonso Gambardella, Nan Jia, Hyunjin Kim, Karim Lakhani, Milan Miric, Tim Meyer, Hakan Ozalp, Christian Peukert, and Phanish Puranam as well as audiences at Bocconi, Erasmus, Norwegian BI, HEC Lausanne, Harvard Business School, the AI Strategy Group, and the Academy of Management and Strategic Management Society conferences.

Endnotes

¹ Our choice of GNMT and ChatGPT as comparative cases is theoretically motivated rather than exhaustive. These two systems represent distinct and widely recognized architectural regimes—task-specific predictive AI and general-purpose generative AI—whose comparison is designed to capture a structural shift in AI capabilities and its implications for task substitution rather than to provide a comprehensive enumeration of translation tools. We do not include earlier or intermediate systems because doing so would weaken identification without sharpening this theoretical contrast.

² Microsoft, as did Google, introduced NMT for multiple language pairs in November 2016, covering English paired with Arabic,

Mandarin Chinese, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish. Seven of these language pairs overlap with those initially supported by GNMT: English paired with Mandarin Chinese, French, German, Portuguese, Japanese, Korean, and Spanish. GNMT additionally covered the Turkish–English pair, which was not included in Microsoft's initial NMT rollout. In our empirical analysis, language pairs not covered by GNMT are retained in the control group. To the extent that Microsoft's NMT affected these language pairs, this approach would bias our estimates toward zero, rendering our findings conservative.

³ The details of the launch of GNMT for different language pairs can be found in Online Appendix A1.

⁴ Details of the task classification procedure, including prompts, validation steps, and robustness checks, are provided in Online Appendix A2.

⁵ We acknowledge the possibility of classification error and its implications for our estimates. To assess this concern, we conducted a supplemental validation exercise. We drew a simple random sample of 500 tasks using uniform probability across (i) task category (regular, technical, transcreation), (ii) language pair, and (iii) calendar year. Two research assistants who were blind to the algorithmic classification and to whether a task fell in the pretreatment or post-treatment period independently coded each observation. Agreement between the algorithm and the human coders was 86.8% and 93.2%, respectively, and intercoder agreement was 94.2%. These results suggest that any remaining misclassification is limited and unlikely to drive our main findings. Within the translation sample, every job posting is assigned to exactly one task category; there is no residual or ambiguous classification that would lead to selective exclusion of observations.

⁶ Details on the timing of GNMT's rollout across language pairs are provided in Online Appendix A1.

⁷ Spillovers between treated and control language pairs are possible in our setting as translators often work across multiple language pairs. Our identification strategy does not require the control group to be entirely unaffected by GNMT's introduction, only that treated and control pairs would have followed parallel trajectories absent treatment. Event study analyses support this assumption. We discuss spillovers further as a limitation in Section 6.2.

⁸ As a robustness check, we implement the bundled task identification strategy used in the ChatGPT analysis on the GNMT sample (Online Appendix A13). Specifically, we define translation tasks as the treated group and performance-related tasks as the control group, mirroring the ChatGPT specification. The resulting estimates are qualitatively similar to our baseline results: regular translation exhibits a negative effect (−7.4%, $p=0.057$), whereas the effect on transcreation remains statistically indistinguishable from zero (−0.8%, not significant). These results indicate that our findings are not sensitive to the choice of identification strategy.

⁹ This pattern is consistent with broader evidence on the translation industry. Frey and Llanos-Paredes (2025) document that U.S. translator employment continued to grow following the diffusion of Google Translate but at a significantly reduced rate relative to the overall economy (3% versus 13% over 2010–2023). Their instrumental variable estimates capture a reduction in the growth rate of translator employment, not an absolute decline, mirroring the dynamics we observe in our setting. In particular, the online translation market expanded overall during our study period, but treated language pairs grew more slowly than untreated pairs following the introduction of GNMT. Because our difference-in-differences design identifies treatment effects from this relative slowdown rather than from absolute levels, control group growth is not a threat to identification but rather reflects the counterfactual growth trajectory against which the treatment effect is measured.

¹⁰ We clarify that this classification relies solely on job descriptions, not on transaction outcomes or performance measures, and is, therefore, orthogonal to treatment timing.

References

- Accenture (2023) A new era of generative AI for everyone. Accessed July 27, 2025, <https://www.accenture.com/content/dam/accelture/final/accelture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf>.
- Acemoglu D, Autor D (2011) Skills, tasks and technologies: Implications for employment and earnings. Ashenfelter O, Card D, eds. *Handbook of Labor Economics*, vol. 4B (Elsevier, Amsterdam), 1043–1171.
- Acemoglu D, Restrepo P (2020) Robots and jobs: Evidence from US labor markets. *J. Political Econom.* 128(6):2188–2244.
- Acemoglu D, Autor D, Hazell J, Restrepo P (2022) Artificial intelligence and jobs: Evidence from online vacancies. *J. Labor Econom.* 40(S1):S293–S340.
- Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Press, Boston).
- Agrawal A, Gans J, Goldfarb A (2023) Do we want less automation? *Science* 381(6654):155–158.
- Agrawal A, Lacetera N, Lyons E (2016) Does standardized information in online markets disproportionately benefit job applicants from less developed countries? *J. Internat. Econom.* 103:1–12.
- Allison PD, Waterman RP (2002) Fixed-effects negative binomial regression models. *Sociol. Methodology* 32(1):247–265.
- Angrist JD, Pischke JS (2009) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ).
- Argote L, Lee S, Park J (2021) Organizational learning processes and outcomes: Major findings and future research directions. *Management Sci.* 67(9):5399–5429.
- Argyres NS, Zenger TR (2012) Capabilities, transaction costs, and firm boundaries. *Organ. Sci.* 23(6):1643–1657.
- Autor D (2014) Polanyi's paradox and the shape of employment growth. NBER Working Paper No. 20485, National Bureau of Economic Research, Cambridge, MA.
- Autor D, Salomons A (2018) Is automation labor-displacing? Productivity growth, employment, and the labor share. NBER Working Paper No. 24871, National Bureau of Economic Research, Cambridge, MA.
- Autor D, Levy F, Murnane RJ (2003) The skill content of recent technological change: An empirical exploration. *Quart. J. Econom.* 118(4):1279–1333.
- Baker AC, Larcker DF, Wang CC (2022) How much should we trust staggered difference-in-differences estimates? *J. Financial Econom.* 144(2):370–395.
- Benson A, Sojourner A, Umyarov A (2020) Can reputation discipline the gig economy? Experimental evidence from an online labor market. *Management Sci.* 66(5):1802–1825.
- Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589–1607.
- Brynjolfsson E, McAfee A (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (W. W. Norton, New York).
- Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *Quart. J. Econom.* 140(2):889–942.
- Brynjolfsson E, Mitchell T, Rock D (2018) What can machines learn, and what does it mean for occupations and the economy? *AEA Papers Proc.* 108:43–47.
- Brynjolfsson E, Rock D, Syverson C (2021) The productivity J-curve: How intangibles complement general purpose technologies. *Amer. Econom. J. Macroeconomics* 13(1):333–372.
- Brzeski C, Burk I (2015) Die Roboter kommen: Folgen der Automatisierung für den deutschen Arbeitsmarkt. Report, ING-DiBa Economic Research, Frankfurt am Main.
- Callaway B, Sant'Anna PH (2021) Difference-in-differences with multiple time periods. *J. Econometrics* 225(2):200–230.
- Cameron AC, Trivedi PK (2013) *Regression Analysis of Count Data*, 2nd ed. (Cambridge University Press, Cambridge, UK).
- Chen DL, Horton JJ (2016) Research note—Are online labor markets spot markets for tasks? A field experiment on the behavioral response to wage cuts. *Inform. Systems Res.* 27(2):403–423.
- Chen J, Roth J (2024) Logs with zeros? Some problems and solutions. *Quart. J. Econom.* 139(2):891–936.
- Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management J.* 41(8):1381–1411.
- Cullen Z, Farronato C (2021) Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms. *Management Sci.* 67(7):3985–4003.
- Dahl M, Magesh V, Suzgun M, Ho DE (2024) Large legal fictions: Profiling legal hallucinations in large language models. *J. Legal Anal.* 16(1):64–93.
- de Chaisemartin C, D'Haultfoeuille X (2026) Difference-in-differences estimators of intertemporal treatment effects. *Rev. Econom. Statist.* 108(4):863–880.
- Dell'Acqua F, Ayoubi C, Lifshitz H, Sadun R, Mollick E, Mollick L, Han Y, et al. (2026) The cybernetic teammate: A field experiment on generative AI and teamwork. *Organ. Sci.* 37(4):1217–1242.
- Demirci O, Hannane J, Zhu X (2025) Who is AI replacing? The impact of generative AI on online freelancing platforms. *Management Sci.* 71(10):8097–8108.
- Dixon J, Hong B, Wu L (2021) The robot revolution: Managerial and employment consequences for firms. *Management Sci.* 67(9):5586–5605.
- Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* 10(28):eadn5290.
- Eichenbaum M, Godinho de Matos M, Lima F, Rebelo S, Trabandt M (2024) Expectations, infections, and economic activity. *J. Political Econom.* 132(8):2571–2611.
- Eloundou T, Manning S, Mishkin P, Rock D (2024) GPTs are GPTs: Labor market impact potential of LLMs. *Science* 384(6702):1306–1308.
- Faraj S, Pachidi S, Sayegh K (2018) Working and organizing in the age of the learning algorithm. *Inform. Organ.* 28(1):62–70.
- Fedyk A, Hodson J, Khimich N, Fedyk T (2022) Is artificial intelligence improving the audit process? *Rev. Accounting Stud.* 27(3):938–985.
- Felten E, Raj M, Seamans R (2021) Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management J.* 42(12):2195–2217.
- Felten E, Raj M, Seamans R (2023) How will language modelers like ChatGPT affect occupations and industries? Preprint, submitted March 2, <https://arxiv.org/abs/2303.01157>.
- Frank MR, Autor D, Bessen JE, Brynjolfsson E, Cebrian M, Deming DJ, Feldman M, et al. (2019) Toward understanding the impact of artificial intelligence on labor. *Proc. Natl. Acad. Sci. USA* 116(14):6531–6539.
- Franko J, Du S, Kallweit S, Duelberg E, Engemann H (2020) Design of a multi-robot system for wind turbine maintenance. *Energies* 13(10):2552.
- Frey CB, Llanos-Paredes P (2025) Lost in translation: Artificial intelligence and the demand for foreign language skills. Working paper, Oxford Martin School, University of Oxford, Oxford, UK.
- Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? *Tech. Forecasting Soc. Change* 114:254–280.

- Garicano L, Wu Y (2012) Knowledge, communication, and organizational capabilities. *Organ. Sci.* 23(5):1382–1397.
- Girotra K, Meincke L, Terwiesch C, Ulrich KT (2023) Ideas are dimes a dozen: Large language models for idea generation in innovation. Preprint, submitted August 2, <https://doi.org/10.2139/ssrn.4526071>.
- Glăveanu V, Lubart T, Bonnardel N, Botella M, De Biais PM, Desainte-Catherine M, Georgsdottir A, et al. (2013) Creativity as action: Findings from five creative domains. *Frontiers Psych.* 4:176.
- Glorioso C (2023) Fake news: ChatGPT has a knack for making up phony anonymous sources. *NBC New York* (February 23), <https://www.nbcnewyork.com/investigations/fake-news-chatgpt-has-a-knack-for-making-up-phony-anonymous-sources/4120307/>.
- Goldberg SG, Johnson GA, Shriver SK (2024) Regulating privacy online: An economic evaluation of the GDPR. *Amer. Econom. J. Econom. Policy* 16(1):325–358.
- Goldfarb A, Taska B, Teodoridis F (2023) Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings. *Res. Policy* 52(1):104653.
- Goodman-Bacon A (2021) Difference-in-differences with variation in treatment timing. *J. Econometrics* 225(2):254–277.
- Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum likelihood methods: Theory. *Econometrica* 52(3):681–700.
- He VF, Shrestha YR, Puranam P, Miron-Spektor E (2024) Creative together: Human-AI ensembling for idea generation. Working paper, INSEAD, Fontainebleau, France.
- Holland D, Quinn N, eds. (1987) *Cultural Models in Language and Thought* (Cambridge University Press, Cambridge, UK).
- Horton JJ, Tambe P (2025) The death of a technical skill. *Inform. Systems Res.* 36(3):1799–1820.
- Hui X, Reshef O, Zhou L (2024) The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Organ. Sci.* 35(6):1977–1989.
- Hunter W (2023) What poets know that ChatGPT doesn't. *The Atlantic* (February 13), <https://www.theatlantic.com/books/archive/2023/02/chatgpt-ai-technology-writing-poetry/673035/>.
- International Organization for Standardization (2017) ISO 18587:2017 Translation services—Post-editing of machine translation output—Requirements. Accessed April 18, 2025, <https://www.iso.org/standard/62970.html>.
- Jia N, Luo X, Fang Z, Liao C (2024) When and how artificial intelligence augments employee creativity. *Acad. Management J.* 67(1): 5–32.
- Jia Y, Carl M, Wang X (2019) Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation* 33(1):9–29.
- Kochhar R (2023) Which U.S. workers are more exposed to AI on their jobs? Accessed February 18, 2025, <https://www.pewresearch.org/social-trends/2023/07/26/which-u-s-workers-are-more-exposed-to-ai-on-their-jobs/>.
- Kokkodis M, Ipeirotis PG (2016) Reputation transferability in online labor markets. *Management Sci.* 62(6):1687–1706.
- Korinek A (2023) Generative AI for economic research: Use cases and implications for economists. *J. Econom. Literature* 61(4): 1281–1317.
- Krakowski S, Luger J, Raich S (2023) Artificial intelligence and the changing sources of competitive advantage. *Strategic Management J.* 44(6):1425–1452.
- Lewis-Kraus G (2016) The great A.I. awakening. *The New York Times Online* (December 14), <https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>.
- Luong MT, Manning CD (2015) Stanford neural machine translation systems for spoken language domains. Federico M, Stüker S, Niehues J, eds. *Proc. 12th Internat. Workshop Spoken Language Translation: Evaluation Campaign (IWSLT 2015)* (Da Nang, Vietnam), 76–79.
- Marcus G, Davis E (2019) *Rebooting AI: Building Artificial Intelligence We Can Trust* (Pantheon, New York).
- Martin W, Pham CS (2020) Estimating the gravity model when zero trade flows are frequent and economically determined. *Appl. Econom.* 52(26):2766–2779.
- Massey G, Kiraly D (2019) The future of translator education: A dialogue. *Cultus* 2019(12):15–34.
- Massey G, Wieder R (2019) Quality assurance in translation and corporate communications: Exploring an interdisciplinary interface. Huertas-Barros E, Vandepitte S, Iglesias-Fernández E, eds. *Quality Assurance and Assessment Practices in Translation and Interpreting* (IGI Global, Hershey, PA), 57–87.
- NLLB Team (2024) Scaling neural machine translation to 200 languages. *Nature* 630(8018):841–846.
- Pajarinen M, Rouvinen P (2014) Computerization threatens one third of Finnish employment. ETLA Brief No. 22, Research Institute of the Finnish Economy, Helsinki.
- Pedersen D (2014) Exploring the concept of transcreation—Transcreation as “more than translation.” *Cultus* 7:57–71.
- Peukert C, Abeillon F, Haese J, Kaiser F, Staub A (2024) Strategic behavior and AI training data. CESifo Working Paper No. 11099, Center for Economic Studies, Munich.
- Powell D, Seabury S (2018) Medical care spending and labor market outcomes: Evidence from workers' compensation reforms. *Amer. Econom. Rev.* 108(10):2995–3027.
- Raj M, Seamans R (2018) Artificial intelligence, labor, productivity, and the need for firm-level data. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 553–565.
- Santos Silva JMC, Tenreiro S (2006) The log of gravity. *Rev. Econom. Statist.* 88(4):641–658.
- Santos Silva JMC, Tenreiro S (2011) Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Econom. Lett.* 112(2):220–222.
- Schwarz D, Manning S, Barry P, Cleveland DR, Prescott JJ, Rich B (2026) AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice. *J. Law Empirical Anal.* 3(1):220–250.
- Spitz-Oener A (2006) Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *J. Labor Econom.* 24(2):235–270.
- Teodoridis F (2018) Understanding team knowledge production: The interrelated roles of technology and expertise. *Management Sci.* 64(8):3625–3648.
- Tong S, Jia N, Luo X, Fang Z (2021) The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management J.* 42(9):1600–1631.
- Tschang FT, Almirall E (2021) Artificial intelligence as augmenting automation: Implications for employment. *Acad. Management Perspect.* 35(4):642–659.
- Vieira LN (2020) Automation anxiety and translators. *Translation Stud.* 13(1):1–21.
- Wang S, Yang DM, Rong R, Zhan X, Fujimoto J, Liu H, Minna J, Wistuba II, Xie Y, Xiao G (2019) Artificial intelligence in lung cancer pathology image analysis. *Cancers* 11(11):1673.
- Webb M (2019) The impact of artificial intelligence on the labor market. Preprint, submitted November 15, <https://doi.org/10.2139/ssrn.3482150>.
- Wooldridge JM (2023) Simple approaches to nonlinear difference-in-differences with panel data. *Econom. J.* 26(3):C31–C66.
- World Economic Forum (2023) The future of jobs report 2023. Accessed October 26, 2025, <https://www.weforum.org/publications/the-future-of-jobs-report-2023/>.
- Wu L, Hitt L, Lou B (2020) Data analytics, innovation, and firm productivity. *Management Sci.* 66(5):2017–2039.
- Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, et al. (2016) Google's neural machine translation system:

- Bridging the gap between human and machine translation. Preprint, submitted September 26, <https://arxiv.org/abs/1609.08144>.
- Yamada M (2019) The impact of Google neural machine translation on post-editing by student translators. *J. Specialised Translation* 31:87–106.
- Yao B, Jiang M, Bobinac T, Yang D, Hu J (2024) Benchmarking machine translation with cultural awareness. Al-Onaizan Y, Bansal M, Chen YN, eds. *Findings Assoc. Comput. Linguistics: EMNLP 2024* (Association for Computational Linguistics, Stroudsburg, PA), 13078–13096.
- Yoon S-W (2017) AI programs take on human translators. *The Korea Times* (February 14), <https://www.koreatimes.co.kr/business/tech-science/20170214/ai-programs-take-on-human-translators>.