“My (22m) Girlfriend (23f) Comes Home and Does Nothing”– Gendered Perceptions of Paid and Unpaid Work in Reddit Relationship Discussions over Time

Draft Version.

This version: July 14, 2026

Birgit Zeyer Gliozzo [a], Johanna Hölzl [b], Gundula Zoch [c] [d], Philipp Doebler [e]

[a] TU Dortmund University, August-Schmidt-Straße 4, 44227 Dortmund, Orcid: 0000-0002-4537-7851

[b] University of Mannheim, A5, 6, 68159 Mannheim, Orcid: 0009-0002-4954-2209

[c] Carl von Ossietzky University Oldenburg, Institute for Social Sciences, Ammerlander Heerstraße 114-118, 26111 Oldenburg, Germany. [d] Leibniz Institute for Educational Trajectories (LIfBi), Department: Educational Decisions and Processes, Migration, Returns to Education, Wilhelmsplatz 3, 96047 Bamberg, Germany; Orcid: 0000-0002-4398-4535

[e] TU Dortmund University, August-Schmidt-Straße 4, 44227 Dortmund, Orcid: 0000-0002-2946-8526

Abstract:

The COVID-19 pandemic reignited longstanding debates around gender inequalities in paid and unpaid work. While survey research has advanced our understanding of these disparities, it typically relies on predefined categories and is susceptible to social desirability and recall bias. Online postings, by contrast, capture intimate experiences in real time but rarely include demographic attributes. We leverage a unique dataset of discussions in Reddit relationship communities that combines rich descriptions of conflicts around (un)paid work with demographic information to examine: Which topics do men and women discuss in relationship conflicts around paid and unpaid work? We use Large Language Models (LLMs) and systematically vary GPT-family models and prompting strategies to classify manifest (demographic information) and latent variables (whether posts discuss romantic relationships, paid, and unpaid work) in Reddit posts. Using classifications from the best-performing approach, we apply Structural Topic Models to explore which topics partners discuss and how discussions evolve over time. We find temporary pandemic shifts in topics, and women more often discuss mental health in paid work-related conflicts, while men more frequently

emphasize career objectives. In conflicts related to unpaid work, men more frequently express concerns about sexual intimacy. Our analysis offers new insights into topics driving relationship conflicts around (un)paid work. We also contribute to the growing literature on LLMs' capabilities and limitations in classifying social-science constructs. By combining demographic information with sensitive narratives, our dataset captures forms of relational conflict rarely accessible in survey or social-media research, opening promising avenues for future research.

Keywords: Gender inequality, Paid and household labor, LLM classification, Reddit data, Structural Topic Model, Relationship conflict

Acknowledgments: The authors would like to thank the Summer Institute in Computational Social Science (SICSS), Munich, 2023, for providing the platform that initiated this project. We would also like to thank Julian Kohne, Paul Binder, and Marc Sparhuber for their contributions during the initial analysis phase. We also acknowledge the efforts of our student assistants, Yannick Hilker, Anna Schürmeyer, Viktor Rudakov, and Marie Therese Thoma, in annotating posts and providing additional support, and Hannah Bartmann for her additional contributions. Finally, we thank the research group of Florian Keusch for their feedback on this paper.

Funding statement: The project "From Prediction to Agile Interventions in the Social Sciences (FAIR)" has received/is receiving funding from the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors. [Förderkennzeichen PROFILNRW-2020-068].

This publication was additionally funded by the TU Dortmund Young Academy.

Data availability statement: The dataset will be made available, subject to the outcome of an ethics proposal, upon a qualified request from the research community.

Conflict of interest: The authors declare no conflict of interest.

Ethics approval statement: Not applicable.

Author Credits:

Birgit Zeyer-Gliozzo: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Johanna Hölzl: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing – review & editing.

Gundula Zoch: Conceptualization, Writing – original draft, Writing – review & editing.

Philipp Doebler: Conceptualization, Funding acquisition, Writing – review & editing.

Introduction

The COVID-19 pandemic has renewed longstanding questions about gender inequalities in the division of paid and unpaid work.[1] As schools and daycare facilities closed worldwide, care demands surged – and were predominantly met by mothers (e.g., Huebener et al., 2024; Zoch et al., 2021). Additionally, social distancing measures and workplace closures contributed to widening gender inequalities in professional advancement (e.g., Haas et al., 2025; Hipp & Konrad, 2022). Together, these developments point to a shift towards more *traditional* gender roles during times of crisis (e.g., Huebener et al., 2024; Zoch et al., 2021). Research further highlights the pandemic's far-reaching consequences as women, particularly mothers, reported higher levels of stress, reduced well-being, and more severe mental health challenges compared to men (Ervin et al., 2024; Zoch et al., 2022). Greater and subjectively perceived unfair shares of maternal care and housework were also positively linked to relationship problems and decreased relationship satisfaction for women (Waddell et al., 2021).

So far, most studies on gender inequalities in the division of labor and associated mental health disparities rely on surveys and interviews facing notable limitations. Standardized surveys typically work with predefined categories, are prone to recall bias, and often underrepresent emotionally charged or stigmatized experiences (Acht & Rebaudo, 2025; Zoch, 2021). By contrast, people write about intimate experiences in considerable detail on anonymous online platforms such as Reddit when seeking advice. Such postings, however, rarely include demographic information and remain challenging to analyze at large scale.

We leverage the features of two relationship communities on the popular social media platform Reddit, which uniquely combine rich descriptions of relationship conflicts with demographic information. Using a large-scale collection of posts from the years before, during, and after the COVID-19 pandemic (2011-2023), we examine which topics women and men discuss in relationship conflicts around paid and unpaid work. We ask: How do women and men on Reddit discuss conflicts between paid work and private life, care and household responsibilities, and relationship tensions over time?

Reddit, launched in 2005, currently hosts more than over 100 million daily active users and more than 100,000 active thematic communities known as *subreddits*. Active users are

[1] In this paper, we use the terms "paid work" and "unpaid work". While the literature sometimes classifies volunteering as unpaid work, we focus specifically on household and caregiving tasks. The definitions and operationalization used for data annotation include related terms such as "paid labor" and "household labor", reflecting terminology often used in the literature on housework.

predominantly male, and almost half are based in the United States (Proferes et al., 2021; Reddit, 2025). On Reddit, users discuss a wide range of topics of personal interest, many of which have been extensively studied by the scientific community, including politics, mental health and drug abuse, sexual identity and gender roles, and popular culture. Owing to its thematic organization and unrestricted post length, the platform is particularly well suited for analyzing extensive textual data on specific topics and subpopulations (Amaya et al., 2021; Britt et al., 2023; Proferes et al., 2021). We focus on the subreddits *r/relationships* and *r/relationship_advice,* where users seek and provide advice on intimate relationship conflicts, including romantic, familial, friendship, and professional relationships. Reddit is known for its anonymity, as registration only requires a valid email address, and users can create multiple accounts, leading to the phenomenon of so-called "throwaway accounts" (Leavitt, 2015). The anonymity afforded by Reddit encourages users to describe sensitive relationship conflicts in great depth as they unfold, including those on paid and unpaid work, thereby reducing the risk of social desirability and recall bias. Reddit might also provide access to groups often underrepresented in conventional studies, particularly individuals experiencing crises, those with low institutional trust, and those reluctant to participate in formal studies. Importantly, the posting rules of these two subreddits require users to report the age and gender of both partners (r/relationship_advice, 2025; r/relationships, 2025), information that is typically not available in other online data sources and allows us to examine gender differences in conflict discussions.

Given the large volume of posts in these subreddits, we employ Large Language Models (LLMs) to automatically extract the demographic information and classify posts based on whether they discuss romantic relationships, paid work, and unpaid work. LLMs offer a flexible, accessible, and cost-efficient approach to text classification (Davidson, 2024). However, researchers applying LLMs to text data classification are confronted with a plethora of models (e.g., Alizadeh et al., 2025) and prompting approaches to experiment with. In addition, the classification of latent constructs has proven particularly challenging with LLMs (e.g., Törnberg, 2023). If human annotators do not agree due to ambiguity of language and subjectivity, then LLMs fine-tuned on human rater data might struggle as well. To identify the best-performing classification approach, we systematically compare models from OpenAI's GPT-family (o3, 4o, 4.1), prompting strategies, fine-tuned vs. base models, and context windows. We evaluate performance against human annotations using best-practice metrics (F1-Score, Cohen's Kappa, Matthews Correlation Coefficient) with bootstrapped confidence intervals. Building on the best-performing LLM classification, we apply Structural Topic

Modeling (STM; Roberts et al., 2016) to examine how gendered discourses evolve over time (2011-2023).

Our analysis provides new exploratory insights into the topics driving relationship conflicts surrounding paid and unpaid work discussed on Reddit, gender differences in these discussions, and the impact of external crises such as the COVID-19 pandemic. To this end, we construct a unique dataset of posts from Reddit relationship communities between 2011 and 2023, including information on relationship type and sociodemographic characteristics of the individuals involved in the conflicts. These data offer great potential for further analyses, for instance, on relationship dynamics in same-sex and opposite-sex couples, conflicts in parent-child relationships, post popularity and comments.[2] Methodologically, we contribute to the growing literature on the capabilities and limitations of LLMs for the classification of complex latent social science constructs as we show that LLMs classify manifest variables with high accuracy but struggle with the classification of rare, latent constructs. Our findings offer practical guidance for researchers seeking to apply LLMs to complex text classification and to critically assess their potential and limitations.

Gender Differences in Paid and Unpaid Work

Research has consistently documented persistent gender inequalities in the division of paid and unpaid work, which tend to widen over the life course, particularly around childbirth (Nitsche & Grunow, 2016). Even when women are equally engaged in paid work, they continue to perform more unpaid work than men (Schober & Zoch, 2019), a pattern shaped by persistent gender norms and institutional frameworks (Cooke & Baxter, 2010, for an overview). Although men's participation in household labor has increased, the division of unpaid work remains highly unequal, with women – both in married and unmarried couples – continuing to shoulder the majority of domestic and care responsibilities (e.g., Pailhé et al., 2021; Zoch & Heyne, 2023). These imbalances are associated with reduced relationship quality and satisfaction, increasing risk of separation, particularly for women (Frisco & Williams, 2003).

External shocks can further intensify these gender inequalities, as the COVID-19 pandemic has shown. As schools and daycare centers closed, care burdens rose sharply, predominantly for mothers, even in dual-earner households (e.g., Zoch et al., 2021). These added burdens led

[2] Out of ethical and privacy considerations, we do not include the post's author (user name), title and text in the published dataset. Upon legitimate request, researchers can additionally request the data set including the post's content.

to a temporary re-traditionalization of roles (Huebener et al., 2024) and had negative consequences for maternal well-being, including increased stress, reduced life satisfaction, and greater mental health challenges (Ervin et al., 2024, for an overview; Zoch et al., 2022).

Building on this body of research, our study addresses key gaps that traditional survey or interview-based studies often cannot fully capture, such as difficulties to continuously track social phenomena over time, predefined response categories, and social desirability and recall bias (Acht & Rebaudo, 2025; Zoch, 2021). In contrast, Reddit posts offer spontaneous, non-reactive, real-time reflections on emotionally charged topics related to gendered inequalities, household labor, parenting stress, work-family conflict, and relationship strain.

Previous research has shown that Redditors do discuss topics related to relationship stressors, paid work, and unpaid work. However, these strands of research have rarely been connected. For instance, Gorrepati et al. (2025) use Reddit data to examine the mental health consequences of relationship stressors such as breakups. Research on paid work has investigated topics such as attitudes toward remote work (Elbaz et al., 2025) and the disproportionate mental health burdens experienced by working mothers during the COVID-19 pandemic (Huang et al., 2023). Studies of unpaid work have focused primarily on parenting strain (e.g., Francisco, 2025; Teague & Shatte, 2021; Westrupp et al., 2022) and elder caregiving (Amorim et al., 2026; Hswen et al., 2024), with some also examining the effects of parenting on relationship quality and work-family conflict (Pilkington & Rominov, 2017). We extend this literature by investigating the topics that drive relationship conflicts arising from both paid and unpaid work. By tracing discourse patterns over time, we shed light on how gendered relationship dynamics evolved under conditions of crisis.

Data and Methods

Data

We focus on two large Reddit communities: *r/relationship_advice* and *r/relationships*, with 16 million and 3.6 million members in 2025, respectively. Both communities are dedicated to seeking and providing advice on relationship issues, including romantic, familial, friendship, and professional relationships. Although the majority of posts in both subreddits are written in English, users from around the world can post and comment. The rules of both subreddits require posts to include information about the age, gender, and length of relationships of those involved, as well as a brief summary (for longer posts) and a specific question or request for

advice (r/relationship_advice, 2025; r/relationships, 2025). Figure 1 shows a synthetic example post.

- Figure 1 here -

We obtained the Reddit data for both subreddits for the period 2009 to 2024 using publicly available torrent files collected by Pushshift and u/raiderbdev and hosted by academictorrents.com (Baumgartner et al., 2020; Watchful1, 2025). These files contain comprehensive archives of Reddit content. After removing erroneous entries (incorrectly formatted) and empty posts, our dataset included 2,351,608 posts.

From this data, we extracted the following variables using LLM annotation evaluated against manual coding (see Table B.1-1 in the Appendix for an overview, including detailed variable descriptions, definitions, range, and levels): Manifest variables include the author's age and gender and the age, gender, and role of the second protagonist mentioned in a post. The role is an open-ended response indicating who is involved in the relationship conflict (e.g., boyfriend, co-worker, mother). Complex variables refer to latent constructs. Posts were categorized as discussing paid work, unpaid work, and/or romantic relationships. We defined constructs based on theoretical considerations (see Section B.1.2. in the Appendix for definitions) and posts were classified by measuring the percentage of sentences addressing the respective topic.[3] Posts were assigned to a topic if at least 15 percent of the sentences addressed paid or unpaid work, or if at least 40 percent addressed romantic relationships. We determined these thresholds through extensive manual annotation, ensuring sufficient coverage of each topic within posts.

As a benchmark for the LLM classification and training data set for semi-automated classification, two researchers and four student assistants (BZG, JH, YH, AS, VR, and MT) manually annotated 1,854 posts, which were drawn randomly from the two subreddits. Following Grimmer et al. (2022), we developed the codebook iteratively by defining constructs and coding guidelines, coding a subset of posts, and refining the rules until no further adjustments were necessary. We then jointly coded 101 posts to assess intercoder reliability using Cronbach's alpha, achieving satisfactory values above 0.7 for all variables (Table B.3-1 in the Appendix).[4] Finally, different team members independently coded all remaining posts.

[3] In the definitions, we use the terms "paid labor" and "household labor", following terminology in the literature on housework.
[4] Cronbach's alpha ranges from 0.79 for the protagonist's role to 0.99 for the author's age.

After removing the handlabelled data, remaining erroneous entries (13 incorrectly formatted posts that had previously gone unnoticed), deleted authorship entries, duplicates, and non-English content, we ended up with a dataset of 2,124,285 posts for automated classification. From this, we took a random sample of 502,000 posts, stratified for the two subreddits, and automatically classified them. For our analyses, we filtered out posts with fewer than ten characters and posts from authors outside the age range of 18-69 years. Due to sparse data in earlier and later years, we restricted our analysis to posts from 2011-2023, leaving us with 474,944 posts.

For topic modeling, we focused on posts without comments discussing romantic relationships with the partner or ex-partner as the second protagonist and addressing paid or unpaid work. This restriction resulted in subsets of 294,715 posts about romantic relationships involving a partner or ex-partner, including 34,736 posts about paid work and 6,180 posts about unpaid work (see Appendix A.1 for full details). Figure 2 illustrates our methodological pipeline, covering all steps from data collection through preprocessing to analysis.

- Figure 2 here -

Methods

Classification of Reddit Posts with LLMs

The large amounts of text data in our dataset makes manual coding infeasible. LLMs' ease of use lowers technical barriers for researchers with less technical background and promises a flexible and efficient alternative to traditional text extraction methods. For semi-automated classification, we applied OpenAI's GPT models, a widely used LLM-family for processing and generating natural language. We decided on models from the GPT-family due to their ease of use, and because larger general-purpose LLMs have been shown to achieve better predictive performance on unseen tasks (e.g., Alizadeh et al., 2025; Chae & Davidson, 2025; Palmer et al., 2023). Moreover, several studies have promisingly applied LLM-based text classification to Reddit data, relying on models from the GPT-family as well (Corral et al., 2024; Li et al., 2024; Radwan et al., 2024). We based our prompt formulation on previous research (e.g., von der Heyde et al., 2025) and OpenAI's recommendations (OpenAI, n.d.).

The field of divergent LLMs and classification approaches remains fragmented, including a wide range of models, prompting strategies, fine-tuning approaches, and context window lengths. While only very few studies systematically compare their effectiveness (Chae & Davidson, 2025), most existing applications have focused on a single model or strategy.

Hence, it remains opaque which strategies yield satisfactory results and which will demonstrate the greatest performance in comparison. The classification of complex, latent constructs, which are often central to social science research, has proven particularly challenging with LLMs (e.g., Stuhler et al., 2025). We therefore tested several approaches for classification with GPT, varying (1) prompting setups (i.e., number of posts per prompt and instructions in the main or system message), (2) few-shot prompting (i.e., providing a few coded examples in the instructions covering a wide range of variable values) versus zero-shot prompting (giving no examples, (3) inclusion of variable descriptions in the instruction, and (4) different answer formats (i.e., instructing GPT to code the topic variables as either metric in line with human annotation, ordinal, or binary). Combining these choices resulted in a total of 36 distinct approaches (see Table C.1-1 in the Appendix).

Figure 3 shows how we constructed the prompts for each of the approaches. Additionally, we varied GPT models (o3, 4o and 4.1), the number of tokens (i.e., units of text that GPT processes) used in one query (30,000 vs. 120,000) to evaluate the impact of the context window length, and fine-tuned vs. non-fine-tuned models (OpenAI, 2025). We fine-tuned GPT-4o and 4.1 on OpenAI's platform using a random 60-20-20 split of the manually annotated data as training, test, and validation sets, three and four epochs, and OpenAI's default batch size and learning rate.[5] Fine-tuning promises improved performance on the specific annotation task, as the LLM's weights get updated as the model learns from the manually annotated data (e.g., Dodge et al., 2020; Radford et al., 2018).

- Figure 3 here -

We based our evaluation of the LLM's annotation performance on a sample of 350 human-annotated posts, using a diverse set of accuracy metrics, as each individual metric has its own strengths and weaknesses (e.g., Hand et al., 2024).[6] We calculated bootstrapped confidence intervals for each metric by drawing 10,000 random samples with replacement from the test data. To test the classification's reliability, we ran each approach twice for each GPT model with one context window length (30,000 tokens for GPT o3 and 4o and 120,000 tokens for GPT 4.1) and calculated intercoder reliability with Cohen's Kappa. We report model performance based on the average classification results of both runs. To decide on an approach for fine-tuning o4 and 4.1, we selected the best-performing approach from each base model across metrics. However, we prioritized metrics performing better on complex variable

[5] At the time of writing, OpenAI does not offer the option to also fine-tune GPT-o3.
[6] We evaluated each approach against around 20 percent of our manually annotated data, similar to the validation set split for the fine-tuned models.

classification (see Appendix Section C.1 for details on the LLM classification). We then compared results from all approaches and used the best-performing model to code all posts from our sample.

Overall, classification performance varied substantially depending on the type of category labeled, prompt design, context window length, GPT model, and whether models are fine-tuned or used in their base form (see Figure 4 and Section D.1 in the Appendix). However, our findings clearly indicate that the LLMs and particularly base models struggle with the classification of rare latent constructs (unpaid work), when providing several posts per prompt, and asking for ordinal responses. In contrast, simpler classification tasks (manifest categories), the classification of more frequently occurring topics (romantic relationship), shorter context windows, binary or metric response formats, and fine-tuning substantially improve classification performance.

- Figure 4 here -

From all approaches, GPT-4.1 fine-tuned for three epochs with few-shot prompting and category descriptions in the main prompt, a single post per prompt, a 30,000 token context window length, and a metric response format proved to be the most effective overall configuration. This approach showed strong classification performance on all manifest variables and in the classification of romantic relationships and paid work. All metrics reach or exceed 0.80, with at most a single lower confidence interval falling below 0.70 for all variables except for paid work, with all lower confidence intervals greater than 0.6. Fine-tuned GPT-4.1 also shows extremely high interrater reliability across two runs. However, even our strongest model configuration fails to achieve satisfactory performance on the latent constructs of paid and especially unpaid work. Notably, the latent constructs most challenging for the LLMs to classify are also those that occur the rarest and for which human annotators exhibit the lowest reliability. In classifying the remaining posts, we nevertheless instructed the fine-tuned GPT-4.1 to annotate all categories included in the initial prompt, as omitting categories could alter predictive accuracy.

Topic models

As the main focus of our analysis lies in how gender impacts the topics discussed in relationship conflicts around (un)paid work, we further restrict our dataset to posts that cover romantic relationships and either paid or unpaid work as classified by the LLM. We then used Structural Topic Models (STMs) to explore the topics in these posts, with separate analysis for

the posts on paid and unpaid work. STMs are based on the logic of Latent Dirichlet Allocation Models (LDA; Blei et al., 2003) to uncover the hidden topic structure in the given documents. STMs extend LDAs by allowing the incorporation of metadata into the model to improve topic estimation (Roberts et al., 2016). Specifically, covariate information can be included that affects the topic prevalence (i.e., a document's proportion devoted to a topic) and the topical content (i.e., the frequency of words used when discussing a topic). Document-topic proportions can thus vary as a function of observed covariates rather than arising from a single prior shared by all documents (Roberts et al., 2016). We included the following variables from the LLM classification as covariates for topic prevalence: the author's gender and age; the partner's gender, age, and role; and the time of posting, measured in quarters (three month intervals). We adopted a more flexible, nonlinear functional form for the time variable and estimated it with a spline (Roberts et al., 2019). We estimated the effects of the covariates using the `estimateEffect` function, which performs a regression with the topic proportions as the outcome variable (Roberts et al., 2014).

In STMs, the researcher must specify the number of topics k. Using a data-driven approach, we assessed k within the range of 5-20, evaluating metrics such as held-out likelihood, residuals, and semantic coherence (Roberts et al., 2019). As the focus of model selection is not on statistical fit, but on achieving substantively relevant results, we additionally reviewed the top candidate models manually, focusing on interpretability (Grimmer & Stewart, 2013; Maier et al., 2021). Based on the highest probability (top) words, FREX words (words weighted by frequency and exclusivity), and a close reading of posts with a high prevalence of a given topic of models with 10, 12, and 15 topics, we selected 15 topics for paid work and 15 topics for unpaid work (Roberts et al. 2019). Further details on the methodology used for STMs can be found in Appendix Section C.2.

For our analyses, we created a document-term matrix by preprocessing the data according to standard procedures, including tokenization, discarding punctuation, URLs, numbers, separators, and symbols, lowercasing, and filtering out stopwords (Blaheta & Johnson, 2001; Grimmer et al., 2022; Maier et al., 2021). Collocations were identified and selected based on their co-occurrence frequency and association strength (e.g., mental health). Subsequently, we applied stemming and removed terms with extremely high or low frequency. We also deleted entries in the token list with information on age and gender (e.g., 27f, 28m, nb35) after filtering out stopwords, as these did not add valuable information to our topic models.

Findings

Relationship Subreddits on Reddit

Of the 474,944 posts in our cleaned dataset, 62.1% concern romantic relationships with a current or former partner (hereafter "relationship posts"). From 2011 to 2023, the percentage of relationship posts out of all posts fluctuated modestly. It rose until early 2013, declined for several years, then climbed sharply to around 70 percent in early 2023. Within these posts, 11.8 percent discuss paid work, while 2.1 percent discuss unpaid work. While unpaid work discussions remained rare, their share gradually increased from 2020 onward. Paid work discussions fluctuated more. They rose until 2017, declined until 2020, and increased again thereafter.

Some of these dynamics overlapped with the onset of the pandemic. Based on WHO declarations and U.S. trends, we define the years up to and including 2019 as pre-pandemic, 2020-2022 as pandemic years, and 2023 as post-pandemic (JHU.edu, 2026; WHO, 2026). Post-pandemic, relationship discussions intensified, while attention to unpaid work began to rise during the pandemic.

Across all posts, women tend to post more than men, and the disparity is greatest in unpaid work posts (women post over twice as much as men). Therefore, Reddit users in relationship subreddits appear to differ from the average Reddit user, the majority of whom are male (Proferes et al., 2021). Most authors are young, with more than half aged 18-29, although paid and especially unpaid work posts come from slightly older authors, with a comparatively high proportion aged between 30 and 39. The age of the person discussed mirrors that of the authors, while the gender dynamic is reversed: women tend to write about men, and vice versa, reflecting that most conflicts are discussed within heterosexual relationships. Missing gender mentions are far less frequent for protagonists than authors (see Appendix Section D.2.1 for a detailed graphical overview of the trends and Section D.2.2 for changes in the demographic structure over time).

Gendered Discussions in Relationship Subreddits Over Time

Paid Work

To understand and label the topics discussed in relationship discourses on Reddit in the context of paid and unpaid work, we examined the top words, FREX words, and most typical posts for each topic (Roberts et al., 2019). Table 1 presents the topics for conversations related to paid work, ordered by expected topic proportions over all documents. The most prevalent

topic is "career and mobility choices" (11.6%). Here, for example, the top and FREX words include “career”, “futur”, “graduat”, and “abroad”. The top posts often include discussions on future career plans, tension between work and family, issues in long-distance relationships, and decisions about relocation for work. "Mental health issues" appear with a similar frequency (11.3%), with posts describing relationship strains linked to mental or physical illnesses experienced by oneself or one's partner and how work contributes to, exacerbates, or is affected by these circumstances. Other commonly discussed topics include “mistrust, jealousy, and insecurity” (9.7%), for instance with regard to coworkers or when both partners share the same workplace, “work and relationship time conflicts” (8.2%), e.g., leisure activities and time management affected by work, and “values, attitudes, appreciation, and boundaries” (8.1%), particularly regarding the separation of professional and private life. Additional topics include “internet and phone communication” (7.9%), the top post of which contains discussion about the (lack of) communication during and outside working hours, “finances and expenses” (7.4%), and “misaligned work schedules and daily routines” (6.8%).

Table 1: Paid labor topics ordered by average topic proportion

Topic	Title	Top Words	FREX Words	Average expected topic proportion across documents
1	Career and mobility choices	career, futur, school, citi, find, current, colleg, countri, graduat, stay	graduat, career, abroad, visa, countri, degre, program, internship, opportun, futur	11.6%
2	Mental health issues	depress, stress, better, support, break, bad, hard, emot, everyth, problem	depress, anxieti, mental_health, mental, therapi, therapist, suffer, cope, pain, medic	11.3%
3	Mistrust, jealousy, and insecurity	guy, girl, girlfriend, cowork, date, gf, ex, met, peopl, hang	cowork, girl, flirt, jealous, femal, guy, crush, co-work, male, insecur	9.7%
4	Work and relationship time conflict	plan, hour, weekend, spend, trip, away, girlfriend, night, date, visit	trip, weekend, birthday, saturday, friday, cancel, flight, sunday, holiday, summer	8.2%
5	Values, attitudes, appreciation, and boundaries	partner, person, issu, peopl, convers, understand, share, seem, howev, situat	partner, valu, boundari, topic, profession, express,	8.1%

			regard, intellig, view, dynam	
6	Internet and phone communication	text, call, phone, messag, went, convers, later, send, repli, answer	text, repli, messag, respond, phone, send, chat, delet, sent, answer	7.9%
7	Finances and expenses	money, pay, financi, save, bill, rent, car, paid, buy, expens	debt, payment, loan, bill, pay, paycheck, expens, credit_card, money, rent	7.4%
8	Misaligned work schedules and daily routines	sleep, hous, clean, bed, night, hour, dog, leav, around, morn	wake, dog, kitchen, wash, dish, breakfast, clean, nap, bed, sleep	6.8%
9	Cheating, lying, separations, and ex-partners	lie, found, cheat, ex, went, leav, left, trust, came, happen	lie, affair, forgiv, john, polic, truth, betray, proof, caught, discov	5.4%
10	Partners‘ leisure time activities	spend, play, hour, game, watch, sometim, around, full_tim, chore, hobbi	game, hobbi, play, video_gam, chore, free_tim, play_video_gam, playing_video_gam, lazi, play_gam	5.0%
11	Sex, desire, and intimacy	sex, anymor, sexual, attract, seem, great, initi, interest, chang, physic	sex, intimaci, intim, weight, sexual, sex_driv, affection, attract, libido, spark	5.0%
12	Extended family issues	famili, parent, mom, dad, hous, gf, mother, sister, brother, stay	brother, sister, dad, mom, parent, sibl, mum, mother, cousin, uncl	4.4%
13	Physical appearance and lifestyle habits	fuck, look, drink, smoke, use, shit, wear, peopl, weed, drug	smoke, weed, wear, drug, jack, hair, shirt, tattoo, model, dress	3.6%
14	Family formation, parenting, and marriage	wife, kid, babi, marri, son, child, daughter, marriag, divorc, children	wife, kid, daughter, daycar, son, babi, child, custodi, children, born	3.4%
15	(Married) partner’s job	husband, marri, compani, new, wed, busi, boss, marriag, manag, fianc	husband, tom, gt, spous, wed, husband’, jame, hubbi, h, hire	2.2%

By estimating correlations between topic proportions, the STM reveals which topics are likely to co-occur in posts, thereby deepening our understanding of their interrelations (Roberts et al. 2016, 2019; see Figure D.3-1 in the Appendix). For instance, “internet and phone

communication" appear together with "mistrust, jealousy, and insecurity" as well as "cheating, lying, separations, and ex-partners", while "partner's leisure time activities" co-occur with "misaligned work schedules and daily routines", and "family formation, parenting, and marriage" with "(married) partner's job". Conversely, "mental health issues" are negatively correlated with "mistrust, jealousy, and insecurity".

To compare topic prevalences between men and women, we examine a first difference estimate, which contrasts the topic proportions of the two groups. Figure 5 shows the differences in topic proportions when changing between men and women (Roberts et al., 2019). Differences are generally minor, i.e., men and women discuss most topics with similar frequency. The most pronounced discrepancies pertain to "mental health issues", which women discuss with greater frequency. Conversely, men engage in discourse on "career and mobility choices" significantly more frequently. These patterns reflect traditional gender norms. Men focus on career-related issues, while women focus on interpersonal topics (Davis & Greenstein, 2009). Notably, these topics tend to be discussed more frequently by younger users (see Figure D.3-2 in the Appendix). These findings might suggest that traditional gender norms persist or even reemerge among the younger generation of Reddit users.

Figure 6 illustrates the prevalence of topics over time as a smooth function of yearly quarters, emphasizing the onset and end of the Covid-19 pandemic. Crisis-related shifts are visible across almost all topics, indicating that external shocks influenced the discourse surrounding paid work in romantic relationships. Some topics appear to have been temporarily replaced by others. For instance, discussions about relationship strain related to "mental health issues" increased noticeably at the beginning of the pandemic, while topics such as "mistrust, jealousy and insecurity" and "work- and relationship-time conflict" decreased. However, by the end of the pandemic, these topics had largely returned to pre-pandemic levels, suggesting short-term crisis effects. Discussions about "career and mobility choices" have decreased steadily, slightly intensified by the pandemic, but recently rose again. This pattern may be linked to the recent withdrawal of many home office regulations and their impact on private life. "Internet and phone communication" have become increasingly important, especially during the pandemic, reflecting general digitalization trends and their impact on relationship conflicts surrounding paid work.

Unpaid work

Regarding relationship discussions related to unpaid work, "arguments" are the most frequently discussed topic (10.1%), i.e., posts about disputes, conflicts, and arguing behavior.

Top and FREX words include “yell”, “apolog”, “scream”, and “upset” (see Figure 7 and Table D.3-1 in the Appendix). This topic is closely followed by discussions about “(mental) health issues” (9.4%), “cleaning and laundry” (9.3%), and “division of household tasks” (9.1%). In most cases, there is a clear reference to the overarching topic of unpaid work, indicating that the LLM classification worked, despite not sufficing our performance criteria. For example, discussions about health issues include problems with one's own mental or physical health or that of one's partner, which entail restrictions in terms of household work. These discussions are followed by concerns about “division of labor, exhaustion, and fatigue” (8.3%), which mainly concern being proactive, laziness related to house- and care work, and exhaustion/tiredness from work and housework, and, slightly less frequently, “sexual intimacy” (7.6%), including conflicts around lack of sexual activity, mismatched sexual needs, feelings of (un)attraction, or feeling unloved, often linked to unequal divisions of work, “childcare and pregnancy” (7.5%), and “finances and expenses” (7.4%). Topic correlations indicate the strongest positive correlation between “childcare and pregnancy” and “(extended) family conflicts”, whereas “arguments” appear less likely with “division of household tasks” and “finances and expenses” (see Figure D.3-3 in the Appendix).

Similarly to paid work, men and women talk about these topics in the context of unpaid work with the same frequency (see Figure 7). This finding somewhat contradicts the expectation that women talk more about, e.g., the division of household tasks. Only “sexual intimacy” is discussed significantly more often by men. The stronger presence of sexual intimacy among men’s posts is consistent with previous research on gender stereotypes, which suggests that behaviours associated with a high level of sexual activity are more expected and viewed more positively in men than in women (Endendijk et al., 2020). Finally, looking at changes over time, times of crisis are also linked to shifts in the topics discussed, with some themes appearing to gain and others to lose importance (see Figure D.3-4 in the Appendix). However, most of the changes observed are statistically uncertain because confidence intervals for mean topic proportions substantially overlap.

Discussion

In this study, we use LLMs to extract demographic and content-related information of Reddit posts, before applying STMs to examine which topics women and men discuss in relationship conflicts around paid and unpaid work over time. To this end, we construct a unique dataset from two Reddit relationship communities, *r/relationships* and *r/relationship_advice,* which

combines rich descriptions of intimate relationship conflicts with demographic information about the partners involved. Current survey-based research is often constrained by predefined categories and underrepresents emotionally charged or stigmatized experiences. By contrast, social media data typically does not provide any sociodemographic information. Our dataset thus offers a combination rarely available in either survey or social-media research.

Our analysis identified 15 topics in discussions about relationship conflicts surrounding paid and unpaid work. In paid work contexts, the most prevalent topics are "career and mobility choices", "mental health issues", "mistrust, jealousy, and insecurity", and "work and relationship time conflicts". While men and women discuss most of these topics with similar frequency, there are significant gender differences aligned with traditional gender norms. Men emphasize professional issues, including career objectives, more often, while women focus more on interpersonal issues, such as mental health. In discussions about relationship problems related to unpaid work, "arguments" emerged as the most frequently discussed topic, followed by "mental health issues", "cleaning and laundry", and "division of household tasks". Correlations between topic prevalence offer additional insights. For instance, relationship conflicts around (lacking) internet and phone communication during or outside working hours are often discussed in conjunction with conflicts relating to (mis)trust, cheating and lying, and highlight the role of social media or smartphone communication in relationship quality. Additionally, in unpaid work discussions, “arguments” appear less often with “finances and expenses” and the “division of household tasks”, which one might assume to be central issues. As in paid work contexts, men and women discuss topics with equal frequency, except for "sexual intimacy," which was significantly more common among men. Pandemic-related shifts in topic prevalence are visible across almost all topics for both paid and unpaid work. This pattern underpins the validity of our findings and is consistent with research findings that point to a return to more traditional gender roles during the pandemic (e.g., Huebener et al., 2024; Zoch et al., 2021), suggesting that times of crisis reshape relationship dynamics. During the pandemic, certain topics were overshadowed by others, with most changes being only temporary. However, most changes in discussions of unpaid work were not statistically significant. Moreover, most pandemic-related changes in unpaid work discussions occurred with a delay, possibly because changes in this area unfold more slowly.

Research on gender inequalities typically centers on the unequal distribution of paid and unpaid work, perceived fairness, and, for example, its impact on health and relationship

quality (e.g., Ervin et al., 2024; Waddell et al., 2021; Zoch et al., 2021). Rather than measuring how labor is divided, our analysis maps which work-related tensions partners foreground in conflict, adding relational texture that surveys with predefined categories tend to miss. Our findings underscore the important role of (mental) health, which can be affected by work and can also affect it, leading to tension within relationships. Here, the mental health of the partner often seems to play a decisive role - a bidirectional, dyadic link rarely captured in surveys that treat respondents' wellbeing as an individual outcome. Unlike discussions about paid work, those about unpaid work tend to focus more on the division of tasks itself, which is consistent with the slow progress toward equality in this area. Nevertheless, somewhat unexpectedly, women do not discuss topics related to task division more often than men, although they do participate more often in discussions about unpaid work in general. Furthermore, conflicts surrounding unpaid work seem to extend beyond the mere division of tasks and their perceived fairness. Frequent discussions about “arguments” suggest that household work is a complex source of relationship tension and a black box that remains under-explored in current research.

Beyond our substantive findings on drivers of relationship conflicts around (un)paid work, we also contribute to the growing literature on the capabilities and limitations of LLMs for classifying social science constructs by contrasting performance on complex latent constructs with that on simpler, manifest ones. Our study shows that LLMs can be a valuable tool for classifying online content, particularly in datasets that exceed feasible human coding capacity. Consistent with prior recommendations (e.g., Chae & Davidson, 2025), researchers should systematically vary model choice and prompt design, including response formats, the provision of category descriptions, context window length, and few-shot versus zero-shot prompting, to identify the configuration that yields the strongest classification performance.

In our study, performance varies substantially across prompting strategies, depending on both the construct being classified and the specific GPT model used. Overall, GPT-family LLMs perform best when classifying simpler, manifest constructs and more frequently occurring latent topics, such as romantic relationships, particularly when shorter context windows are used, when response formats are binary or metric rather than ordinal, and when models are fine-tuned. By contrast, the LLMs struggle to classify infrequently occurring latent topics for which human annotators exhibit the lowest reliability as well, such as unpaid work, particularly in ordinal response formats, and when asked to classify multiple posts within a single prompt.

Using Reddit data to investigate gender disparities in relationship conflicts constitutes both a key strength and an important limitation of this study, rooted in the nature of the data source and the ethical considerations involved. The most fundamental limitation concerns the lack of generalizability and potential selection bias inherent in Reddit data. Reddit users represent a specific subpopulation, typically younger, more male, and predominantly based in the United States (Proferes et al., 2021; Reddit, 2025). In the two subreddits examined in this study, the predominance of English-language posts may also constitute a barrier for non-native speakers, even though users from around the world can, in principle, participate. Consequently, the findings cannot be generalized to the broader population. The gradual change in the composition of Reddit's user base over time, i.e., what Salganik (2017) terms population drift, complicates comparisons of topic prevalence across time. This issue is further exacerbated by evidence from Veselovsky and Anderson (2023), who show that Reddit gained new users during the COVID-19 pandemic who differ systematically from earlier cohorts in their usage patterns and interests.

Another limitation stems from the reliance on self-reported information in posts. Although Reddit's community rules require users to provide demographic information such as age and gender, these self-reports cannot be verified. Given the platform's anonymity, there is always a risk of trolling, exaggeration, or fictional storytelling. The reported relationship conflicts also reflect only one side of the interaction. The posts' authors present their subjective perspective and interpretation of a situation but we do not have access to their partners' perspectives. These factors may introduce noise into the dataset and must be considered when interpreting the results. Users who post in subreddits such as r/relationship_advice are also likely to do so during periods of conflict or crisis. The study thus captures discourses of relationship strain, rather than everyday experiences and more stable patterns of work-family organization. Even though users write about conflicts as they unfold, they also sometimes include experiences that happened across a longer period of time. We therefore cannot rule out some recall bias in how the users present past experiences.

Using social media data further raises ethical concerns as Reddit users are potentially unaware that the sensitive information they provide in the posts may be used for research purposes (Gliniecka, 2023; Proferes et al., 2021; Salganik, 2017). In this study, we take steps to mitigate the risk of deanonymization, including reporting only aggregate measures such as topic proportions and refraining from sharing usernames or reproducing posts verbatim.

Beyond the choice of data, our study also comes with several methodological limitations. The comparatively low predictive performance for unpaid work constitutes an important limitation for our subsequent analyses. The prevalence of specific topics cannot be determined prior to manual annotation and classification and researchers cannot anticipate which constructs may occur only rarely and therefore prove especially challenging for the LLM classification. Accordingly, substantive conclusions regarding housework-related conflicts derived from these classifications should be interpreted with caution. Additionally, we examine only a subset of GPT-family models, as larger, general-purpose LLMs have been shown to achieve stronger predictive performance on unseen tasks (e.g., Alizadeh et al., 2025; Chae & Davidson, 2025; Palmer et al., 2023). However, using OpenAI's proprietary, closed-source models is costly and yields issues regarding the findings' transparency and replicability. Future research should therefore assess the extent to which open-source LLMs and alternative machine learning classifiers can match or exceed the performance of closed-source LLMs in the classification of latent social science constructs. Finally, the STM serves as an exploratory tool capturing broad thematic structures. In the next step, future research should examine specific topics, such as "arguments" or discussions about mental health, in greater depth.

Overall, this study offers new, exploratory insights into gendered patterns in relationship conflicts around paid and unpaid work. By combining demographic information with highly sensitive narratives, such as partners' mental health, our dataset provides information rarely available in either survey or social-media research. Men and women discuss most relationship conflict topics with similar frequency. However, the differences that do emerge in topic prevalence are consistent with traditional gender stereotypes, despite Reddit's relatively young user demographic. The data's high temporal granularity allowed us to trace how crises shaped relationship problems concerning paid and unpaid work, mostly as short-term effects, and which issues gained or lost significance over time. These data also offer great potential for further analyses, for instance, on hard-to-reach populations and intimate relationship conflicts, such as same sex couple conflicts and parent-children-conflicts. Methodologically, our findings also provide practical insights into LLMs' capabilities and limitations in classifying complex social science constructs in social media data. Most notably, we show that LLMs excel at classifying manifest variables, such as age and gender, but struggle with the classification of rare, latent constructs, such as discussions of unpaid work in Reddit posts.

References

Acht, M., & Rebaudo, M. (2025). A gender gap in gender gaps: Social norms and housework reporting. *Empirical Economics*, *68*(6), 2977–3029. https://doi.org/10.1007/s00181-024-02710-z

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2025). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, *8*(17). https://doi.org/10.1007/s42001-024-00345-9

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New data sources in social science research: Things to know before working with reddit data. *Social Science Computer Review*, *39*(5), 943–960. https://doi.org/10.1177/0894439319893305

Amorim, D., Lam, K., Chary, A. N., Plys, E., & Shi, S. (2026). Positive aspects of caregiving: A qualitative analysis of reddit posts. *Journal of the American Geriatrics Society*, jgs.70358. https://doi.org/10.1111/jgs.70358

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Blaheta, D., & Johnson, M. (2001). Unsupervised learning of multi-word verbs. *Proceedings of the 39th Annual Meeting of the ACL*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, (3), 993–1022.

Britt, R. K., Franco, C. L., & Jones, N. (2023). Trends and challenges within reddit and health communication research: A systematic review. *Communication and the Public*, *8*(4), 402–417. https://doi.org/10.1177/20570473231209075

Chae, Y., & Davidson, T. (2025). Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*, 00491241251325243. https://doi.org/10.1177/00491241251325243

Cooke, L. P., & Baxter, J. (2010). "families" in international context: Comparing institutional effects across western societies. *Journal of Marriage and Family*, *72*(3), 516–536. https://doi.org/10.1111/j.1741-3737.2010.00716.x

Corral, P. G., Green, A., Meyer, H., Stoll, A., Yan, X., & Reuver, M. (2024). A few hypocrites: Few-shot learning and subtype definitions for detecting hypocrisy accusations in online climate change debates. In C. Klamm, G. Lapesa, S. P. Ponzetto, I. Rehbein, & I. Sen (Eds.), *Proceedings of the 4th workshop on computational*

linguistics for the political and social sciences: Long and short papers (pp. 45–60). Association for Computational Linguistics. https://aclanthology.org/2024.cpss-1.4/

Davidson, T. (2024). Start generating: Harnessing generative artificial intelligence for sociological research. *Socius: Sociological Research for a Dynamic World*, *10*, 23780231241259651. https://doi.org/10.1177/23780231241259651

Davis, S. N., & Greenstein, T. N. (2009). Gender ideology: Components, predictors, and consequences. *Annual Review of Sociology*, *35*(Volume 35, 2009), 87–105. https://doi.org/10.1146/annurev-soc-070308-115920

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2002.06305

Elbaz, S., Petracca, R., Argiropoulos, N., & Provost Savard, Y. (2025). Exploring reddit discussions of remote work and return to office in the canadian federal public service. *Community, Work & Family*, 1–20. https://doi.org/10.1080/13668803.2025.2561941

Endendijk, J. J., van Baar, A. L., & Deković, M. (2020). He is a stud, she is a slut! A meta-analysis on the continued existence of sexual double standards. *Personality and Social Psychology Review*, *24*(2), 163–190. https://doi.org/10.1177/1088868319891310

Ervin, J., Alfonzo, L. F., Taouk, Y., Maheen, H., & King, T. (2024). Unpaid caregiving and mental health during the COVID-19 pandemic—A systematic review of the quantitative literature. *PLOS ONE*, *19*(4), e0297097. https://doi.org/10.1371/journal.pone.0297097

Francisco, S. C. (2025). Hey mama, we hear you: An analysis of online parenting support conversations on reddit. *Families in Society: The Journal of Contemporary Social Services*, *106*(4), 1207–1228. https://doi.org/10.1177/10443894241254128

Frisco, M. L., & Williams, K. (2003). Perceived housework equity, marital happiness, and divorce in dual-earner households. *Journal of Family Issues*, *24*(1), 51–73. https://doi.org/10.1177/0192513X02238520

Gliniecka, M. (2023). The ethics of publicly available data research: A situated ethics framework for reddit. *Social Media + Society*, *9*(3), 20563051231192021. https://doi.org/10.1177/20563051231192021

Gorrepati, L. P., Sonani, R., Velugubantla, V., Potla, R. T., & Sathvik, M. (2025). Mental health and relations: Detection of mental health disorders related to relationship issues

through reddit posts. *Companion Proceedings of the ACM on Web Conference 2025*, 1885–1889. https://doi.org/10.1145/3701716.3717742

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data. A new framework for machine learning and the social sciences*. Princeton University PRESS.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Haas, C., Zoch, G., & Kleinert, C. (2025). *Gender gaps in job-related training during COVID-19? Longitudinal evidence on supply and demand changes from germany* (K54sy_v1). SocArXiv. https://doi.org/10.31235/osf.io/k54sy_v1

Hand, D. J., Christen, P., & Ziyad, S. (2024). *Selecting a classification performance measure: Matching the measure to the problem* (arXiv:2409.12391). arXiv. https://doi.org/10.48550/arXiv.2409.12391

Hipp, L., & Konrad, M. (2022). Has covid-19 increased gender inequalities in professional advancement? Cross-country evidence on productivity differences between male and female software developers. *Journal of Family Research*, *34*(1), 134–160. https://doi.org/10.20377/jfr-697

Hswen, Y., Xiong, J., Hurley, M., & T. Nguyen, T. (2024). Experiences of alzheimer's disease and related dementia family caregivers on reddit communities: A topic modeling and sentiment analysis. *Artificial Intelligence in Health*, *1*(3), 127. https://doi.org/10.36922/aih.3075

Huang, C., Bandyopadhyay, A., Fan, W., Miller, A., & Gilbertson-White, S. (2023). Mental toll on working women during the COVID-19 pandemic: An exploratory study using reddit data. *PLOS ONE*, *18*(1), e0280049. https://doi.org/10.1371/journal.pone.0280049

Huebener, M., Danzer ,Natalia, Pape ,Astrid, Schober ,Pia, Spiess ,C. Katharina, & and Wagner, G. G. (2024). Cracking under pressure? Gender role attitudes toward maternal employment during COVID-19 in germany. *Feminist Economics*, *30*(3), 217–254. https://doi.org/10.1080/13545701.2024.2349295

JHU.edu. (2026). *United states—Overview*. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/region/united-states

Leavitt, A. (2015). "this is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community. *Proceedings of the 18th*

ACM Conference on Computer Supported Cooperative Work & Social Computing, 317–327. https://doi.org/10.1145/2675133.2675175

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2), 1–36. https://doi.org/10.1145/3643829

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., & Häussler, T. (2021). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. In *Computational Methods for Communication Science*. Routledge.

Nitsche, N., & Grunow, D. (2016). Housework over the course of relationships: Gender ideology, resources, and the division of housework from a growth curve perspective. *Advances in Life Course Research*, *29*, 80–94. https://doi.org/10.1016/j.alcr.2016.02.001

OpenAI. (n.d.). *Prompt engineering*. OpenAI Platform. Retrieved https://platform.openai.com/docs/guides/prompt-engineering?prompt-example=prompt

OpenAI. (2025, August 1). *What are tokens and how to count them?* OpenAI Help Center. https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

Pailhé, A., Solaz, A., & Stanfors, M. (2021). The great convergence: Gender and unpaid work in europe and the united states. *Population and Development Review*, *47*(1), 181–217. https://doi.org/10.1111/padr.12385

Palmer, A., Smith, N. A., & Spirling, A. (2023). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Pilkington, P. D., & Rominov, H. (2017). Fathers' worries during pregnancy: A qualitative content analysis of reddit. *The Journal of Perinatal Education*, *26*(4), 208–218. https://doi.org/10.1891/1058-1243.26.4.208

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, *7*(2), 20563051211019004. https://doi.org/10.1177/20563051211019004

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radwan, A., Amarneh, M., Alawneh, H., Ashqar, H. I., AlSobeh, A., & Magableh, A. A. A. R. (2024). Predictive analytics in mental health leveraging LLM embeddings and

machine learning models for social media analysis: *International Journal of Web Services Research*, *21*(1), 1–22. https://doi.org/10.4018/IJWSR.338222

Roberts, M. E., Stewart, B. M., & and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B., & Tingley, D. (2014). *stm: Estimation of the structural topic model* (p. 1.3.7) [Dataset]. https://doi.org/10.32614/CRAN.package.stm

r/relationship_advice. (2025, June 26). *R/relationship_advice*. https://www.reddit.com/r/relationship_advice/

r/relationships. (2025, June 26). */R/relationships*. https://www.reddit.com/r/relationships/wiki/index/#wiki_.2Fr.2Frelationships

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schober, P. S., & Zoch, G. (2019). Change in the gender division of domestic work after mothers or fathers took leave: Exploring alternative explanations. *European Societies*, *21*(1), 158–180. https://doi.org/10.1080/14616696.2018.1465989

Stuhler, O., Ton, C. D., & Ollion, E. (2025). From codebooks to promptbooks: Extracting information from text with generative large language models. *Sociological Methods & Research*, *54*(3), 794–848. https://doi.org/10.1177/00491241251336794

Teague, S. J., & Shatte, A. B. R. (2021). Peer support of fathers on reddit: Quantifying the stressors, behaviors, and drivers. *Psychology of Men & Masculinities*, *22*(4), 757–766. https://doi.org/10.1037/men0000353

Törnberg, P. (2023). *How to use LLMs for text analysis* (arXiv:2307.13106). arXiv. https://doi.org/10.48550/arXiv.2307.13106

Veselovsky, V., & Anderson, A. (2023). Reddit in the time of COVID. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 878–889.

von der Heyde, L., Haensch, A.-C., Weiß, B., & Daikeler, J. (2025). Using large language models for coding german open-ended survey responses on survey motivation. *Survey Research Methods*, *19*(4), 355–370. https://doi.org/10.18148/srm/2025.v19i4.8568

Waddell, N., Overall, N. C., Chang, V. T., & Hammond, M. D. (2021). Gendered division of labor during a nationwide COVID-19 lockdown: Implications for relationship

problems and satisfaction. *Journal of Social and Personal Relationships*, *38*(6), 1759–1781. https://doi.org/10.1177/0265407521996476

Watchful1. (2025, February 20). *Separate dump files for the top 40k subreddits, through the end of 2024* [Reddit Post]. R/Pushshift. https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/

Westrupp, E. M., Greenwood, C. J., Fuller-Tyszkiewicz, M., Berkowitz, T. S., Hagg, L., & Youssef, G. (2022). Text mining of reddit posts: Using latent dirichlet allocation to identify common parenting issues. *PLOS ONE*, *17*(2), e0262529. https://doi.org/10.1371/journal.pone.0262529

WHO. (2026). *Coronavirus disease (COVID-19) pandemic*. https://www.who.int/europe/emergencies/situations/covid-19

Zoch, G. (2021). Gender-of-interviewer effects in self-reported gender ideologies: Evidence based on interviewer change in a panel survey. *International Journal of Public Opinion Research*, *33*(3), 626–636. https://doi.org/10.1093/ijpor/edaa017

Zoch, G., Bächmann, A.-C., & Vicari, B. (2021). Who cares when care closes? Care-arrangements and parental working conditions during the COVID-19 pandemic in Germany. *European Societies*, *23*(S1), S576–S588. https://doi.org/10.1080/14616696.2020.1832700

Zoch, G., Bächmann, A.-C., & Vicari, B. (2022). Reduced well-being during the COVID-19 pandemic – The role of working conditions. *Gender, Work & Organization*, *29*(6), 1969–1990. https://doi.org/10.1111/gwao.12777

Zoch, G., & Heyne, S. (2023). The evolution of family policies and couples' housework division after childbirth in Germany, 1994–2019. *Journal of Marriage and Family*, *85*(5), 1067–1086. https://doi.org/10.1111/jomf.12938

My (27F) partner (30M) assumes I should handle all housework because he pays more. How do I push back?

We both have demanding full-time jobs, but he earns significantly more, so he covers most of the rent while I handle smaller bills. Lately, he's started acting like that financial gap means I owe him domestic work.

He rarely helps with cleaning, says it "makes sense" for me to take care of the apartment since he's the primary earner. I'm constantly tidying up after both of us while he relaxes, and when I bring it up, he says I should contribute more financially if I want things to be "equal."

I'm exhausted and frustrated. Does a bigger paycheck justify opting out of shared responsibilities at home?

Figure 1: Exemplary Reddit post. Synthetic illustration based on r/relationship_advice (2025) and text created with the help of GPT.

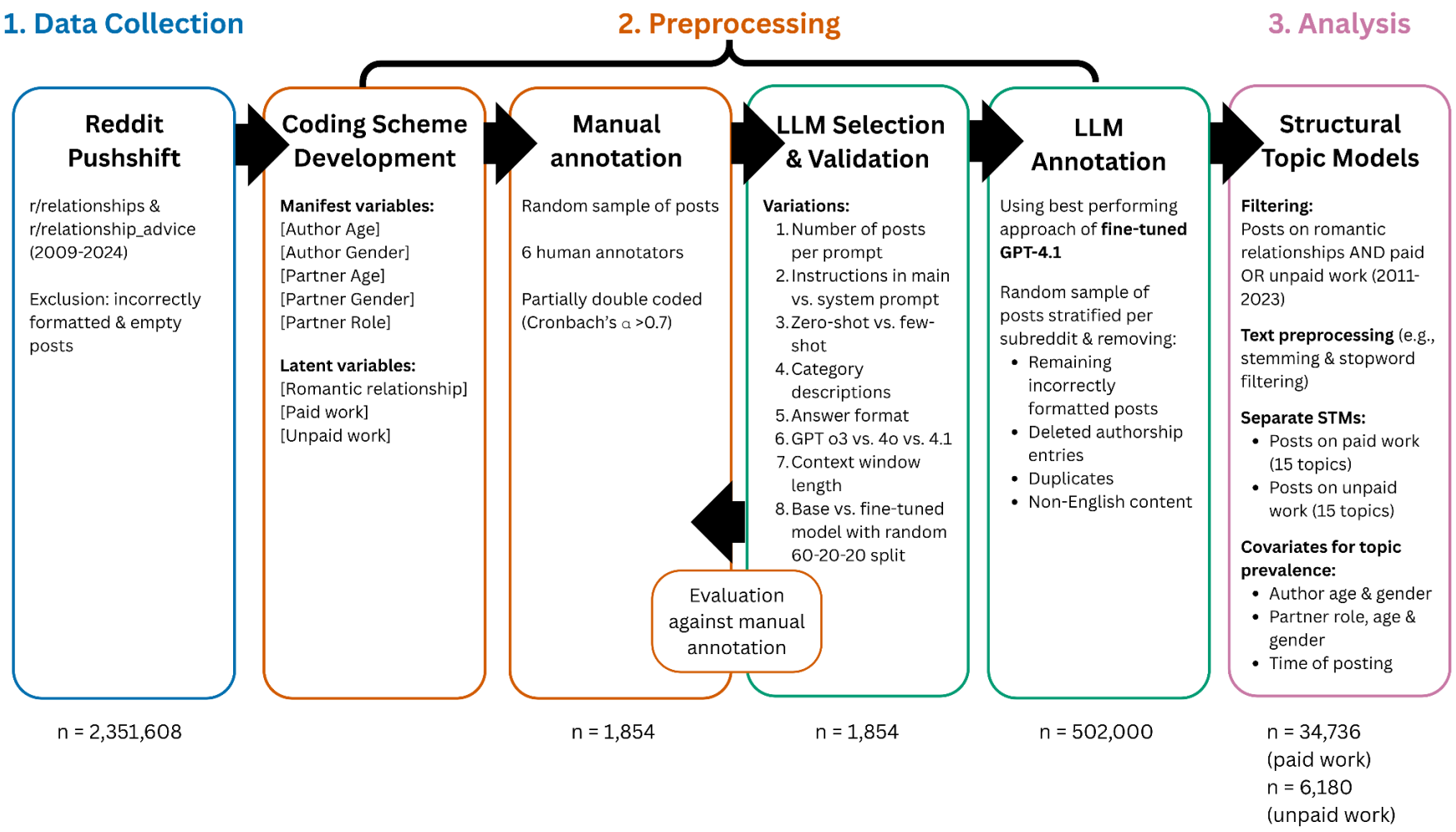

Figure 2: Workflow for collecting, preprocessing, and analyzing Reddit posts on romantic relationships, paid and unpaid work.
Note. Overview of the study workflow, including data collection (blue), preprocessing (orange: human annotation; green: large language model [LLM] annotation), and analysis (pink). N shown beneath each stage represents the number of posts included in that phase of the workflow. Data were collected from Pushshift and include posts from the subreddits r/relationships and r/relationship_advice (2009–2024), downloaded in August and September 2024.

System prompt:

You will be provided with posts from the social media platform Reddit. You are an expert in classifying the content of the posts.

Prompt:

Classify the following categories for each Reddit post:

AUTHOR'S AGE: [Description],
AUTHOR'S GENDER: [Description],
SECOND PROTAGONIST'S ROLE: [Description],
SECOND PROTAGONIST'S AGE: [Description],
SECOND PROTAGONIST'S GENDER: [Description],
WHETHER/TO WHAT EXTENT THE POST IS ABOUT PAID LABOR: [Description],
WHETHER/TO WHAT EXTENT THE POST IS ABOUT HOUSEHOLD LABOR: [Description],
WHETHER/TO WHAT EXTENT THE POST IS ABOUT ROMANTIC RELATIONSHIP: [Description]

General coding rules: Return a response in the following format: "Message {NR}: Age1; Gender1; Role2; Age2; Gender2; Paid_labor; Household_labor; Romantic_relationship". Respond NA in case of missing information for that category or respond NA for each category if a post is in a language other than English. Do not respond anything other than the string with the given format to each message. Respond separately for each message and separate messages with new line symbols.

Example 1: Message 1: "...", Response: "Message 1: 21; f; husband; 21; m; yes; yes; yes"'
Example 2: Message 2: "...", Response: "Message 2: 23; m; wife; 22; f; no; yes; yes"')
Example 3: Message 3: "...", Response: "Message 3: 24; f; boyfriend; 28; m; yes; no; yes"')
Example 4: Message 4: "...", Response: "Message 4: 25; m; partner; 21; f; no; yes; yes"')
Example 5: Message 5: "...", Response: "Message 5: 25; f; boyfriend; 32; m; yes; yes; yes"')
Example 6: Message 6: "...", Response: "Message 6: NA; NA; ex; NA; m; no; no; yes"')
Example 7: Message 7: "...", Response: "Message 7: NA; f; NA_p; NA_p; NA_p; no; no; yes")

Post/Posts: …

Figure 3: Example prompt for GPT's classification of Reddit posts from r/relationship_advice and r/relationship.
Note: The prompt varied in whether category descriptions (blue) and examples (green) were included, as well as in the number of posts (pink) shown following the instructions. The requested answer format for the topic categories differed by approach: binary (yes/no), ordinal ("not at all," "slightly," "moderately," "strongly"), or metric (0–100 scale). For full details on all variations of category descriptions and examples, see Online Appendix Section C.1.

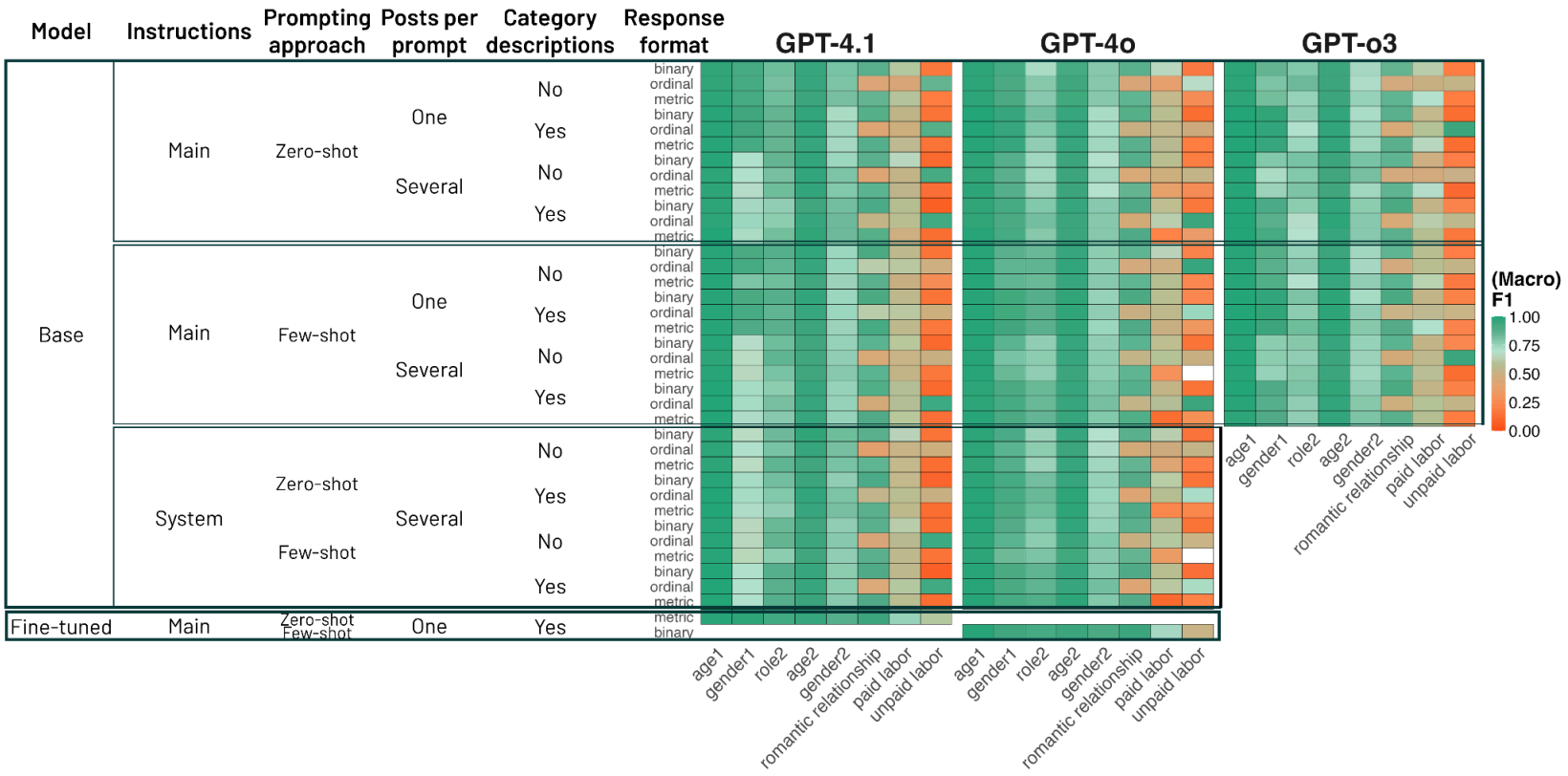

Figure 4. Heatmap of the (Macro) F1 showing the predictive performance of each prompting approach per model and category

Note: Each cell reports the F1 score (binary classification) and Macro F1 score (multiclass classification) for OpenAI models GPT-4.1, 4o, and o3. Darker green shades indicate stronger predictive performance (F1 > 0.7), whereas darker red shades indicate weaker performance (F1 < 0.7). Columns represent the GPT models within each category, and rows correspond to prompting strategies. We ran each approach using context window lengths of 30,000 and 120,000 tokens. As predictive performance was nearly identical across context window sizes, we report the slightly better result for each model (30,000 tokens for GPT-4.1 and o3, and 120,000 tokens for 4o). To test the classification's reliability, we ran each approach twice for each GPT model, with one context window length (30,000 tokens for GPT o3 and 4o and 120,000 tokens for GPT 4.1) and calculated intercoder reliability with Cohen's kappa. Where two runs were available, we report the mean scores across both runs.

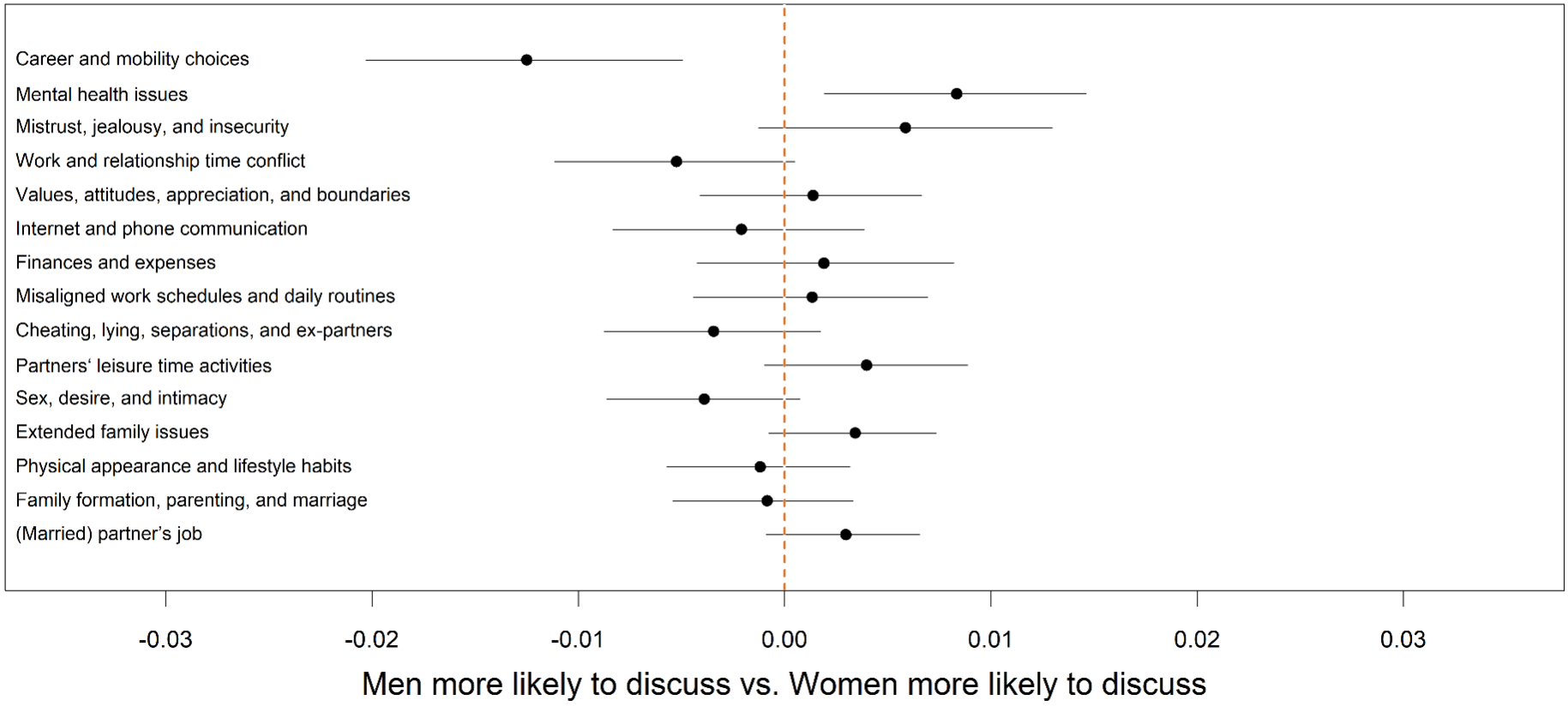

Figure 5: Structural Topic Model results showing the contrast of topical prevalence by gender for the top 15 topics in posts about paid work conflicts in romantic relationships

Note: Structural Topic Models with k=15 topics on the y-axis, ordered from most to least frequent and using age and gender of the author and partner, marital status to the partner, and time as covariates for topic prevalence. Confidence intervals are shown for the estimated differences in topic prevalence between men (negative values) and women (positive values) on the x-axis, indicating whether one gender discusses a given topic significantly more frequently than the other.

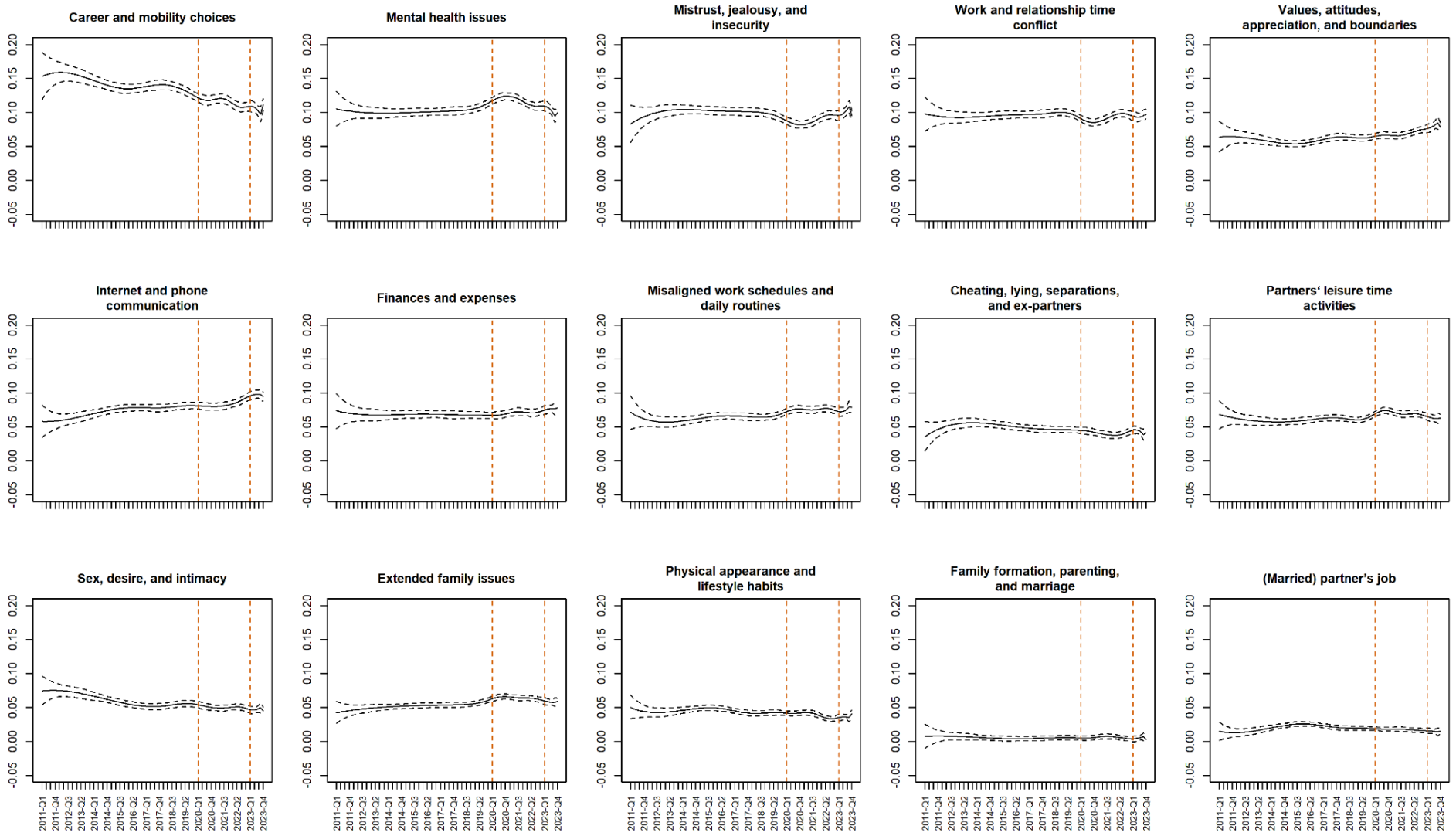

Figure 6. Expected topic proportion (y-axis) over time (x-axis) for topics occurring in conflicts about paid work in romantic relationships.
Note: Structural Topic Models with k=15 topics, using age and gender of the author and partner, marital status to the partner, and time as covariates for topic prevalence. The vertical lines indicate the time periods before, during, and after the COVID-19 pandemic.

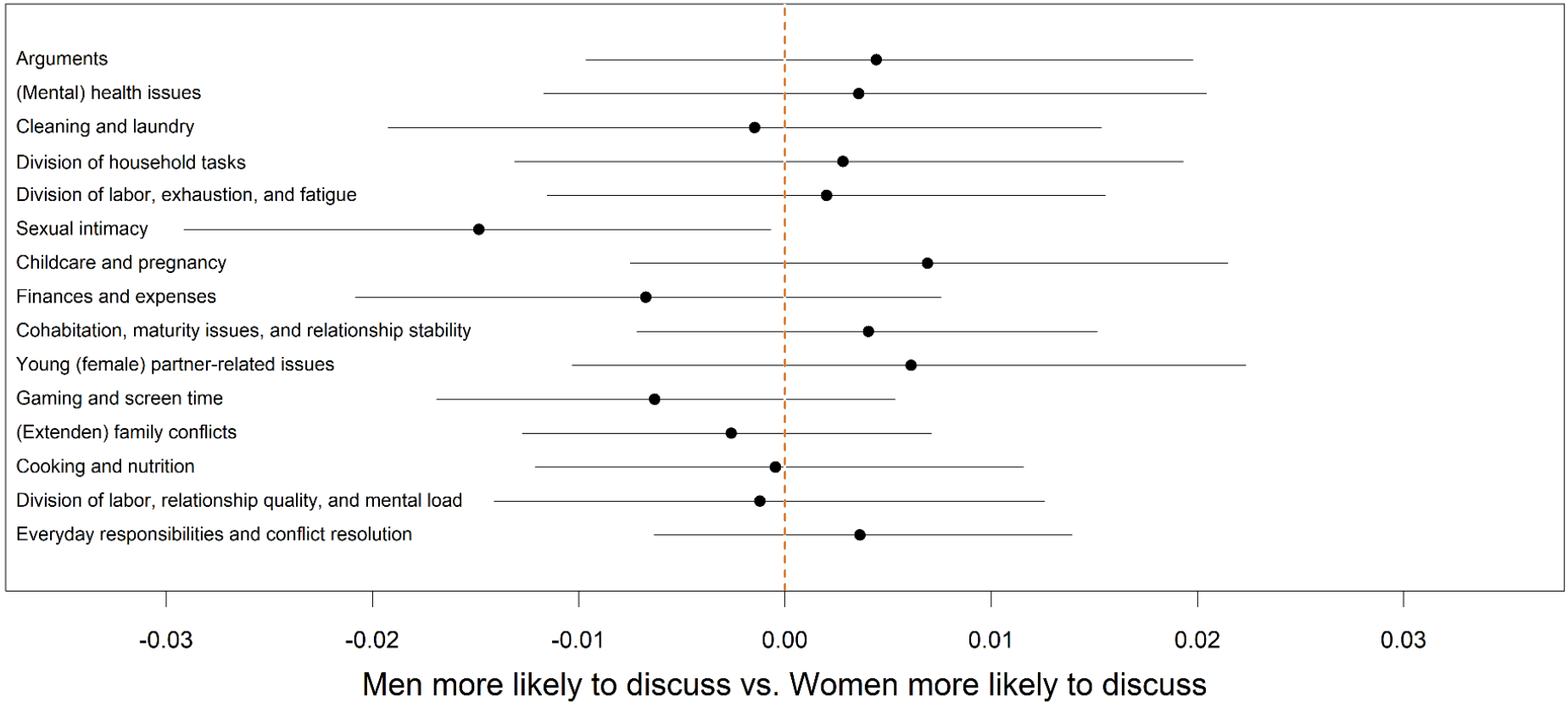

Figure 7. Structural Topic Model results showing the contrast of topical prevalence by gender for the top 15 topics in posts about unpaid work conflicts in romantic relationships

Note: Structural Topic Models with k=15 topics on the y-axis, ordered from most to least frequent, and using age and gender of the author and partner, marital status to the partner, and time as covariates for topic prevalence. Confidence intervals are shown for the estimated differences in topic prevalence between men (negative values) and women (positive values) on the x-axis, indicating whether one gender discusses a given topic significantly more frequently than the other.

Online Supplement for

“My (22m) Girlfriend (23f) Comes Home and Does Nothing” – Gendered Perceptions of Paid and Unpaid Work in Reddit Relationship Discussions over Time

Table of contents

A Data

A.1 Sample

We collected Reddit data from the two subreddits r/relationships and r/relationship_advice covering the period 2009 to 2024. We obtained the data using publicly available torrent files compiled by Pushshift and u/raiderbdev, hosted on academictorrents.com (Baumgartner et al., 2020; Watchful1, 2025). These files contain comprehensive archives of Reddit content. After removing empty entries, our dataset comprised a total of 2,351,608 posts: 791,051 from r/relationships and 1,560,557 from r/relationship_advice.

From this dataset, we randomly sampled 1,854 posts for manual annotation, preserving the proportional distribution of posts across both subreddits. The remaining 2,349,754 posts were set aside for automated classification. During preprocessing, we removed: (a) 13 invalid entries remaining in the dataset; (b) all posts by accounts later deleted (214,743 cases); and (c) duplicates identified based on identical full text (title + text), which accounted for 7,835 cases plus another 16 duplicates overlapping with the manually annotated data. After these steps, we retained a total of 2,124,705 posts. Next, we filtered out non-English content using Google's Compact Language Detector 2 package. This step removed another 420 posts from the dataset. The resulting dataset contained 2,124,285 posts.

For automated classification using GPT-4.1, we drew random samples while preserving the observed distribution between the two subreddits (relationship_advice: ~68.5%; relationships: ~31.5%). In total, 502,000 posts were classified automatically.

To construct our analysis dataset, we combined the manually annotated data (excluding two non-English cases that were retained during annotation to train GPT on handling missing values) with the GPT-classified data. This approach resulted in 503,852 coded posts.

We then applied final restrictions to refine our analysis data set. We excluded all entries where (a) title + text combined had fewer than ten characters (1 case) or (b) the author’s reported age fell outside 18–69 years (26,527 cases), in line with our focus on working-age individuals and labor dynamics in romantic relationships. We also limited the dataset to 2011–2023, excluding earlier years due to sparse posting activity (2009: 202 posts; 2010: 2,162 posts) and 2024 due to minimal data (9 posts). To address duplicate content resulting from post edits (e.g., when users updated their original posts), we checked for duplicates after removing any added "edit" sections and comparing the remaining text parts. Seven such duplicates were identified; in

each case we retained only the edited version. This refinement process resulted in a final dataset of 474,944 posts for analysis.

For our topic modeling, we applied additional thematic filters. We retained only posts that: (a) centered around romantic relationships (where at least 40% of the sentences discuss this theme), (b) featured a partner or ex-partner as the main protagonist, and (c) included discussion of either paid labor (≥15% of sentences) or household labor (≥15%). Out of our analysis data set, 294,715 posts met the criteria for romantic relationships involving a partner or ex-partner. Among these, 34,736 further addressed paid labor, while only 6,180 discussed household labor.

A.2 Data preparation

The automated classification of data contained some erroneous classifications, such as incorrect coding of gender or unreasonable numbers for the share of sentences related to household labor. These cases were rare, however, and made it possible to manually check and correct the classifications.

B Coding scheme

B.1 Overview

Table B.1-1: Variables to be extracted from r/relationship_advice and r/relationships

Variable	Content	Type	Range/Levels	Missings
Simple/manifest variables				
age1	age of author	metric	0-116	NA (not available); NA_l (text is not in English)
gender 1	gender of author	nominal	f (female); m (male); mtf (male to female); ftm (femalte to male); trans; non-binary; mf (m/f or f/m); other	NA (not available); NA_l (text is not in English)
role2	role of protagonist	nominal	Open answer OR partner/ex-partner (when talking about someone with whom one has/had a romantic relationship without naming that person)	NA_p (more than one protagonist), NA (not available), NA_l (text is not in English)
age2	age of protagonist	metric	0-116	NA (not available); NA_l (text is not in English); NA_p (more than one protagonist)
gender2	gender of protagonist	nominal	f (female); m (male); mtf (male to female); ftm (femalte to male); trans; non-binary; mf (m/f or f/m); other	NA (not available); NA_l (text is not in English); NA_p (more than one protagonist)
Complex/latent variables (topics)				
hh_labor	share of the post sentences that is concerned with household labor	metric	0-100 (rounded in steps of 5)	NA_l (text is not in English)
paid_labor	share of the post sentences that is concerned with paid labor	metric	0-100 (rounded in steps of 5)	NA_l (text is not in English)
rom_rel	share of the post sentences that is concerned with romantic relationship	metric	0-100 (rounded in steps of 5)	NA_l (text is not in English)

B.2 Topic definitions

Paid labor: We define paid labor (synonyms paid work, employment, work in the formal market economy, participation in work/labor force, paid work outside home, market work) as any work for pay. Working for pay can be done by having a job and earning a salary, or by earning as a self-employed with one's own business. We code paid labor if a post mentions everything related paid labor and labor market activities, including experiences, situations and interactions during and in the context of work. We code paid labor as a continuum indicating the share of the post sentences that is concerned with paid labor.

Household labor: We define household labor (synonyms family work, housework, domestic work) as "unpaid work done to maintain family members and/or home" (Shelton & John, 1996, p. 300). Various types of cognitive labor, falling under the category of non-physical or "invisible" types of work, such as anticipating basic household and/or family needs, household management, or decision-making fall under this definition (Daminger, 2019; Shelton & John, 1996). Examples for housework tasks include routine housework tasks such as cooking, cleaning, and shopping and intermittent or occasional housework tasks such as home repairs, gardening, and paying bills (Coltrane, 2000; Lachance-Grzela & Bouchard, 2010). We also include childcare tasks in our definition. We code unpaid labor as a continuum indicating the share of the post sentences that is concerned with household labor.

Romantic relationship: We define romantic relationships as "mutually acknowledged ongoing voluntary interactions", which "typically have a distinctive intensity, commonly marked by expressions of affection and current or anticipated sexual behavior" (Collins et al., 2009, p. 632). In our definition, romantic relationships must have three common features: passion, intimacy and commitment (Connolly & McIsaac, 2013, p. 181; Sternberg, 1986). Forms of romantic relationships in this definition include marital relationships, coresidential relationships such as cohabitation relationships, part-time cohabitation or stayover relationships (in between dating and cohabitation), living-together-apart relationships (ex-partners living together), and dating, at least if it has been going on for some time (e.g., a month) (Furman & Hand, 2015; Tillman et al., 2019). Sexual activity can be part of romantic relationships. We do not include sexual relationships that occur outside of the dating context or outside of committed relationships, even if they are permanent sexual relationships (Wentland & Reissing, 2011, p. 75). As romantic relationships have to be mutually acknowledged, we also do not code unilateral crushes or desires for a relationship with another person as romantic relationships. We code past and current romantic relationships as a

continuum. We code everything concerning a romantic relationship as such (e.g., everyday life problems), regardless of who is involved (e.g., not the author).

B.3 Intercoder reliability

Table B.1-2: Cronbach's alpha of 101 human-labelled posts

Variable	IRR	Variable	IRR
Author age	0.994	Household labor (metric)	0.932
Author gender	0.928	Paid labor (metric)	0.845
Protagonist age	0.982	Romantic relationship (metric)	0.877
Protagonist gender	0.907	All variables	0.917
Protagonist role	0.791		

C Methods

C.1 Classification approaches

Studying social inequalities using social media data, like Reddit, offers promising new opportunities but requires classification of large amounts of text data. Given the data volume (2.3M posts), manual coding is impossible, making automated or semi-automated coding necessary. Common approaches include dictionary-based methods, which are easy to implement but often lack precision. Although supervised classification methods (e.g., Naive Bayes Classifiers, Support Vector Machines, and tree-based methods) generally deliver results with high accuracy, these approaches are very costly in terms of monetary, time, and computational resources due to the extensive pre-processing required and the large amounts of manually annotated training data needed (Grimmer et al., 2022; Haensch et al., 2022; von der Heyde et al., 2025). By offering simple and more cost-effective classification options, LLMs promise to provide a flexible and efficient alternative (Chae & Davidson, 2025; Davidson, 2024; Törnberg, 2024). Within social media research, scholars have frequently applied LLMs to stance and sentiment detection (e.g., Chae & Davidson, 2025; Gilardi et al., 2023), to classify hateful content (e.g., Li et al., 2024), mental health disorders (e.g., Radwan et al., 2024), politicians' political affiliation (Törnberg, 2024), and hypocrisy accusations (Corral et al., 2024). The large majority of these studies work with LLMs from the GPT- followed by the LlaMa-family (Chae & Davidson, 2025; Corral et al., 2024; Gilardi et al., 2023; Li et al., 2024; Radwan et al., 2024; Törnberg, 2024). While most existing work has focused on Twitter/X (e.g., Gilardi et al., 2023; Törnberg, 2024) and Facebook (e.g., Chae & Davidson, 2025), several studies have also promisingly applied LLM-based text classification to Reddit data (Corral et al., 2024; Li et al., 2024; Radwan et al., 2024).

LLMs' ease of use lowers technical barriers for researchers with less technical background. Yet, the field of divergent models and approaches remains fragmented, including a wide range of models, prompting strategies, fine-tuning approaches, and context window lengths. While only very few studies systematically compare their effectiveness (e.g., Chae & Davidson, 2025), most existing applications have focused on a single model or strategy. Hence, it remains opaque which strategies yield satisfactory results and which will demonstrate the greatest performance in comparison. Lastly, while manifest constructs are easier for LLMs to classify, the classification of complex, latent constructs, which are often central to social science research, has proven particularly challenging with LLMs (e.g., Fan et al., 2024; Stuhler et al., 2025). We therefore tested several approaches for classification with GPT,

varying four key dimensions in our prompting strategy. First, we tested different prompting setups, including (a) instructions with one post per query, (b) instructions with multiple posts per query, and (c) instructions placed in the system message, while the prompt contained multiple posts. Second, we compared prompting approaches, using either (a) few-shot prompting, giving a few coded examples covering a wide range of variable values, or (b) zero-shot prompting, giving no examples. Third, we varied whether a category description was included or not, i.e., whether GPT received our definitions and coding rules, including the levels of the codebook, or not. Fourth, we tested different answer formats, instructing GPT to code the topic variables as either metric (in line with human annotation), ordinal, or binary. Combining these choices resulted in a total of 36 distinct approaches (see Table C.1-1).

Additionally, we varied GPT models (o3, 4o and 4.1), the number of tokens (i.e., units of text that GPT processes) used in one query (30,000 vs. 120,000) to evaluate the impact of the context window length, and fine-tuned vs. non-fine-tuned models (OpenAI, 2025). We fine-tuned GPT-4o and 4.1 on OpenAI's platform using a random 60-20-20 split of the manually annotated data as training, test, and validation sets, three and four epochs, and OpenAI's default batch size and learning rate. Fine-tuning promises improved performance on the specific annotation task, as the LLM's weights get updated as the model learns from the manually annotated data (Dodge et al., 2020; Radford et al., 2018). For GPT-4.1 and -4o, and all prompting approaches, we set a seed and used a temperature of 0 to lead to more deterministic annotations and reduce variability. For GPT-o3, the temperature is automatically set to 1 and cannot be changed by the researchers.

We based our evaluation of the LLM's annotation performance on a sample of 350 human-annotated posts, using a diverse set of accuracy metrics, as each individual metric has its own strengths and weaknesses (e.g., Hand et al., 2024). For nominal, ordinal, and metric variables, we calculated Accuracy, Macro F1, Cohen's Kappa, and Matthews Correlation Coefficient (MCC). For our binary variables, we additionally calculated Balanced Accuracy, Sensitivity, Specificity, Precision, Recall, and F1 Score. We calculated bootstrapped confidence intervals for each metric by drawing 10,000 random samples with replacement from the test data. To test the classification's reliability, we ran each approach twice for each GPT model with one context window length (30,000 tokens for GPT o3 and 4o and 120,000 tokens for GPT 4.1) and calculated intercoder reliability with Cohen's Kappa.

To decide on an approach for fine-tuning, we first defined criteria for "satisfactory" performance for the non-fine-tuned models. Since they performed differently, we distinguish

between the manifest, simple to classify variables and complex variables entailing latent constructs. For manifest variables, satisfactory performance requires that no more than one metric falls below 0.7 (including lower confidence intervals). For complex variables, predictive accuracy was overall lower, leading to a threshold of 0.6 to be considered for fine-tuning. Reliability values above 0.7 were classified as fully satisfactory.

We decided on the most promising prompting approach, GPT model, and context window length based on a three-step process. First, we selected the top five prompting approaches for each model and token number based on the arithmetic mean of the average of all metrics across all variables, as well as across complex variables only, due to their lower overall performance. For models tested twice, we used the respective means across classifications. Additionally, we included approaches that performed satisfactorily in at least one complex variable. For this selection, we second examined whether an approach met the satisfactory threshold for all variables. Finally, we compared the best-performing prompting approaches across token numbers within each model to identify the optimal approach per model. Overall, we prioritized metrics related to complex variable classification.

Table C.1-1: Classification approaches -- differentiation according to promting setup, approach, category description and answer options

ID	Prompting setup	Prompting Approach	Category description	Answer options for topics
main_few_catdes_bi	instructions + one post per query	few shot promting	yes	binary
main_sev_few_catdes_bi	instructions + multiple post per query	few shot promting	yes	binary
sev_few_catdes_bi	multiple post per query / instructions in the system	few shot promting	yes	binary
main_zero_catdes_bi	instructions + one post per query	zero shot promting	yes	binary
main_sev_zero_catdes_bi	instructions + multiple post per query	zero shot promting	yes	binary
sev_zero_catdes_bi	multiple post per query / instructions in the system	zero shot promting	yes	binary
main_few_bi	instructions + one post per query	few shot promting	no	binary
main_sev_few_bi	instructions + multiple post per query	few shot promting	no	binary
sev_few_bi	multiple post per query / instructions in the system	few shot promting	no	binary
main_zero_bi	instructions + one post per query	zero shot promting	no	binary
main_sev_zero_bi	instructions + multiple post per query	zero shot promting	no	binary
sev_zero_bi	multiple post per query / instructions in the system	zero shot promting	no	binary
main_few_catdes_ordinal	instructions + one post per query	few shot promting	yes	ordinal
main_sev_few_catdes_ordinal	instructions + multiple post per query	few shot promting	yes	ordinal
sev_few_catdes_ordinal	multiple post per query / instructions in the system	few shot promting	yes	ordinal
main_zero_catdes_ordinal	instructions + one post per query	zero shot promting	yes	ordinal
main_sev_zero_catdes_ordinal	instructions + multiple post per query	zero shot promting	yes	ordinal
sev_zero_catdes_ordinal	multiple post per query / instructions in the system	zero shot promting	yes	ordinal
main_few_ordinal	instructions + one post per query	few shot promting	no	ordinal
main_sev_few_ordinal	instructions + multiple post per query	few shot promting	no	ordinal
sev_few_ordinal	multiple post per query / instructions in the system	few shot promting	no	ordinal
main_zero_ordinal	instructions + one post per query	zero shot promting	no	ordinal
main_sev_zero_ordinal	instructions + multiple post per query	zero shot promting	no	ordinal
sev_zero_ordinal	multiple post per query / instructions in the system	zero shot promting	no	ordinal
main_few_catdes_prop	instructions + one post per query	few shot promting	yes	metric
main_sev_few_catdes_prop	instructions + multiple post per query	few shot promting	yes	metric
sev_few_catdes_prop	multiple post per query / instructions in the system	few shot promting	yes	metric

main_zero_catdes_prop	instructions + one post per query	zero shot promting	yes	metric
main_sev_zero_catdes_prop	instructions + multiple post per query	zero shot promting	yes	metric
sev_zero_catdes_prop	multiple post per query / instructions in the system	zero shot promting	yes	metric
main_few_prop	instructions + one post per query	few shot promting	no	metric
main_sev_few_bi	instructions + multiple post per query	few shot promting	no	metric
sev_few_prop	multiple post per query / instructions in the system	few shot promting	no	metric
main_zero_prop	instructions + one post per query	zero shot promting	no	metric
main_sev_zero_prop	instructions + multiple post per query	zero shot promting	no	metric
sev_zero_prop	multiple post per query / instructions in the system	zero shot promting	no	metric

C.2 Topic models

We used topic modeling algorithms to explore the topics discussed in all posts classified as covering romantic relationships and household or paid labor. These algorithms analyze unstructured document collections and discover themes without requiring prior annotations or labeling (Blei, 2012). Unlike clustering algorithms, topic models are mixed-membership models, meaning that each document can belong to multiple topics simultaneously (Grimmer et al., 2022).

One widely used topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is a Bayesian hierarchical model that assumes a specific process for generating text (Grimmer et al., 2022). The documents are characterized by distributions over latent topics, and the topics are defined as distributions over words (both of which are assumed to be drawn from Dirichlet distributions). The generative process involves two main steps: 1) Randomly draw a distribution over topics for each document, 2a) randomly choose a topic for the *n*th word in a document based on the distribution over topics in step one, and 2b) randomly choose an actual word from the corresponding distribution over the vocabulary, conditional on the topic assignment and the topic-specific distribution (Blei, 2012; Blei et al., 2003; Grimmer et al., 2022). The computational goal is to compute the posterior distribution of the hidden topic structure given the observed documents (Blei, 2012; Blei et al., 2003). When interpreting the topic model results, we follow Grimmer et al.'s (2022) by combining a close reading of exemplar texts that are strongly associated with a given topic, with quantitative methods that identify the words characterising each topic.

In many social science applications, researchers estimate LDA topic models for a corpus of documents and then compare how topic proportions vary with external covariates such as author’s characteristics. However, LDA assumes that documents are interchangeable, which contradicts findings on the influence of covariates on the text content (Roberts et al., 2016). Structural Topic Models (STM) address this issue by extending LDA to incorporate metadata into the model (Roberts et al., 2016). We applied STM using the R package `stm` (Roberts et al., 2014, 2019). With STM, additional covariates are included to improve topic estimation. Specifically, covariate information can be incorporated that affects the topic prevalence (i.e., a document's proportion devoted to a topic) and the topical content (i.e., the frequency of words used when discussing a topic). Thereby, for example, document-topic proportions can vary as a function of observed covariates rather than arising from a single prior shared by all

documents (Roberts et al., 2016). Accordingly, STM helps determine how topic prevalence and content differ across documents based on their metadata (Grimmer et al., 2022).

In our study, we included the following as covariates for topic prevalence: the author's gender and age; the protagonist's gender, age, and role; and the time, measured in quarters (three months intervals). We adopted a more flexible, nonlinear functional form for the time variable and estimated it with a spline (Roberts et al., 2019). We estimated the effects of the covariates using the estimateEffect function, which performs a regression with the topic proportions as the outcome variable (Roberts et al., 2014).

In STMs, the researcher must specify the number of topics *k*. In the first step, we selected the optimal number of topics *k* from the range 5-20 using a data-driven approach. Here, we evaluated the key metrics provided by the searchK() function in the STM R package (Roberts et al., 2019), including held-out log-likelihood, residual analysis, average exclusivity and semantic coherence. For example, exclusivity quantifies the degree to which the highest-probability words of a topic are unique to a given topic, while semantic coherence is maximized when the most probable words in a given topic frequently co-occur (Roberts et al., 2019). Second, based on these diagnostics, we selected three candidate values of K with the most favorable trade-off (K = 10, 12, and 15) and subjected them to manual evaluation. This qualitative assessment focused on topic interpretability (Grimmer & Stewart, 2013; Maier et al., 2021). For each topic, we examined the highest-probability words and the FREX metric. The FREX metric combines word frequency and exclusivity to a topic to penalize frequent yet non-exclusive words, providing a clearer picture of the topic (Grimmer et al., 2022; Roberts et al., 2016, 2019). In addition, we read the original posts with the highest proportions for each topic (Grimmer et al., 2022).

For data collection, preprocessing and the analysis, we made use of the following packages and software: We collected the Reddit data from pushshift using Python, especially the module `zstandard` (Szorc, 2025), while we conducted the data (pre-)processing, calls to the GPT API for the posts' classification, and Structural Topic Modeling in R, particularly using the `tidyverse` (Wickham et al., 2019), `httr`(Wickham, 2012), `quanteda`(Benoit et al., 2015), `stm` (Roberts et al., 2014, 2019), and `stminsights` (Schwemmer & Guyt, 2024) packages.

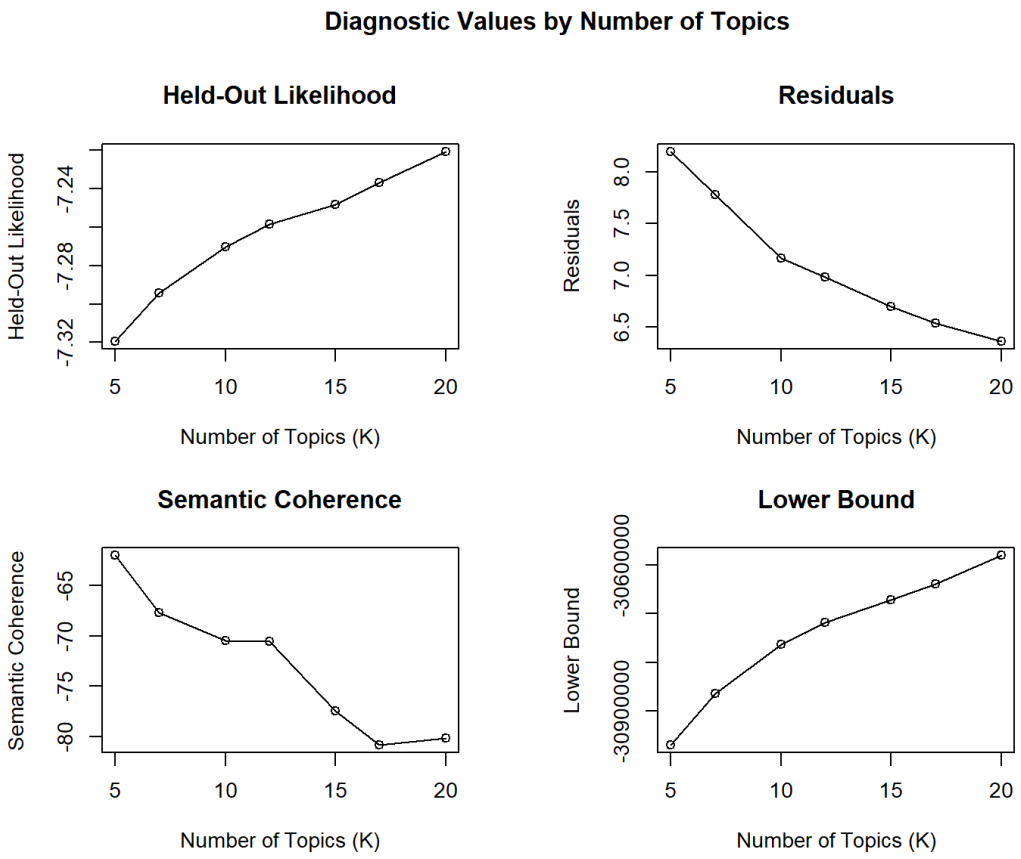

Figure C.1-1: k-diagnostic for STMs for paid labor

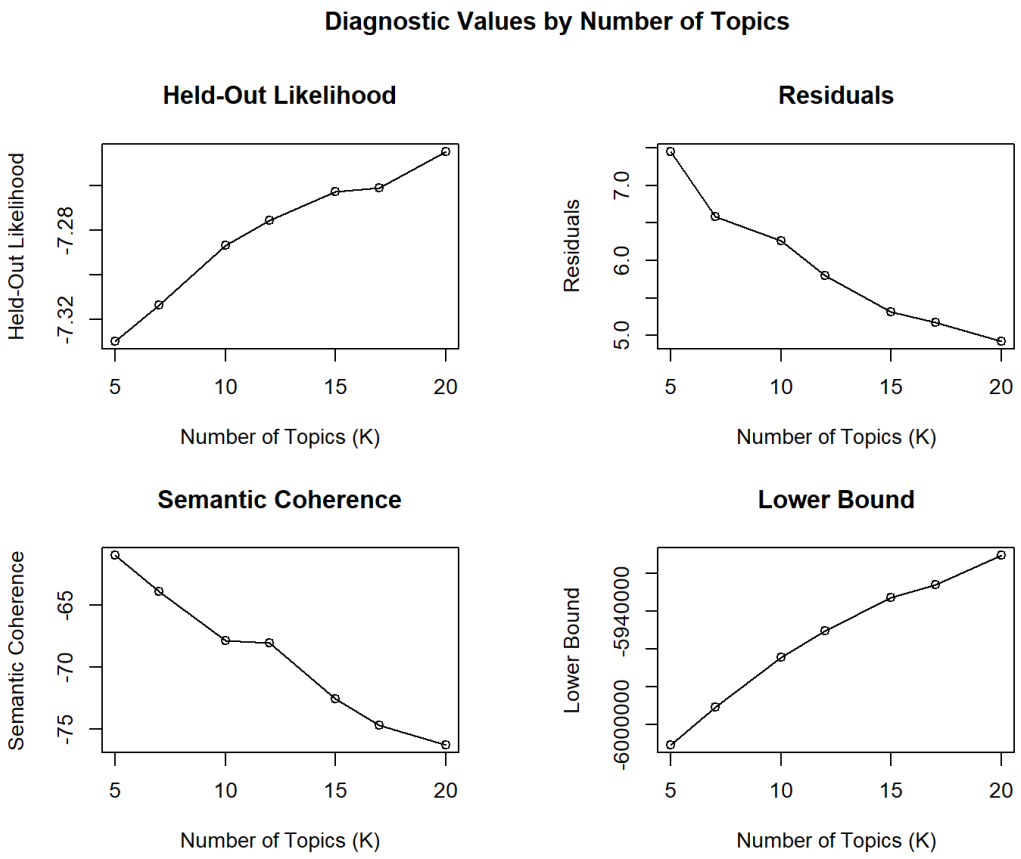

Figure C.1-2: k-diagnostic for STMs for household labor

Table C.1-1: Mean exclusivity by k for paid and household labor

	Mean exclusivity	
k	Paid labor	Household labor
5	9.184	8.937
7	9.422	9.141
10	9.557	9.434
12	9.629	9.490
15	9.715	9.565
17	9.742	9.611
20	9.782	9.651

Table C.1-2: Top words and FREX words for topics for different k (paid labor posts)

	K=10		K=12		K=15	
Topic	Top Words	FREX Words	Top Words	FREX Words	Top Words	FREX Words
1	depress, stress, better, support, bad, hard, emot, everyth, leav, issu	depress, anxieti, mental_health, mental, therapi, therapist, suicid, pain, suffer, medic	depress, stress, support, better, bad, break, hard, emot, everyth, leav	depress, anxieti, mental_health, mental, therapist, therapi, medic, suffer, pain, suicid	depress, stress, better, support, break, bad, hard, emot, everyth, problem	depress, anxieti, mental_health, mental, therapi, therapist, suffer, cope, pain, medic
2	text, call, phone, messag, went, later, night, happen, left, upset	text, messag, repli, phone, sent, delet, email, respond, send, call	text, call, phone, messag, went, later, night, saw, happen, look	messag, text, repli, delet, phone, respond, chat, instagram, sent, send	text, call, phone, messag, went, convers, later, send, repli, answer	text, repli, messag, respond, phone, send, chat, delet, sent, answer
3	guy, girl, ex, cheat, girlfriend, cowork, date, gf, trust, met	girl, ex, cowork, cheat, flirt, co-work, guy, drunk, jealous, trust	guy, girlfriend, girl, gf, ex, date, cowork, cheat, broke, met	girl, ex, gf, jealous, guy, flirt, cowork, crush, girlfriend, broke	guy, girl, girlfriend, cowork, date, gf, ex, met, peopl, hang	cowork, girl, flirt, jealous, femal, guy, crush, co-work, male, insecur
4	career, famili, citi, futur, school, find, stay, countri, current, plan	career, graduat, countri, visa, abroad, opportun, citi, program, dream, across_countri	career, citi, futur, school, famili, find, current, countri, colleg, stay	graduat, career, countri, visa, abroad, program, opportun, degre, citi, dream	career, futur, school, citi, find, current, colleg, countri, graduat, stay	graduat, career, abroad, visa, countri, degre, program, internship, opportun, futur
5	hous, sleep, clean, hour, bed, around, night, dog, cook, eat	dish, laundri, clean, wash, wake, cook, dog, kitchen, chore, nap	hous, sleep, clean, hour, bed, around, night, dog, cook, tire	dish, laundri, clean, wake, cook, dog, chore, wash, kitchen, game	sleep, hous, clean, bed, night, hour, dog, leav, around, morn	wake, dog, kitchen, wash, dish, breakfast, clean, nap, bed, sleep
6	partner, person, peopl, issu, seem, look, convers, interest, etc, kind	partner, boundari, tattoo, profession, gender, men, polit, art, creativ, ben	partner, peopl, person, issu, convers, seem, busi, understand, share, etc	partner, boundari, profession, team, project, valu, role, client, art, topic	partner, person, issu, peopl, convers, understand, share, seem, howev, situat	partner, valu, boundari, topic, profession, express, regard, intellig, view, dynam
7	money, pay, financi, save, car, hous, parent, bill, rent, buy	rent, payment, loan, debt, bill, pay, money, expens, credit_card, paycheck	money, pay, financi, save, car, bill, rent, hous, paid, buy	loan, debt, rent, payment, bill, pay, expens, money, paycheck, credit_card	money, pay, financi, save, bill, rent, car, paid, buy, expens	debt, payment, loan, bill, pay, paycheck, expens, credit_card, money, rent
8	husband, wife, kid, marri, famili, babi, mom, marriag, son, child	wife, husband, pregnant, daughter, divorc, son,	wife, kid, famili, mom, babi, son, mother, child, parent, daughter	wife, daughter, son, babi, daycar, kid, custodi, child, pregnanc, born	wife, kid, babi, marri, son, child, daughter, marriag, divorc, children	wife, kid, daughter, daycar, son, babi, child, custodi, children, born

		pregnanc, babi, custodi, child				
9	spend, girlfriend, hour, plan, weekend, date, busi, gf, away, trip	weekend, trip, schedul, cancel, spend, holiday, saturday, quality_tim, summer, clingi	spend, hour, plan, weekend, trip, away, busi, night, date, girlfriend	weekend, trip, cancel, saturday, friday, sunday, holiday, birthday, schedul, anniversari	plan, hour, weekend, spend, trip, away, girlfriend, night, date, visit	trip, weekend, birthday, saturday, friday, cancel, flight, sunday, holiday, summer
10	sex, use, sexual, smoke, watch, anymor, initi, look, weed, stop	porn, sex, smoke, weed, sex_driv, jack, drug, libido, addict, masturb	drink, fuck, car, smoke, look, use, shit, fianc, stop, went	smoke, weed, drug, alcohol, wear, addict, sober, hair, shirt, cop	fuck, look, drink, smoke, use, shit, wear, peopl, weed, drug	smoke, weed, wear, drug, jack, hair, shirt, tattoo, model, dress
11			sex, seem, anymor, interest, attract, chang, sexual, great, use, initi	sex, intimaci, weight, sex_driv, affection, intim, attract, sexual, spark, libido	sex, anymor, sexual, attract, seem, great, initi, interest, chang, physic	sex, intimaci, intim, weight, sexual, sex_driv, affection, attract, libido, spark
12			husband, marri, lie, found, marriag, cheat, leav, went, trust, came	husband, affair, john, email, deploy, lie, bob, l, wed, counsel	lie, found, cheat, ex, went, leav, left, trust, came, happen	lie, affair, forgiv, john, polic, truth, betray, proof, caught, discov
13					husband, marri, compani, new, wed, busi, boss, marriag, manag, fianc	husband, tom, gt, spous, wed, husband', jame, hubbi, h, hire
14					spend, play, hour, game, watch, sometim, around, full_tim, chore, hobbi	game, hobbi, play, video_gam, chore, free_tim, play_video_gam, playing_video_gam, lazi, play_gam
15					famili, parent, mom, dad, hous, gf, mother, sister, brother, stay	brother, sister, dad, mom, parent, sibl, mum, mother, cousin, uncl

Table C.1-2: Top words and FREX words for topics for different k (household labor posts)

	K=10		K=12		K=15	
Topic	Top Words	FREX Words	Top Words	FREX Words	Top Words	FREX Words
1	upset, call, wrong, argument, went, fight, angri, right, happen, today	door, yell, apolog, grab, fuck, argument, respond, calm, scream, angri	upset, wrong, fight, argument, angri, right, call, yell, today, happen	apolog, yell, door, anger, calm, argument, angri, fuck, scream, grab	went, call, upset, came, angri, wrong, fight, left, today, argument	yell, apolog, scream, text, drunk, door, angri, fuck, piss, anger
2	eat, sleep, dinner, bed, night, food, wake, meal, morn, usual	chicken, eat, pm, gym, hungri, wake, breakfast, asleep, lunch, meal	eat, dinner, sleep, bed, night, food, wake, morn, play, usual	eat, chicken, pm, breakfast, wake, dinner, hungri, gym, asleep, lunch	eat, food, dinner, meal, made, lunch, buy, weight, night, prepar	chicken, eat, diet, meat, ingredi, soup, recip, pizza, pasta, meal
3	pay, money, bill, financi, partner, support, school, respons, save, contribut	contribut, incom, career, financi, expens, financ, cost, degre, bill, earn	chore, partner, respons, expect, housework, etc, household, around_hous, task, fair	task, equal, chore, household_chor, housework, household, respons, list, full-tim, free_tim	chore, partner, respons, expect, household, housework, etc, task, fair, around_hous	task, equal, household_chor, chore, household, full-tim, housework, fair, respons, resent
4	husband, wife, marri, kid, marriag, divorc, spend, famili, seem, etc	wife, husband, marriag, divorc, marri, wife', golf, toddler, husband', sahm	husband, wife, marri, kid, marriag, divorc, famili, school, spend, seem	wife, husband, spous, marriag, divorc, marri, counsel, husband', wife', sahm	wife, kid, marri, marriag, divorc, seem, famili, school, give, issu	wife, divorc, marriag, spous, wife', marri, in-law, kid, sahm, typic
5	dish, laundri, chore, wash, cloth, mess, trash, floor, dirti, kitchen	dirti, vacuum, pile, sink, messi, dishwash, dish, bathroom, litter, trash	dish, laundri, wash, cloth, mess, trash, floor, kitchen, dirti, bathroom	dirti, pile, sink, bathroom, trash, mop, toilet, counter, floor, put_away	sleep, tire, dog, school, night, laundri, wake, bed, lazi, full_tim	lazi, dog, wake, shift, tire, studi, exhaust, sleep, class, full_tim
6	boyfriend, bf, dog, play, game, friend, spend, littl, date, break	boyfriend, bf, roommat, living_togeth, dog, game, video_gam, playing_video_gam, boyfriend', puppi	boyfriend, bf, dog, play, apart, break, littl, game, date, place	boyfriend, bf, roommat, living_togeth, dog, leas, video_gam, boyfriend', mum, puppi	boyfriend, bf, apart, place, break, friend, date, littl, guy, living_togeth	boyfriend, bf, roommat, living_togeth, mum, boyfriend', leas, j, apart, guy
7	kid, babi, son, child, daughter, children, pregnant, famili, sleep, parent	babi, pregnanc, pregnant, son, born, child, birth, daycar, month_old, year_old	kid, babi, son, child, daughter, children, pregnant, sleep, famili, year_old	babi, pregnant, pregnanc, son, born, diaper, year_old, child, month_old, daycar	husband, kid, babi, son, child, children, pregnant, marri, two, stay_hom	babi, pregnant, son, pregnanc, husband, daycar, born, children, diaper, month_old

8	sex, depress, issu, life, chang, seem, better, person, partner, anymor	sex, depress, attract, affect, sexual, sex_lif, intimaci, interest, kiss, physic	sex, anymor, chang, happi, seem, better, life, date, person, give	sex, sexual, attract, sex_lif, kiss, romant, intimaci, interest, affection, hug	sex, anymor, chang, friend, happi, date, seem, interest, better, cheat	sex, sex_lif, kiss, attract, sexual, intimaci, cheat, affection, connect, cuddl
9	famili, parent, mom, friend, girlfriend, gf, stay, mother, money, place	gf, brother, sister, girlfriend, ex, aunt, mom, sibl, n, girl	parent, mom, girlfriend, famili, mother, gf, stay, dad, sister, ex	gf, brother, sister, girlfriend, fiancé, mom, parent, sibl, ex, aunt	mom, parent, famili, daughter, mother, dad, stay, sister, brother, father	sister, brother, daughter, mom, mother, sibl, dad, parent, father, visit
10	amp, partner, birthday, gift, adhd, gt, new, issu, life, famili	gt, amp, gift, adhd, nbsp, symptom, birthday, lt, autism, childhood	pay, money, bill, rent, save, buy, amp, financi, car, paid	gt, payment, debt, mortgag, money, pay, bill, credit, paid, expens	amp, birthday, gift, gt, new, christma, card, anniversari, holiday, buy	gt, amp, gift, birthday, anniversari, lo, d, christma, lt, holiday
11			support, depress, life, issu, partner, struggl, anxieti, stress, hard, due	medic, mental_health, anxieti, diagnos, depress, disabl, struggl, health, ill, med	depress, support, life, partner, issu, struggl, anxieti, emot, stress, hard	medic, mental_health, depress, anxieti, struggl, disabl, diagnos, med, adhd, support
12			friend, went, drink, found, made, call, text, new, left, night	drink, drunk, birthday, anniversari, messag, text, christma, alcohol, found, cowork	pay, money, bill, rent, save, financi, buy, car, paid, groceri	debt, pay, bill, money, paid, cost, rent, save, mortgag, expens
13					girlfriend, gf, friend, apart, break, person, ex, place, issu, problem	girlfriend, gf, ex, nbsp, she'll, joe, she'd, girl, breakup, apart
14					dish, wash, cloth, laundri, mess, floor, dirti, trash, kitchen, bathroom	sink, dirti, wash, bathroom, floor, pile, put_away, messi, cloth, toilet
15					play, watch, game, spend, friend, phone, tv, sit, video_gam, comput	game, play, video, tv, comput, video_gam, watch, movi, playing_video_gam, xbox

D Results

D.1 Classification statistics

In the following, we outline the LLMs' classification performance in detail. We interpret mean classification results for the cases in which models classified each post twice (i.e., GPT-4.1 with a 120,000-token context window and GPT-4o and o3 with a 30,000-token context window). We first describe the results for the non-fine-tuned models in detail, as these inform our selection of prompting approaches for subsequent fine-tuning.

Across all models, manifest categories are classified more accurately than latent topic categories. Among the manifest variables, GPT-o3 performs best, followed by GPT-4.1, while GPT-4o exhibits the lowest predictive performance. GPT-4.1 and o3 achieve particularly strong performance for the manifest categories of the author's and protagonist's age (averages of all metrics across prompting approaches ≥ 0.95 and 0.91, respectively) and meet acceptable performance thresholds for the second protagonist's gender (averages ≥ 0.81, with at most one metric including a lower confidence interval below 0.70). For GPT-o3, all approaches also satisfy the criteria for the author's gender (no metric with lower confidence intervals < 0.70). In contrast, most GPT-4.1 prompting approaches fall below acceptable thresholds for this category (averages ≥ 0.65), although all approaches involving exactly one post per prompt perform satisfactorily. The second protagonist's role is the most challenging manifest variable (averages ≥ 0.58 for GPT-4.1 and ≥ 0.69 for o3). Yet several approaches, primarily few-shot prompts with category descriptions, still meet our defined criteria. Except for a small number of zero-shot prompting approaches, GPT-4o also attains satisfactory performance for author and protagonist age and gender (averages ≥ 0.90, 0.66, 0.85, and 0.71, respectively). As with the other models, the protagonist's role performs least well (averages ≥ 0.66, with several approaches showing more than one metric below 0.70), although few-shot prompts with category descriptions again tend to meet the criteria. Overall, few-shot prompts with category descriptions and a single post per prompt yield strong classification performance across all GPT models, context window lengths, and manifest categories. Approaches using the shorter 30,000-token context window also perform marginally better than those using 120,000 tokens.

For the latent topic categories paid labor, household labor, and romantic relationship, GPT-4.1 outperforms both GPT-4o and o3, and GPT-4o consistently performs worse than GPT-o3. Overall, the latent topic categories show substantially weaker performance than the manifest variables. No combination of prompting strategy and GPT model achieves satisfactory performance across all topic categories. Predictive performance is lowest for household labor

(averages ≥ 0.39 for GPT-4.1, 0.30 for GPT-4o, and ≥ 0.42 for o3), followed by paid labor (averages ≥ 0.43, 0.42, and 0.48) and romantic relationships (averages ≥ 0.43, 0.40, and 0.45). Across all models, prompts requesting ordinal classifications perform substantially worse than those requesting binary or metric responses.

Although no prompting strategy achieves satisfactory performance for household or paid labor, each model attains satisfactory classification on the romantic relationship category with at least one approach. For GPT-4.1, two approaches meet the criteria with a binary or metric response format and a 120,000-token context window. Both place instructions in the main rather than system prompt and include category descriptions. One uses zero-shot prompting to classify multiple posts (average 0.82), while the other uses few-shot prompting to classify a single post (average 0.80). With a 30,000-token context window, two metric-format prompting approaches also meet the criteria, whereas none of the ordinal or binary ones do. Across these approaches, few-shot prompting with category descriptions in the main prompt, regardless of whether one or several posts are classified per prompt, performs best. For GPT-4.1, few-shot prompts with category descriptions, a single post per prompt, and a binary or metric response format perform well also across manifest categories. The metric format slightly outperforms the binary one for the most challenging topic category, household labor. We therefore select the few-shot prompt with category descriptions, a single post per prompt, and a metric response format as our fine-tuning approach for GPT-4.1. Because this approach always classifies a single post per prompt, the context window is never fully utilized. Resulting performance differences between context window lengths are thus likely attributable to random variance rather than systematic effects. This interpretation is supported by extremely high interrater reliability across the two runs (Cohen's Kappa ranging from 0.93 for romantic relationships to 1.0 for author and protagonist age). Given the slightly better performance on the topic categories, we select the 120,000-token context window for fine-tuning GPT-4.1.

For GPT-o3, only one approach meets the performance criteria for romantic relationships, namely a zero-shot prompt with category descriptions in the main prompt requesting a binary classification of multiple posts in a 120,000-token context window (average 0.79). As fine-tuning is not available for o3, we retain this approach for later comparison with the fine-tuned GPT-4.1 and GPT-4o models.

For GPT-4o, two approaches perform satisfactorily on the romantic relationship category with a 120,000-token context window, and one approach with a 30,000-token window. With the

longer window, both satisfactory approaches involve zero-shot prompts with category descriptions requesting metric classifications of multiple posts, with instructions placed either in the main or the system prompt (average of 0.83 and 0.81, respectively). With the shorter context window, the only satisfactory approach uses few-shot prompting with category descriptions in the main prompt for binary classification of a single post (average 0.80). Based on stronger performance across the manifest categories, we select this latter approach for fine-tuning GPT-4o. Interrater reliability is again extremely high (Cohen's Kappa 0.93 ranging from paid labor to 1.0 for age categories), further supporting this decision.

Fine-tuning substantially improves classification performance for both GPT-4.1 and GPT-4o, regardless of the number of training epochs. For GPT-4.1, the largest improvement arises with three epochs, whereas GPT-4o performs best on several categories when fine-tuned with four epochs. Because increasing the number of epochs leads to declining performance for GPT-4.1 and to slight decreases in accuracy on some manifest categories for GPT-4o, we refrain from increasing the number of epochs further to avoid overfitting. We validate the fine-tuned models by combining the validation and test sets and having each model classify each post twice, achieving high interrater reliability. We interpret the mean performance metrics below.

For GPT-4.1 fine-tuned on three epochs, we observe strong performance across all manifest categories, including the protagonist's role. Accuracy and related metrics reach or exceed 0.80, with at most a single lower confidence interval falling above 0.70. Topic classifications also improve substantially. In particular, the romantic relationship category shows robust performance, with all metrics including lower confidence intervals meeting or exceeding 0.80. Nonetheless, limitations remain for the topics of household and paid labor. While classification for paid labor performs notably better with all lower confidence intervals > 0.60, predictive accuracy for household labor remains insufficient, with a recall of 0.50 and a lower recall confidence interval as low as 0.25. Increasing the number of epochs to four yields largely comparable predictive performance but introduces a slight decline for the protagonist's role (e.g., MCC dropping below 0.80). Given these patterns, we select the GPT-4.1 model fine-tuned for three epochs for subsequent comparisons of approaches.

For GPT-4o fine-tuned on three and four epochs, the overall pattern of predictive performance mirrors that of GPT-4.1. The strongest results again appear for the manifest categories. However, the protagonist's role shows reduced stability, with three metrics falling below 0.80 for both the three-epoch and four-epoch models. As with GPT-4.1, romantic relationship predictions perform consistently well across epochs, with no metrics below 0.80. Paid labor

classifications also surpass a minimum lower confidence interval of 0.60. Yet, performance for household labor remains insufficient across both fine-tuned variants, reflected in a lower recall confidence interval of 0.18 for three epochs and 0.23 for four epochs.

Comparing the number of epochs across GPT-4o models, we observe slight improvements from three to four epochs for the topic categories and the protagonist's role, accompanied by marginal reductions in accuracy for other manifest categories. This marginal decline might indicate emerging overfitting.

Taken together, our results identify GPT-4.1 fine-tuned for three epochs as the most effective overall configuration, followed by GPT-4o fine-tuned for four epochs. All fine-tuned models outperform the non-fine-tuned counterparts across GPT-4.1, GPT-4o, and o3. Nonetheless, even our strongest model configuration fails to achieve satisfactory performance on the latent constructs of paid and especially household labor. Notably, the latent constructs most challenging for the LLMs to classify are also those for which human annotators exhibit the lowest reliability. In classifying the remaining posts, we nevertheless instructed the fine-tuned GPT-4.1 to annotate all categories included in the initial prompt, as omitting categories could alter predictive accuracy.

D.2 Descriptive statistics

D.2.1 Graphical overview

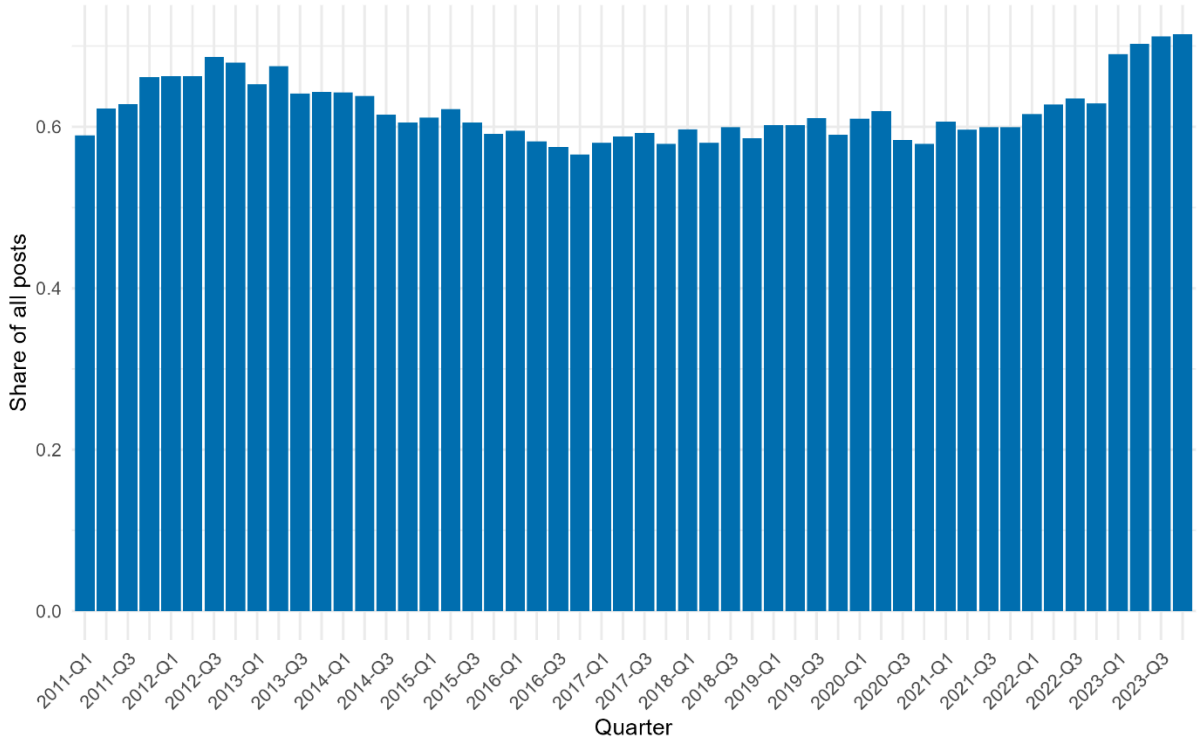

Figure D.2-1: Share of relationship posts of all posts over time

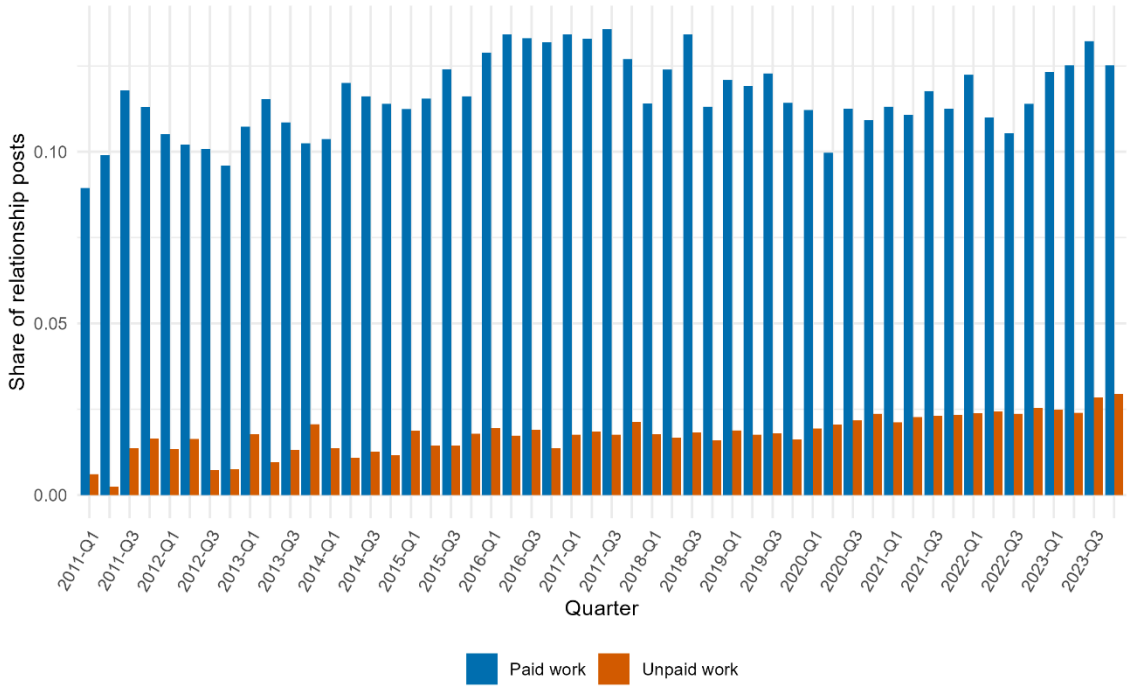

Figure D.2-2: Share of paid and unpaid work discussions within relationship posts over time

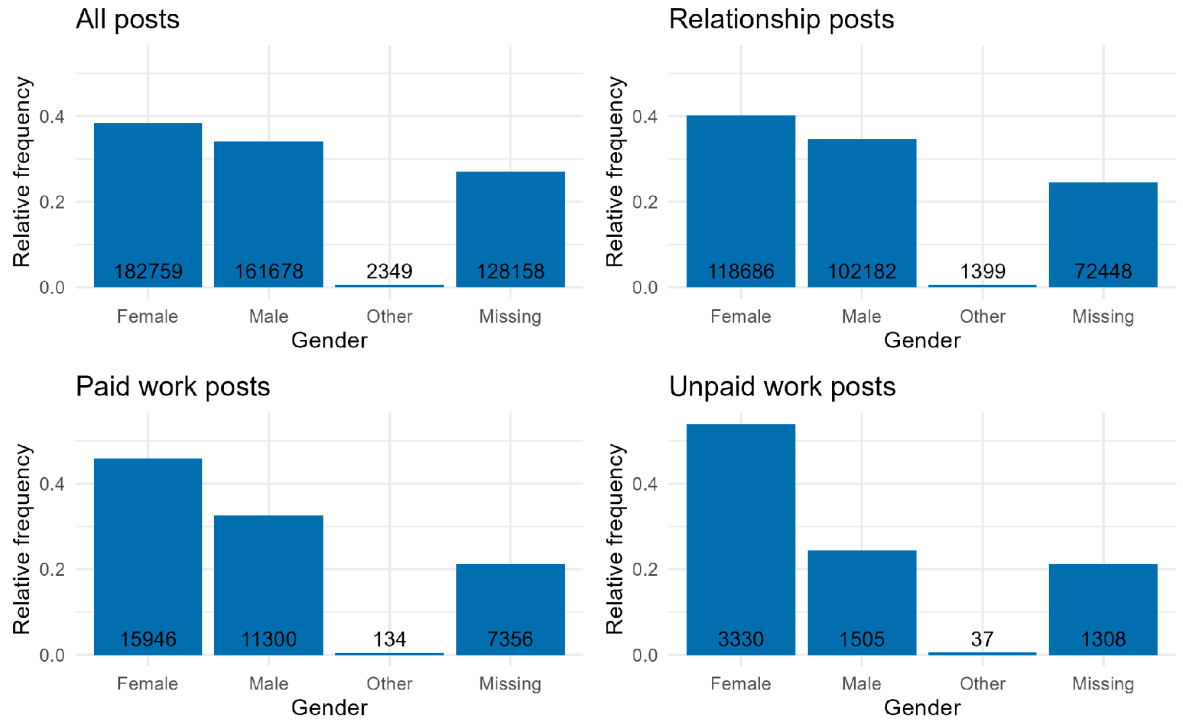

Figure D.2-3: Gender distribution of authors

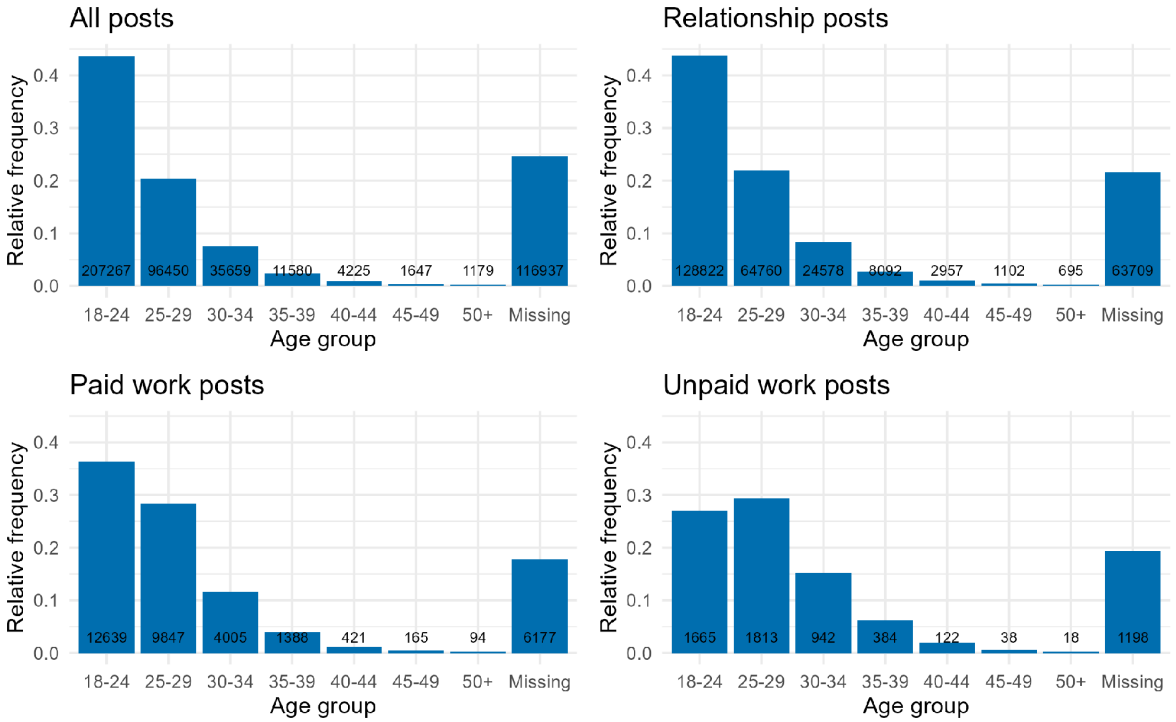

Figure D.2-4: Age distribution of authors

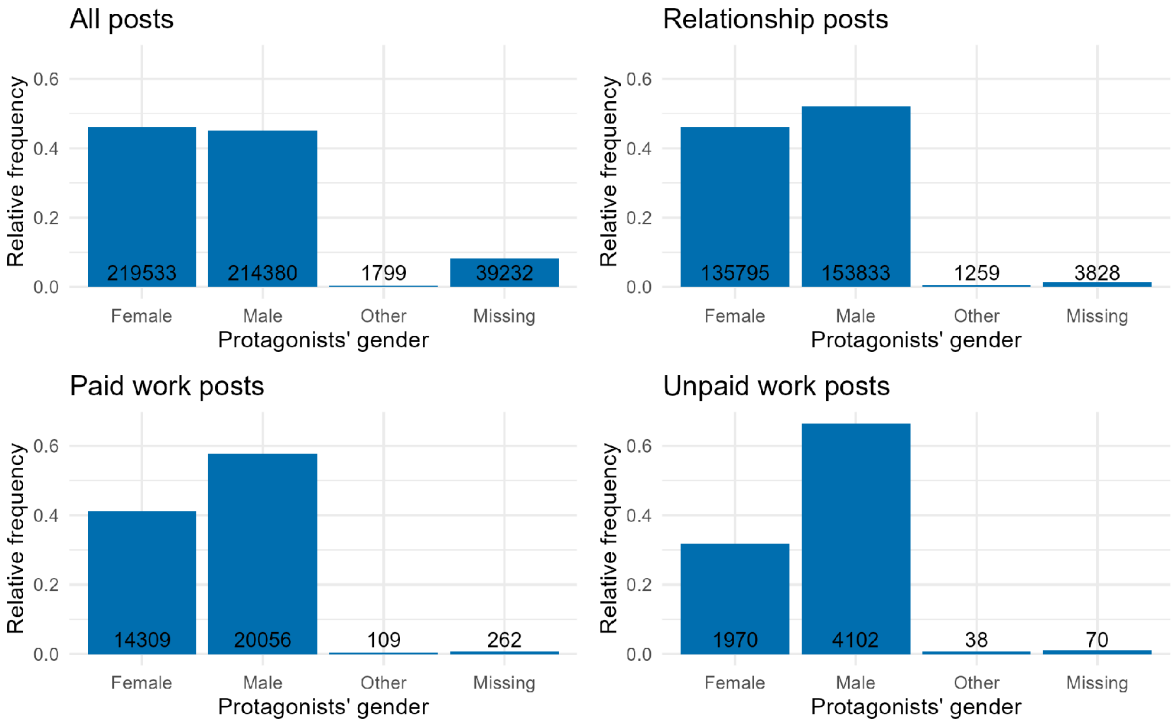

Figure D.2-5: Gender distribution of protagonists

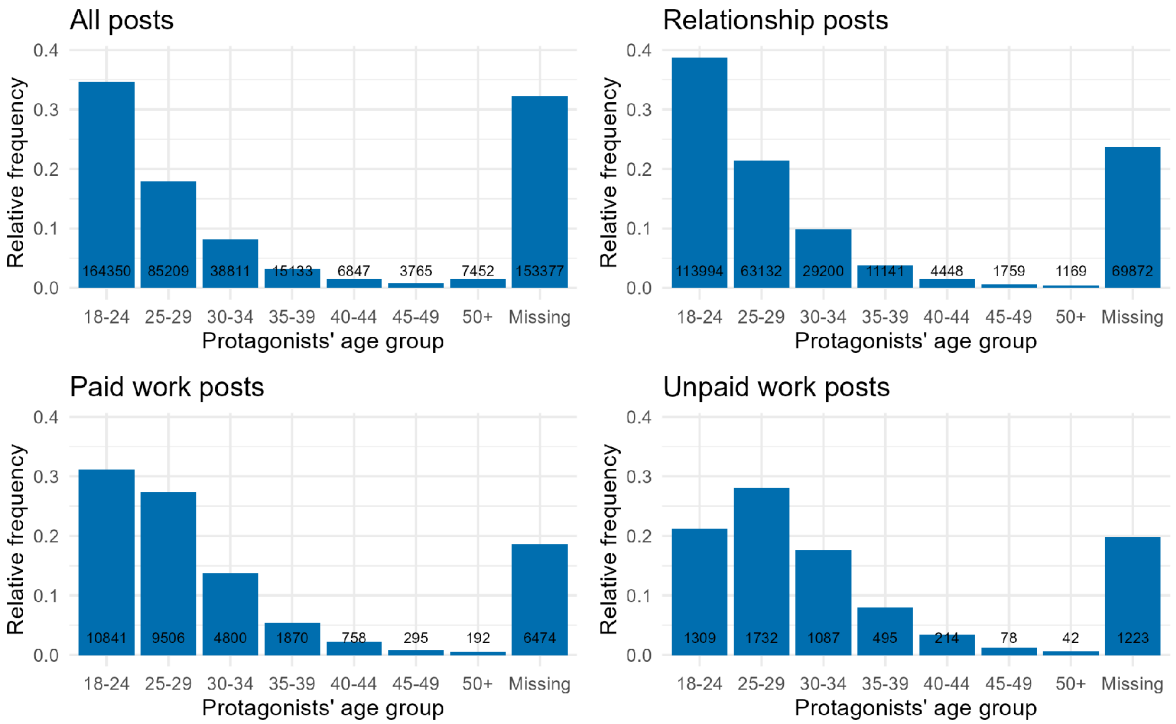

Figure D.2-6: Age distribution of protagonists

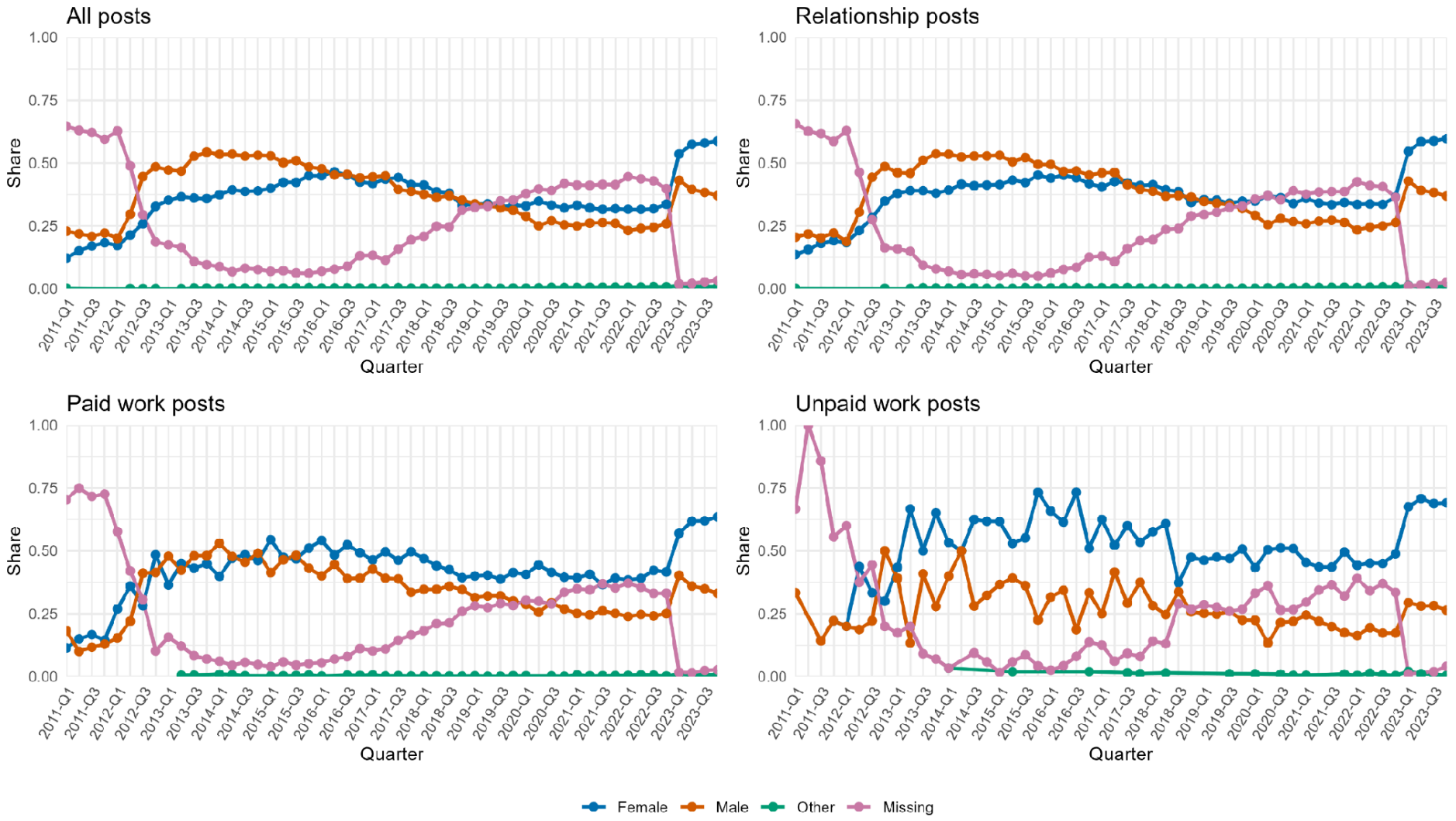

Figure D.2-7: Gender distribution of authors over time

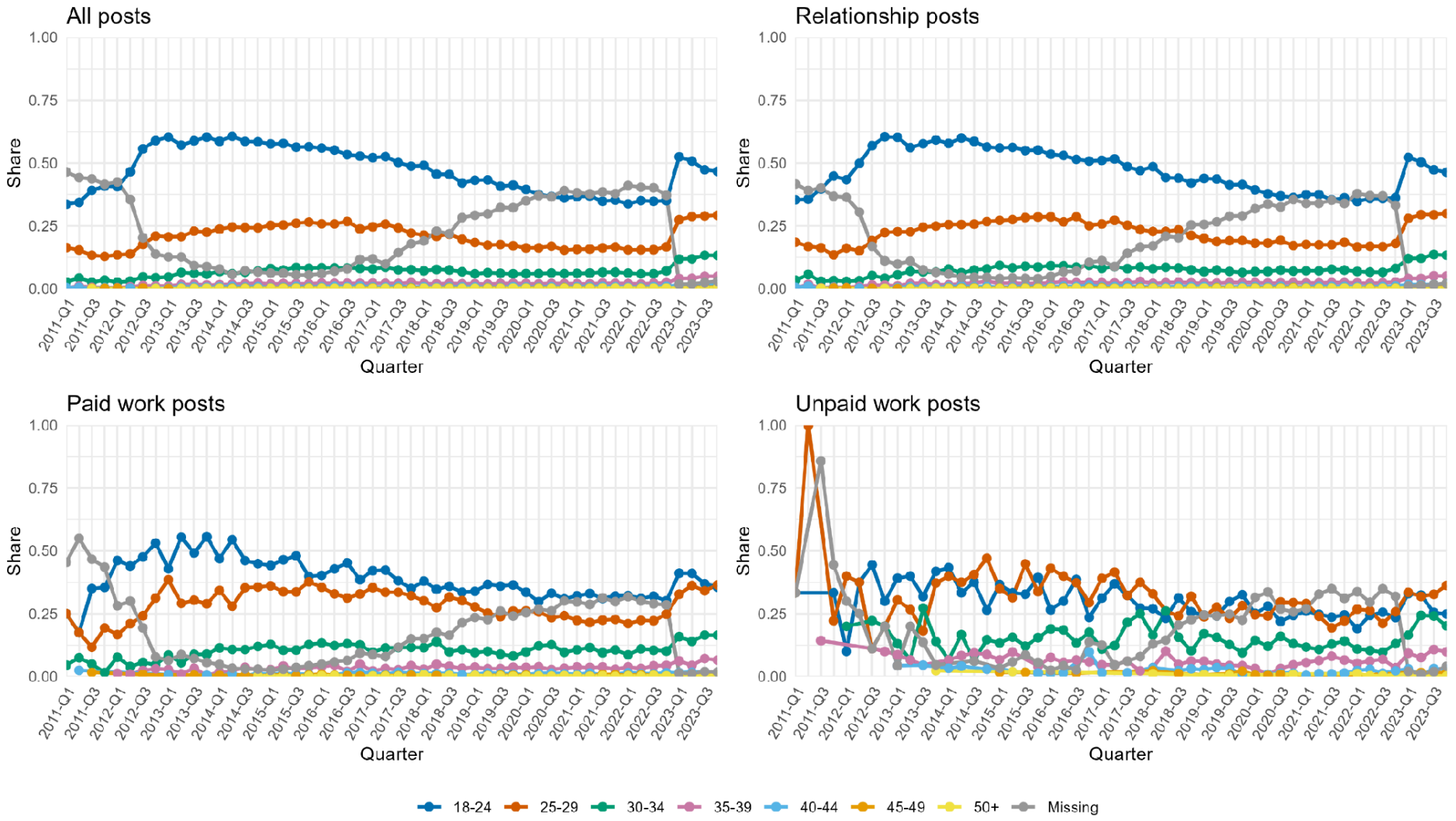

Figure D.2-8: Age distribution of authors over time

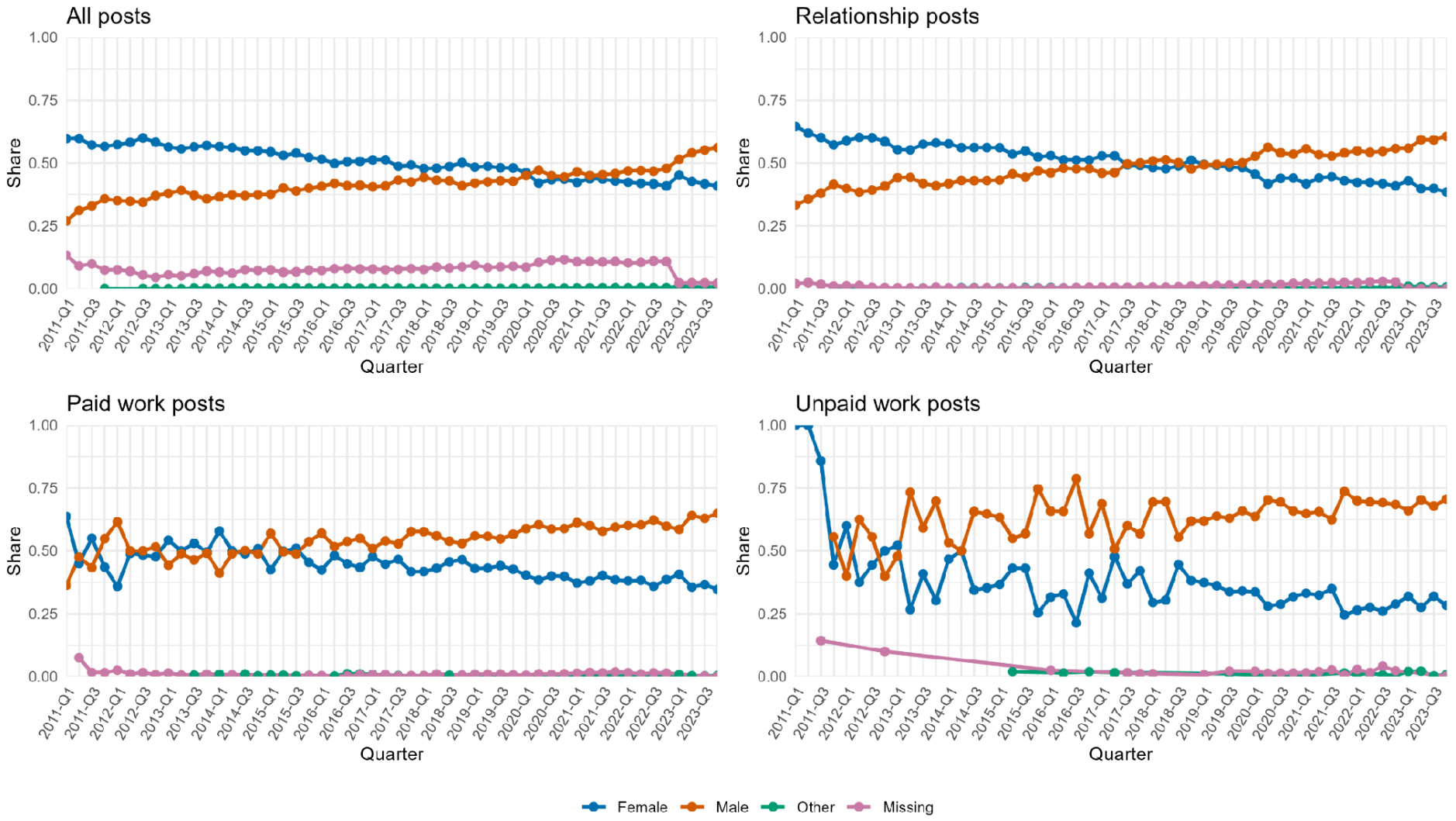

Figure D.2-9: Gender distribution of protagonists over time

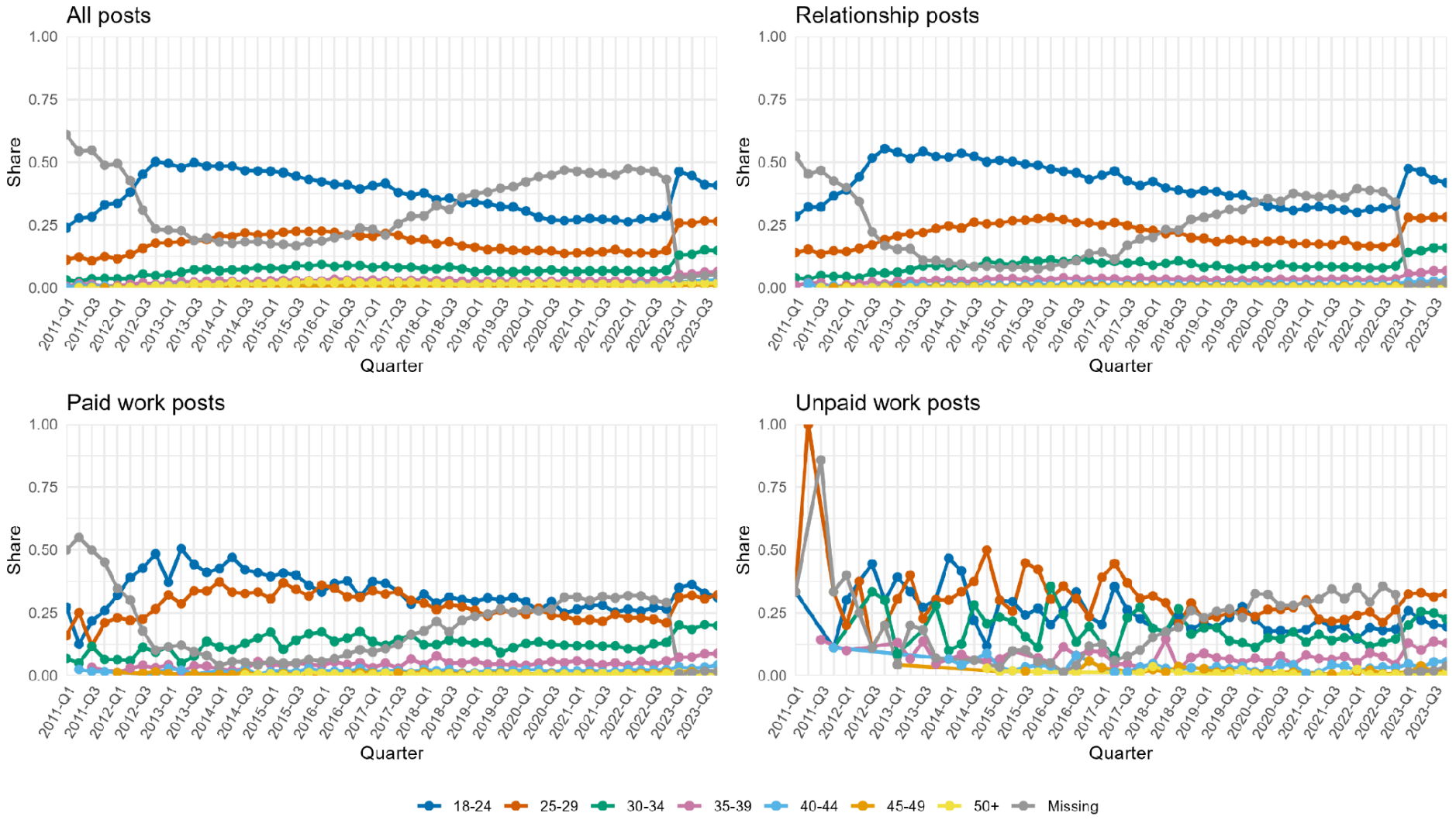

Figure D.2-10: Age distribution of protagonists over time

D.2.2 Changes in the demographic structure over time

Gender patterns have shifted. Since 2016, women have been more represented than men as authors in all posts and relationship posts, and in paid work posts since 2015. They have consistently dominated men in unpaid work household labor posts since 2013. Since the pandemic, men have become more frequent protagonists in all posts, and even more so in relationship posts. In paid and unpaid work posts, they have predominated since 2015 and 2013, respectively, but we also see a widening gender gap since the pandemic. The amount of missing demographic information was high from 2011 to 2013, then declined, and then rose again until early 2023, when it dropped sharply. This finding probably reflects changes in community rules and moderation. Age distributions show a gradual decline in the youngest groups and a slight rise in the 30-34 age group, with more fluctuations in paid and unpaid work posts.

D.3 Relationship subreddits on Reddit

D.3.1 Paid work

Topic correlations

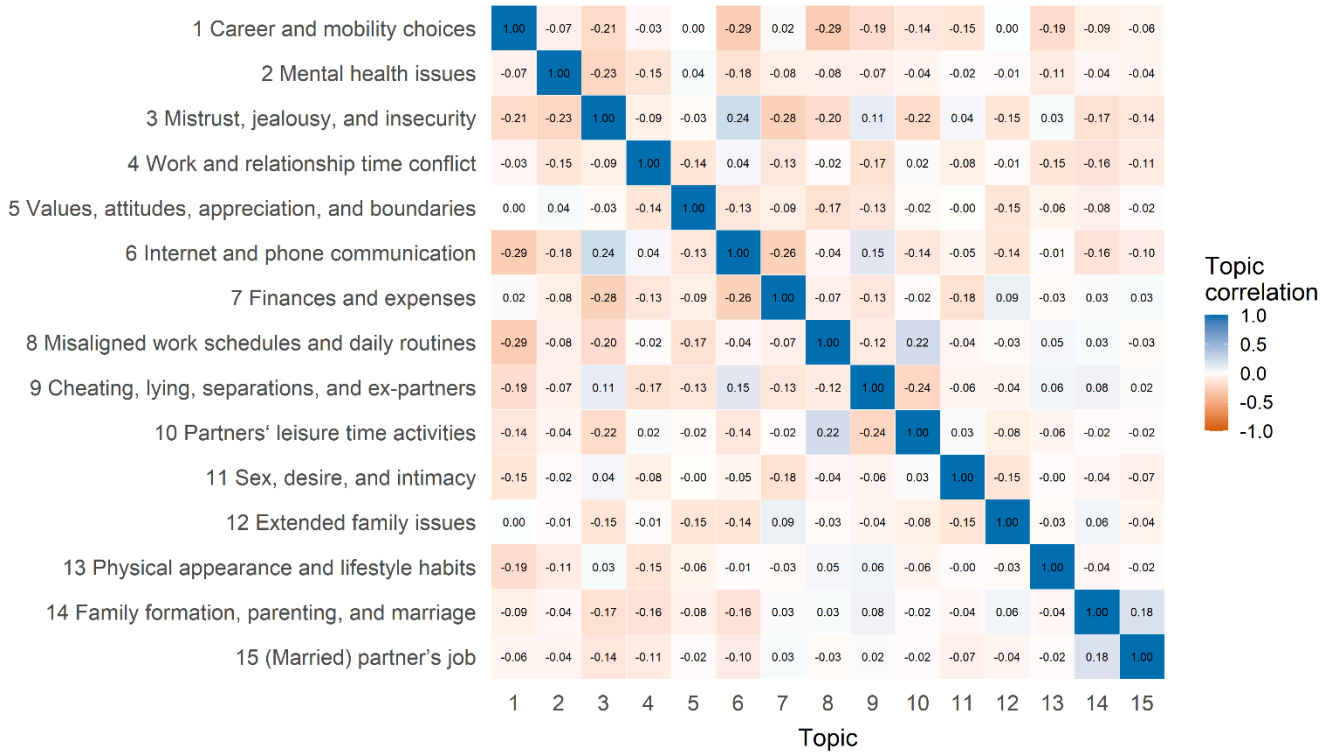

Figure D.3-1: Topic correlations for topics occurring in conflicts about paid work in romantic relationships.

Note: Structural Topic Models with k=15 topics, ordered from most to least frequent and using age, gender, of the author and partner, marital status to the partner, and time as covariates for topic prevalence.

Age patterns

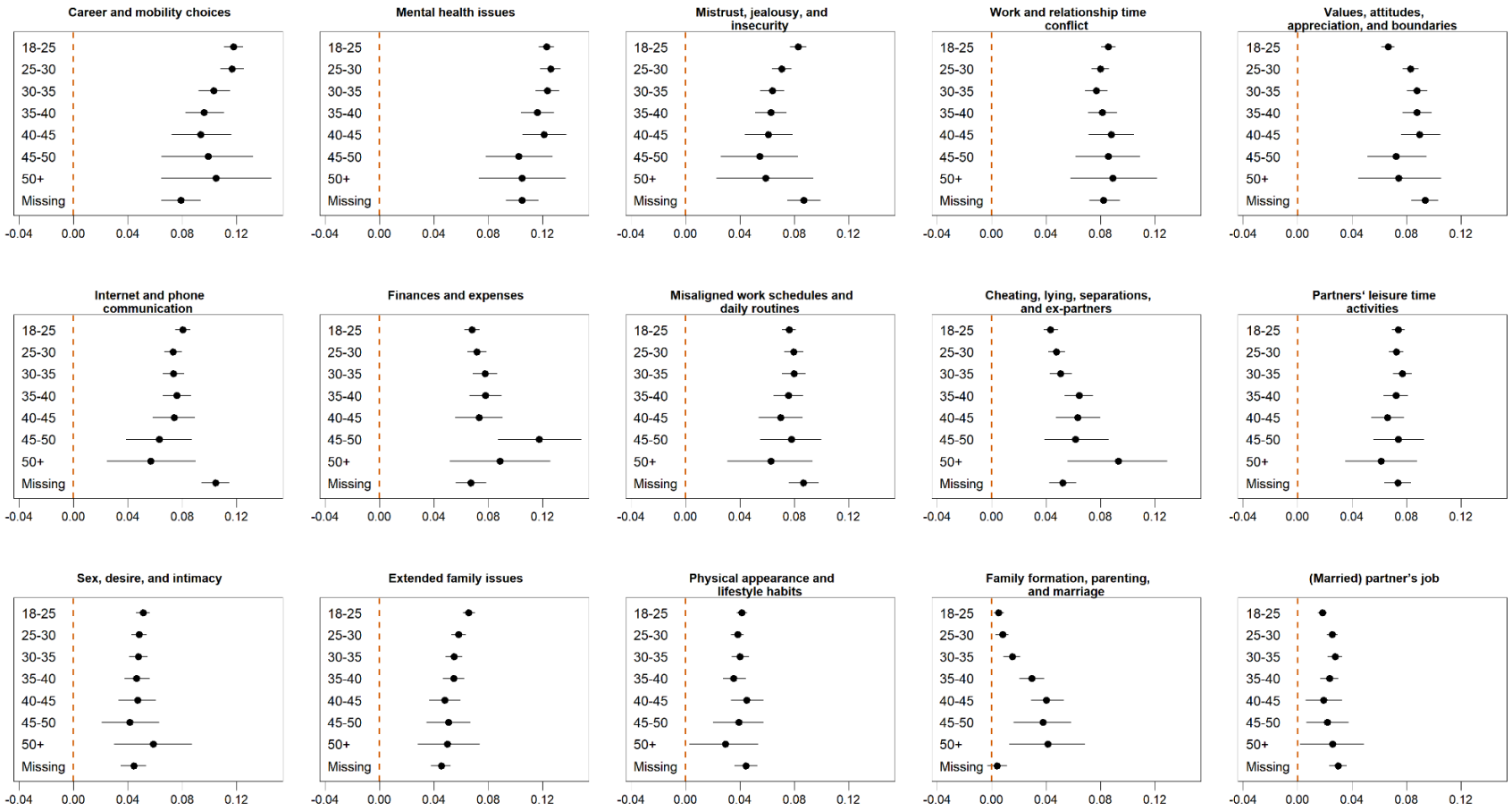

Figure D.3-2: Expected topic proportion (x-axis) by age (y-axis) for topics occurring in conflicts about paid work in romantic relationships.

Note: Structural Topic Models with k=15 topics, ordered from most to least frequent and using age, gender, of the author and partner, marital status to the partner, and time as covariates for topic prevalence.

D.3.2 Unpaid work

Topics

Table D.3-1: Unpaid work topics

Topic	Title	Top Words	FREX Words	Average expected topic proportion across documents
1	Arguments	went, call, upset, came, angri, wrong, fight, left, today, argument	yell, apolog, scream, text, drunk, door, angri, fuck, piss, anger	10.1%
2	(Mental) health issues	depress, support, life, partner, issu, struggl, anxieti, emot, stress, hard	medic, mental_health, depress, anxieti, struggl, disabl, diagnos, med, adhd, support	9.4%
3	Cleaning and laundry	dish, wash, cloth, laundri, mess, floor, dirti, trash, kitchen, bathroom	sink, dirti, wash, bathroom, floor, pile, put_away, messi, cloth, toilet	9.3%
4	Division of household tasks	chore, partner, respons, expect, household, housework, etc, task, fair, around_hous	task, equal, household_chor, chore, household, full-tim, housework, fair, respons, resent	9.1%
5	Division of labor, exhaustion and fatigue	sleep, tire, dog, school, night, laundri, wake, bed, lazi, full_tim	lazi, dog, wake, shift, tire, studi, exhaust, sleep, class, full_tim	8.3%
6	Sexual intimacy	sex, anymor, chang, friend, happi, date, seem, interest, better, cheat	sex, sex_lif, kiss, attract, sexual, intimaci, cheat, affection, connect, cuddl	7.6%
7	Childcare and pregnancy	husband, kid, babi, son, child, children, pregnant, marri, two, stay_hom	babi, pregnant, son, pregnanc, husband, daycar, born, children, diaper, month_old	7.5%
8	Finances and expenses	pay, money, bill, rent, save, financi, buy, car, paid, groceri	debt, pay, bill, money, paid, cost, rent, save, mortgag, expens	7.4%
9	Cohabitation, maturity issues, and relationship stability	boyfriend, bf, apart, place, break, friend, date, littl, guy, living_togeth	boyfriend, bf, roommat, living_togeth, mum, boyfriend', leas, j, apart, guy	6.6%
10	Young (female) partner-related issues	girlfriend, gf, friend, apart, break, person, ex, place, issu, problem	girlfriend, gf, ex, nbsp, she'll, joe, she'd, girl, breakup, apart	5.5%

11	Gaming and screen time	play, watch, game, spend, friend, phone, tv, sit, video_gam, comput	game, play, video, tv, comput, video_gam, watch, movi, playing_video_gam, xbox	4.7%
12	(Extended) family conflicts	mom, parent, famili, daughter, mother, dad, stay, sister, brother, father	sister, brother, daughter, mom, mother, sibl, dad, parent, father, visit	4.5%
13	Cooking and nutrition	eat, food, dinner, meal, made, lunch, buy, weight, night, prepar	chicken, eat, diet, meat, ingredi, soup, recip, pizza, pasta, meal	4.2%
14	Division of labor, relationship quality, and mental load	wife, kid, marri, marriag, divorc, seem, famili, school, give, issu	wife, divorc, marriag, spous, wife', marri, in-law, kid, sahm, typic	4.1%
15	Everyday responsibilities, and conflict resolution	amp, birthday, gift, gt, new, christma, card, anniversari, holiday, buy	gt, amp, gift, birthday, anniversari, lo, d, christma, lt, holiday	1.9%

Topic correlations

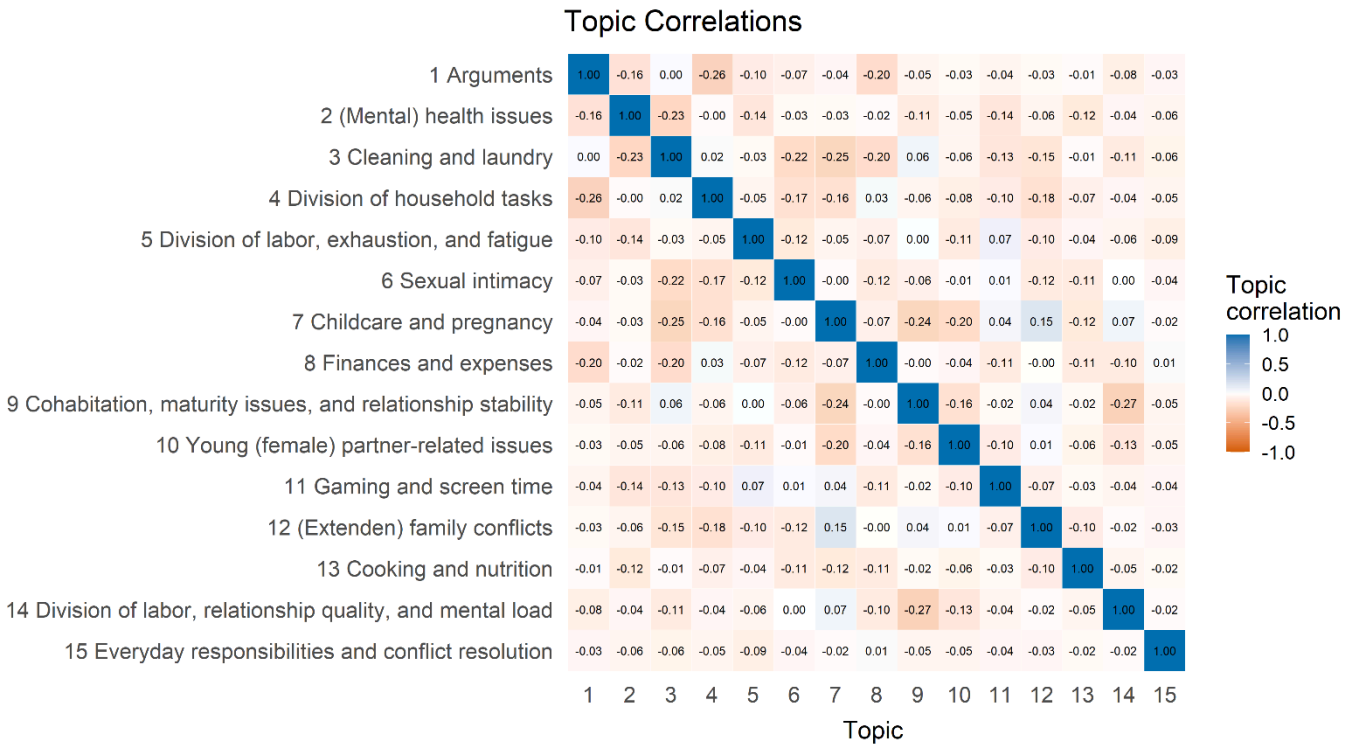

Figure D.3-3: Topic correlations for topics occurring in conflicts about unpaid work in romantic relationships.

Note: Structural Topic Models with k=15 topics, ordered from most to least frequent and using age, gender, of the author and partner, marital status to the partner, and time as covariates for topic prevalence.

Changes over time

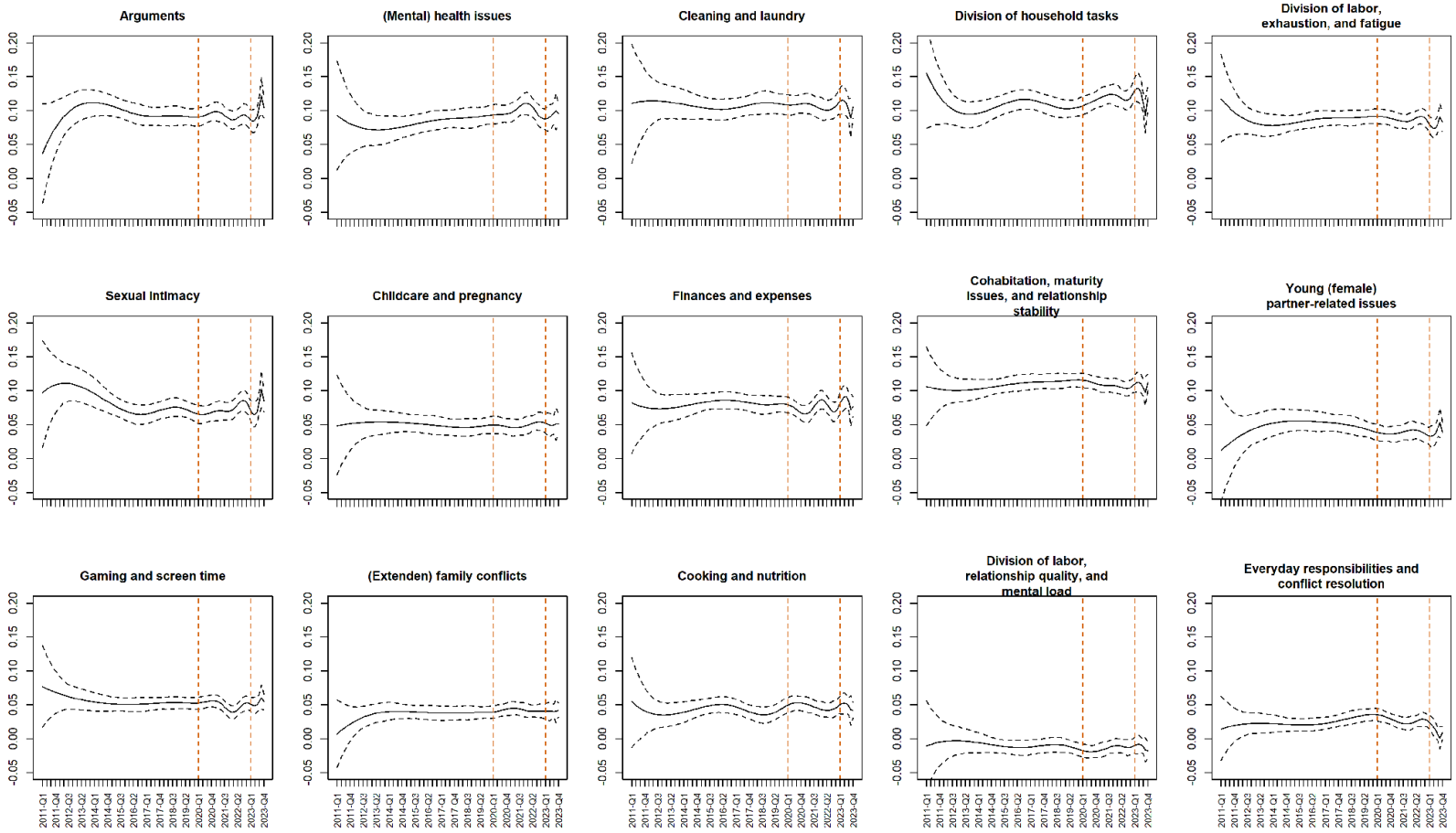

Figure D.3-5: Expected topic proportion (y-axis) over time (x-axis) for topics occurring in conflicts about paid work in romantic relationships.

Note: Structural Topic Models with k=15 topics, using age, gender of the author and partner, marital status to the partner, and time as covariates for topic prevalence. The vertical lines indicate the time periods before, during, and after the COVID-19 pandemic.

Looking at changes over time, times of crisis have led to shifts in the topics discussed (see Figure D.3-2). While some topics such as “everyday responsibilities and conflict resolution”, or “gaming and screen time”, have become less important during the pandemic, others, such as “arguments”, “division of household tasks” and “(mental) health issues”, have become more important. In contrast to the context of paid work, however, most changes occurred with a certain delay. For example, discussions about mental health issues increased in posts about paid work at the very beginning of the pandemic and then declined steadily, while corresponding discussions about mental health in connection with unpaid work only emerged some time after the start of the pandemic. Here, changes may take longer to unfold, whereas in paid work they are immediately noticeable.

E References

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A., & Lowe, W. (2015). *quanteda: Quantitative analysis of textual data* (p. 4.3.1) [Dataset]. https://doi.org/10.32614/CRAN.package.quanteda

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, (3), 993–1022.

Chae, Y., & Davidson, T. (2025). Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*, 00491241251325243. https://doi.org/10.1177/00491241251325243

Collins, W. A., Welsh, D. P., & Furman, W. (2009). Adolescent romantic relationships. *Annual Review of Psychology*, *60*(1), 631–652. https://doi.org/10.1146/annurev.psych.60.110707.163459

Coltrane, S. (2000). Research on household labor: Modeling and measuring the social embeddedness of routine family work. *Journal of Marriage and Family*, *62*(4), 1208–1233. https://doi.org/10.1111/j.1741-3737.2000.01208.x

Connolly, J., & McIsaac, C. (2013). Romantic relationships in adolesence. In M. K. Underwood & L. H. Rosen (Eds.), *Social Development: Relationships in Infancy, Childhood, and Adolescence* (pp. 180–203). Guilford Publications.

Corral, P. G., Green, A., Meyer, H., Stoll, A., Yan, X., & Reuver, M. (2024). A few hypocrites: Few-shot learning and subtype definitions for detecting hypocrisy accusations in online climate change debates. In C. Klamm, G. Lapesa, S. P. Ponzetto, I. Rehbein, & I. Sen (Eds.), *Proceedings of the 4th workshop on computational linguistics for the political and social sciences: Long and short papers* (pp. 45–60). Association for Computational Linguistics. https://aclanthology.org/2024.cpss-1.4/

Daminger, A. (2019). The cognitive dimension of household labor. *American Sociological Review*, *84*(4), 609–633. https://doi.org/10.1177/0003122419859007

Davidson, T. (2024). Start generating: Harnessing generative artificial intelligence for sociological research. *Socius: Sociological Research for a Dynamic World*, *10*, 23780231241259651. https://doi.org/10.1177/23780231241259651

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2002.06305

Fan, J., Ai, Y., Liu, X., Deng, Y., & Li, Y. (2024). Coding latent concepts: A human and LLM-coordinated content analysis procedure. *Communication Research Reports*, *41*(5), 324–334. https://doi.org/10.1080/08824096.2024.2410263

Furman, W., & Hand, L. S. (2015). The slippery nature of romantic relationships: Issues in defi nition and differentiation. In A. Booth, A. C. Crouter, & A. Snyder (Eds.), *Romance and Sex in Adolescence and Emerging Adulthood* (0 ed., pp. 199–206). Routledge. https://doi.org/10.4324/9781315652344-21

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data. A new framework for machine learning and the social sciences*. Princeton University PRESS.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Haensch, A.-C., Weiß, B., Steins, P., Chyrva, P., & Bitz, K. (2022). The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis. *Frontiers in Big Data*, *5*, 880554. https://doi.org/10.3389/fdata.2022.880554

Hand, D. J., Christen, P., & Ziyad, S. (2024). *Selecting a classification performance measure: Matching the measure to the problem* (arXiv:2409.12391). arXiv. https://doi.org/10.48550/arXiv.2409.12391

Lachance-Grzela, M., & Bouchard, G. (2010). Why do women do the lion's share of housework? A decade of research. *Sex Roles*, *63*(11–12), 767–780. https://doi.org/10.1007/s11199-010-9797-z

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2), 1–36. https://doi.org/10.1145/3643829

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., & Häussler, T. (2021). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. In *Computational Methods for Communication Science*. Routledge.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radwan, A., Amarneh, M., Alawneh, H., Ashqar, H. I., AlSobeh, A., & Magableh, A. A. A. R. (2024). Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis: *International Journal of Web Services Research*, *21*(1), 1–22. https://doi.org/10.4018/IJWSR.338222

Roberts, M. E., Stewart, B. M., & and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B., & Tingley, D. (2014). *stm: Estimation of the structural topic model* (p. 1.3.7) [Dataset]. https://doi.org/10.32614/CRAN.package.stm

Schwemmer, C., & Guyt, J. (2024). *Stminsights: A shiny application for inspecting structural topic models. R package version 0.4.3.* [Computer software]. https://github.com/cschwem2er/stminsights

Shelton, B. A., & John, D. (1996). The division of household labor. *Annual Review of Sociology*, *22*(1), 299. https://doi.org/10.1146/annurev.soc.22.1.299

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review*, *93*(2), 119–135. (1986-21992-001). https://doi.org/10.1037/0033-295X.93.2.119

Stuhler, O., Ton, C. D., & Ollion, E. (2025). From codebooks to promptbooks: Extracting information from text with generative large language models. *Sociological Methods & Research*, *54*(3), 794–848. https://doi.org/10.1177/00491241251336794

Szorc, G. (2025). *Zstandard* (Version 0.25.0) [Python]. https://pypi.org/project/zstandard/

Tillman, K. H., Brewster, K. L., & Holway, G. V. (2019). Sexual and romantic relationships in young adulthood. *Annual Review of Sociology*, *45*(Volume 45, 2019), 133–153. https://doi.org/10.1146/annurev-soc-073018-022625

Törnberg, P. (2024). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 08944393241286471. https://doi.org/10.1177/08944393241286471

von der Heyde, L., Haensch, A.-C., Weiß, B., & Daikeler, J. (2025). Using large language models for coding german open-ended survey responses on survey motivation. *Survey Research Methods*, *19*(4), 355–370. https://doi.org/10.18148/srm/2025.v19i4.8568

Watchful1. (2025, February 20). *Separate dump files for the top 40k subreddits, through the end of 2024* [Reddit Post]. R/Pushshift. https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/

Wentland, J. J., & Reissing, E. D. (2011). Taking casual sex not too casually: Exploring definitions of casual sexual relationships. *The Canadian Journal of Human Sexuality*, *20*(3), 75–91.

Wickham, H. (2012). *httr: Tools for working with URLs and HTTP* (p. 1.4.7) [Dataset]. https://doi.org/10.32614/CRAN.package.httr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686