ChatGPT, why can't anyone afford a house? On the Effects of LLM pre-annotation on Annotator Subjectivity

Emilie Francis[†], **Céline Leuzinger**[†], **Ricardo Muñoz Sánchez**[†], **Lee Gauthier**[‡]
[†]University of Gothenburg, Sweden
{emilie.francis,celine.leuzinger,ricardo.munoz.sanchez}@gu.se
[‡]Independent Researcher
lee.d.gauthier@gmail.com

Abstract

Large language models (LLMs) have often been proposed as substitutes for human annotators in a variety of tasks. At the same time, there has been increased focus on the role that human subjectivity and perspective plays in data annotation. To avoid eliminating the human role in annotation entirely, the use of LLMs for pre-annotation has been suggested as an alternative approach. In this paper, we explore to which degree this approach affects subjectivity of social media annotation in English. We focus on comments regarding the current status of the housing market and label them for concern level, factors affecting housing affordability, and aspects that authors claim either exacerbate or improve the situation. To investigate this, we design an experiment involving two rounds of annotation: the first, a dataset annotated by humans only; and the second, a dataset with LLM pre-annotations curated by the same human annotators. We observe that the second setting leads to much higher agreement, as well as significant changes in label distribution and co-occurrence. Similar shifts do not appear in the LLM labels. Our findings show that use of LLMs in the annotation process leads to convergence in annotations and, thus, to an erosion of human subjectivity.

Keywords: subjectivity, annotation, agreement, LLMs, social media analysis

1. Introduction

NLP has experienced a drastic paradigm shift with the advent of highly performant Large Language Models (LLMs). Autoregressive language models have become more commonplace throughout the entire machine learning pipeline, reducing the need for multiple interconnected pieces. Although this is not a new trend, autoregressive generative language models have been increasingly used to label data (Karim et al., 2025) and automated evaluation (Bavaresco et al., 2025).

However, the use of LLMs comes with its own risks and challenges. For instance, they are known to hallucinate Ji et al. (2024) and reproduce social biases Gallegos et al. (2024). There have been multiple proposed solutions when using these systems for tasks like data labelling to ensure the quality and reliability of the system's outputs. Two of these are 'human in the loop' (Pangakis and Wolken, 2024; Wang et al., 2025) and 'hybrid intelligence' (Dellermann et al., 2019).

A major flaw of such approaches is 'anchoring bias', which shows that humans tend to favour the first option presented (Tversky and Kahneman, 1974). Another influencing effect is 'automation bias', which specifies that humans overly rely on machine output for automated systems they consider trustworthy (Dzindolet et al., 2003). Such biases can have a strong impact on annotation outcomes, as human reviewers may be tempted to agree with LLM suggestions. This could in turn lead to what Schroeder et al. (2025) describe as a 'homogenization of insights', reducing annotation diversity and impacting downstream performance. This is particularly problematic for subjective tasks, where an accurate depiction of reality would naturally entail a variety of annotator opinions and judgments (Wan et al., 2023).

To determine the impact of anchoring and/or automation biases, we measure the effect of LLM annotation on human subjectivity. We focus on three research questions: *(i) Does usage of LLMs as a tool for pre-annotation result in convergence of human-annotated labels? (ii) Are different sets of labels more likely to be used when LLMs are involved in the annotation process? (iii) Are humans more efficient at data labelling in terms of time when curating LLM-generated labels?*

To answer these questions, we gather a dataset of social media comments discussing housing affordability in major cities of six English speaking countries. We engage three community annotators in two experimental annotation settings. In the first experiment, annotators are asked to annotate 500 comments for four label categories (*concern score, factor, aspect-improvement, aspect-exacerbate*). In the second setting, GPT-4.5 is used to label a different set of 500 comments whose output is curated by the same annotators.

Our results show that use of an LLM for pre-annotation leads to a drastic decrease in terms of annotation time. However, this comes at a price concerning annotation quality. There is a non-

Proceedings of NLPerspectives @ LREC 2026, pages 98–111
12 May 2026.

negligible increase in inter-annotator agreement and a significant shift in terms of label distribution.

For instance, labels that denote objective factors impacting housing prices are much more common when LLMs are involved compared to human-only annotations. These results suggest that usage of LLMs in the annotation process strongly influences human subjectivity. Our contributions can be summed up as follows:

1. A dataset annotated with labels for subjective measures of concern and commonly mentioned features of housing affordability in several major cities
2. An evaluation of the use of LLM pre-annotated data for annotation in a subjective task, for both prescriptive and descriptive ordinal and categorical labels
3. An exploration of changes in annotation patterns when annotators are provided with LLM pre-annotated data

The following section presents and overview of the use of LLMs for dataset labelling, both as a replacement for human annotators and as assistants, as well as commonly noted challenges in the use of LLMs for such work.

2. Background

In the past few years, generative language models have been proposed as a alternative for human annotators. The core argument behind this is that it takes less resources to automate the annotation process, both in terms of time (Choi et al., 2024) and money (Wang et al., 2021). This approach has been followed to label data in fields such as argumentation mining (Lindahl, 2025) and the social sciences (Ziems et al., 2024), among others. (Gao et al., 2023) describe the LLM as a mediator facilitating discussion between experts, noting that its usage leads to increased agreement.

However, several arguments have been leveraged against taking LLMs as if they were human annotators. Atreja et al. (2024) and Baumann et al. (2025) show that they are highly sensitive to prompt variation, which can lead to major changes in label distribution. Not only that, but LLMs are known to suffer from hallucinations (Ji et al., 2024) and to reproduce social biases (Gallegos et al., 2024). Further, the use of LLM-generated labels can lead to lower performance across several kinds of tasks (Li et al., 2023; Pangakis and Wolken, 2024)

Both 'human in the loop' and 'hybrid intelligence' approaches have been suggested as a way to get around some of these issues (Schroeder et al., 2025; Wang et al., 2025) For instance, Ziems et al. (2024) describe human annotators as 'key' to avoid LLM biases and hallucination in social sciences datasets. However, the use of humans as curators (as opposed to annotators) poses a new set of problems. Humans are known to suffer from different cognitive biases, among them anchoring and automation biases.

Anchoring bias refers to the human tendency of relying on the first piece of information presented, even if it is irrelevant to the task at hand (Tversky and Kahneman, 1974). This phenomenon is widely studied, and has been shown occur in a variety of decision making processes, such as purchasing decisions, legal judgments or time estimation (Furnham and Boo, 2011).

Automation bias refers to the human tendency to favour suggestions coming from automated systems, even if they are in direct contradiction with information that is known to be true by the decision-maker (Dzindolet et al., 2003). Wilcox (2023) argues that human in the loop approaches lead to increased risks of automation bias while reducing accountability.

Choi et al. (2024) showed that human curators heavily rely on LLM annotation even in highly specialized settings. Meanwhile, Schroeder et al. (2025) found that human annotators have a strong tendency to follow LLM suggestions even in subjective tasks, which in turn leads to higher agreement and significant shifts in terms of label distribution.

High agreement between annotators has been historically regarded as the gold standard in machine learning (Basile et al., 2020). However, this ignores the fact that most NLP tasks have a subjective aspect to them. As such, data cleaning and harmonization leads to less rich datasets and risks erasing underrepresented voices (Klenner et al., 2020). Thus, factors that force higher agreement can be detrimental to our data, among them anchoring and automation biases.

This can be seen when taking LLM-generated labels as if they were actual annotations. Gao et al. (2023) show that using humans to curate LLM-generated labels strengthened agreement and diminished the diversity of labels. Schroeder et al. (2025) echo these insights, noting that it homogenizes labels and reduces the diversity of judgments. Moreover, Choi et al. (2024) observe that topics selected by the LLM tend to be broader and less nuanced than topics selected by humans.

3. Methodology and Data

We use a three step process to compare patterns in annotation before and after exposure to data augmented with LLM pre-annotations. Taking a raw dataset (D) of 1000 English comments on Reddit from local forums for several major cities experienc-

ing challenges with housing affordability, we split into two sets of 500 (D^1 and D^2).

In the first step, a team of three annotators (*A*, *B*, and *C*) annotate the first set of comments without LLM pre-annotation (D_h^1). We define four distinct categories for annotation:

1. **Concern Score:** how concerned the comment's author appears to be about not being able to find or maintain adequate housing in the city
2. **Factor:** an objective set of measures for housing affordability used in OECD and EU countries
3. **Aspect-Exacerbate:** aspects that a comment's author claims worsen affordability in the housing market
4. **Aspect-Improve:** aspects that a comment's author claims improve affordability in the housing market

The first two categories follow a prescriptive paradigm (Röttger et al., 2022), while the remaining two are descriptive and developed collaboratively by the annotators during annotation. Details of the annotation process are elaborated on in Section (3.2).

In the second step, annotators and researchers agree on standardized guidelines which include the prescriptive categories and a refined set of the descriptive labels developed in step one. We use few-shot prompting with the standardized guidelines to generate labels for an unseen set of 500 comments (D_l^2) and the initial 500 comments with GPT-4.5 (D_l^1). As LLMs tend to struggle with multi-task prompts (Ma et al., 2025), this step was broken down into four tasks using one few-shot prompt for each label category.

In the final step, annotators independently curated the model output from step two based on the standardized guidelines. We then compare inter-annotator agreement between annotators in the human-only annotated dataset (D_h^1) with that of the human-curated dataset (D_h^2).

Measuring Agreement: Inter-annotator agreement for ordinal labels (*concern score*) was computed with Krippendorff's alpha and agreement between categorical labels was computed with Sørensen-Dice similarity. While Jaccard similarity is more widely used to measure agreement for multi-label categorical annotation tasks in NLP, it is designed for only two annotators. As averaging pairwise Jaccard scores to report agreement between three annotators risks information loss, we use three-way Sørensen-Dice similarity instead (Diserud and Ødegaard, 2007; Magurran, 2003).

Sørensen-Dice similarity, widely used in ecology to measure similarity of site composition, has been extended from a two-way measure to accommodate three or more entities (Diserud and Ødegaard, 2007). The general formula for a multiple entity similarity measure is defined as:

$$S_T = \sum_{i<j} a_{ij} - \sum_{i<j<k} a_{ijk} + \sum_{i<j<k<l} a_{ijkl} - \cdots$$

$$C_S^T = \frac{T}{T-1}\left(\frac{S_T}{\sum_i a_i}\right)$$

Where a_i is the number of variables for entity A_i, a_{ij} is the number of variables shared by entities A_i and A_j, and so on. T is the total number of entities. The Sørensen-Dice similarity measure lies between 0 and 1, where 0 is no agreement and 1 is perfect agreement. For a $T = 3$ measure of similarity between annotators A, B, and C, we calculate Sørensen-Dice similarity with:

$$S = \frac{3}{2}\left(\frac{ab + ac + bc - abc}{a + b + c}\right)$$

Where ab is the number of labels shared by annotators A and B, ac is the number of labels shared by A and C, etc. Inter-annotator agreement is calculated for each comment and averaged over the entire annotation set for both D_h^1 and D_h^2.

Label Patterns: We explore differences in label patterns between D_1 and D_2 across two parameters: frequency and label co-occurence. We compare the frequencies of each label, for each annotator for D_h^1 and D_h^2. We check for statistical significance and effect size using chi-square test. We also compare the frequencies for each label for the LLM-annotated data, D_l^1 and D_l^2.

To analyze differences in label co-occurrence, we use Fisher's Exact test in pairs of labels where at least one showed statistically significant changes for that annotator. We then calculate the effect size of this co-occurrence by calculating φ of the statistic:

$$\varphi = \sqrt{\frac{\text{statistic}}{\text{\# of observations}}}$$

We can interpret values of φ as follows: 0.1 and lower is small, between 0.1 and 0.3 the effect is medium, and an effect size of 0.3 or more is large.

Time: Finally, we also compare time in hours taken to complete steps one and three by each annotator. The following sections describe the dataset and the annotation process.

Local Subreddit	Comments
Vancouver	198
Toronto	189
Los Angeles	154
San Francisco	102
Melbourne	95
London	88
Sydney	88
Bristol	35
Auckland	29
Dublin	22

Table 1: Number of comments per local subreddit included in the full dataset of 1000 comments.

3.1. Data

We select six English speaking countries identified among the top locales for housing cost burden in OECD reports: *U.S.A., Canada, U.K., Ireland, Australia, and New Zealand* (OECD, 2025a).

For each country, we select the largest two cities by population and collect threads from the local subforums (subreddits) from the Reddit PushShift Dataset (Baumgartner et al., 2020). All comments are in English, dating from 2013 to 2023, and belong to one of the following subreddits: *r/LosAngeles, r/SanFrancisco, r/Vancouver, r/Toronto, r/London, r/Bristol, r/Sydney, r/Melbourne, r/Dublin, and r/Auckland*. For Ireland and New Zealand, only the most populous city was included due to lack of subreddit content for the second most populous city.

To restrict data to discussions on housing and affordability, we filter thread titles based on a list of housing terms. This list was developed by a community panel of volunteers, at least one panel member local to each country in the study. Threads whose title did not contain at least one term from the list were removed.

After automatic filtering, we selected ten threads from each city based on number of responses. The titles of these threads were manually reviewed by the authors to remove any that did not explicitly discuss housing affordability, such as advertisements for roommates or apartments. This resulted in five threads for each local subreddit.

For the remaining threads, we remove all comments with a token count lower than 100[1] to ensure there would be enough content for annotators to judge. This yielded 5,000 comments from which 1000 were randomly selected, 500 for human-only (D_h^1) annotation and 500 for pre-annotation with the LLM (D_h^2).[2]

All comments were cleaned to remove Reddit markdown formatting and replace personally identifiable information, such as usernames, with placeholders.

3.2. Dataset Annotation

Annotation was carried out by a team of three volunteers, each of which have spent one or more years living in one of the cities in the study. Two annotators are women in their late 20s to early 30s from a middle-class background in mid-sized cities, while the third is a man in his thirties from a small-town working class background.

The dataset includes four categories of labels: *concern score, factor, aspect-improve, and aspect-exacerbate*. As described in Section 3, two categories (*concern score* and *factor*) are prescriptive labels based on the *More Effective Social Protection for Stronger Economic Growth Survey* (OECD, 2025b) and *Building for a better tomorrow: Policies to make housing more affordable Brief* (OECD, 2021) reports respectively. For our dataset, we combined the survey responses 'not so concerned' and 'somewhat concerned' from OECD (2025b) into the 'mixed' label. The four labels for concern score are defined below:

1. **Off Topic:** The comment does not talk about housing/housing affordability in any way.
2. **No Concern:** The comment does not appear to express concern toward housing at all. May deny there is a problem or attempt to invalidate others' concerns and refute claims.
3. **Mixed:** The comment appears to express some concern, but is mixed. Usually agrees there is a problem, but might discuss the topic more analytically.
4. **Concern:** The comment expresses clear concern toward the housing situation and affordability.

The categorical labels, *factor*, *aspect-exacerbate* and *aspect-improve*, and their corresponding definitions are presented in Appendix A. The two descriptive categories were developed in a collaborative document using open coding (Khandkar, 2009). Additionally, our annotation procedure followed the perspectivist paradigm (Basile et al., 2020; Cabitza et al., 2023). That is, we kept all labels and did not aggregate annotations in any manner.

Two labels were annotator specific. The first, 'money', was used by one annotator in the aspect-improve category, but was later merged into the 'bootstraps'[3] label. The second, 'demand', was only

[1] Based on the 50th percentile for comment length.

[2] And 7 to the Dwarf Lords who became wealthy beyond measure.

[3] From the phrase "pull oneself up by one's bootstraps", meaning "getting oneself out of a difficult situation only with one's own effort".

Label Category	Krippendorff's Alpha (α)	
	human-only (D_h^1)	LLM-assisted (D_h^2)
Concern	0.63	0.72

(a) Inter-annotator agreement for ordinal labels.

Label Category	Sorensen-Dice Similarity (S)	
	human-only (D_h^1)	LLM-assisted (D_h^2)
Factor	0.70	0.86
Aspect (Exacerbate)	0.78	0.84
Aspect (Improve)	0.82	0.88

(b) Inter-annotator agreement for categorical labels.

Table 2: Inter-annotator agreement for the different labels. Note that agreement for ordinal labels is reported in terms of Krippendorff's Alpha, while Sørensen-Dice Similarity is used for categorical ones.

used by one annotator in the *aspect-exacerbate* category.

We created two versions of the dataset, a standardized and non-standardized version. The non-standardized version contains all labels used by annotators to preserve the full variety of annotator perspective for future tasks. The standardized version was created for ease of comparing annotator agreement. This version includes *concern score*, *factor*, and the set of aspect labels agreed upon by all annotators mentioned in Section (3). Labels with fifteen or fewer instances were removed from the standardized label set and subsequent analysis. Annotators used only the the standardized label set for curation of the LLM output in D_h^2. The set of standardized labels and corresponding definition are presented in Appendix A.

In addition, annotators were asked to record time taken to complete both D_h^1 and D_h^2.

In the following section, we compare the results of inter-annotator agreement between D_h^1 and D_h^2, as well as the difference in time taken to complete each task. We also present an analysis of label distribution to explore differences in annotation patterns between D^1 and D^2. Finally, we note specific instances in which annotators believed their individual experience or contextual knowledge were an asset in annotation and how their annotations differed from the LLM and each other.

4. Experimental Results

4.1. Inter-Annotator Agreement

We observe a large increase in inter-annotator agreement between D_h^1 and D_h^2 for both ordinal labels with Krippendorff's Alpha and categorical labels with Sørensen-Dice Similarity. The results of inter-annotator agreement are reported in Table 2.

For both D_h^1 and D_h^2, *concern score* had the lowest agreement. Unlike labels in the other categories, which may be triggered by specific vocabulary, *concern score* is purely subjective as it requires annotators judge the emotional state of a comment's author. The agreement for this category increased by nearly ten points from D_h^1 ($\alpha = 0.63$) to D_h^2 ($\alpha = 0.72$), showing that annotations became less varied when annotators were provided with pre-annotated data.

The *factor* category showed the biggest increase comparing the D_h^1 ($S = 0.7$) with D_h^2 ($S = 0.86$). A potential explanation for this may be that, unlike the aspect labels which are specific, *factor* labels represent broader groups of factors for housing affordability identified by the OECD. Additionally, annotators may have been less certain of their interpretation of these labels as they were pre-defined rather than developed based on their own understanding.

The *aspect-exacerbate* and *aspect-improve* categories showed the highest agreement in D_h^1 ($S = 0.78$ and $S = 0.82$) and the smallest increase when comparing D_h^1 with D_h^2 ($S = 0.84$ and $S = 0.88$). High agreement for these categories overall is likely due to the presence of triggering vocabulary, such as 'NIMBY' for the 'NIMBYism'[4] label. However, there is small (0.06) increase which suggests that annotators have a slight tendency to accept the LLM output for these labels as well.

Comparing inter-annotator agreement between D_h^1 and D_h^2 reveals that disagreement is considerably reduced between annotators when provided with pre-annotated data. This suggests that annotators are more likely to accept the provided labels, resulting in a loss of subjectivity as annotations become overly influenced by the LLM.

[4]NIMBY - 'not in my backyard'. Refers to people who oppose development near their owned property.

Label	Odds Ratio		
	A	B	C
Real Price	**1.43**	**1.53**	**3.70**
Quality	**1.56**	**2.81**	**7.56**
Availability	**1.64**	**1.90**	**5.30**
Housing to Income	1.18	1.35	**3.25**
Building	**1.78**	**2.07**	**4.92**
Bootstraps	1.28	**2.74**	**0.59**
Government Policy (I)	1.21	1.42	**2.05**
Relocation	**2.71**	**3.58**	**2.53**
Wage Price Imbalance	1.48	**1.60**	**5.46**
Government Policy (E)	1.37	**1.78**	**1.60**
Cost of Living	0.90	1.14	**4.40**
Foreign Investment	**0.52**	0.70	0.64
NIMBYism	0.64	0.84	0.86
The Rich	1	1.20	1.48

Table 3: Odds ratio per label and annotator (A, B, C). Bold values are statistically significant.

Label	Odds ratio
Availability	0.97
Housing to Income	0.96
Quality	1.23
Real Price	1.16
Bootstraps	0.70
Building	0.98
Government Policy (I)	0.88
Relocation	**1.66**
Cost of Living	1.02
Foreign Investment	0.63
Government Policy (E)	1.05
NIMBYism	0.71
The Rich	0.82
Wage Price Imbalance	1.03

Table 4: Odds ratio per label for the LLM-only annotation. Bold values are statistically significant.

4.2. Label Frequency

We also take a look at changes in terms of individual labels between D_h^1 and D_h^2. To study changes in label frequency, we compute the frequency of each label in both experimental settings for each annotator. We then perform a chi-square test (Pearson, 1900) to check for statistically significant changes in label frequency. We also compute the odds ratio for effect size. We report these values in Table 3.

To ensure these effects are not due to differences in the data, we also compare label frequency between D_l^1 and D_l^2. The odds ratio for each label is reported in Table 4.

The odds ratios comparing D_l^1 and D_l^2 are relatively close to 1, with values ranging from 0.70 to 1.66. The odds ratios for the human annotations, D_h^1 and D_h^2, display a wider variation, with values ranging from 0.52 to 7.56 across annotators. This suggests that the LLM was more consistent across datasets than humans.

Three out of the four labels that are significantly more frequent for all annotators are factors ('real price', 'quality', 'availability'). We also observe cross-annotator differences in label frequency. For annotator A, there are six labels whose distributions more closely approximated those from the LLM in D_h^2. The number of such labels are nine and twelve for annotators B and C, respectively. This shows that while all annotators were influenced by the LLM, the degree of such influence varies between individuals.

We also look at label co-occurrence to determine whether these changes in labels led to changes in how often they appear with each other. We focus on labels where the chance of them co-occurring was statistically significant and the co-occurrence rate was medium or large. These labels can be found in Table 5 For all annotators, we see an almost complete change of labels that are likely to co-occur, with none of these changes including the label 'relocation', which was the only one that showed a statistically significant change in label distribution for the LLM.

4.3. Annotation Time

Table 6 shows the time in hours taken by each annotator to complete each of the two experiment sets. As we can see, there is a noticeable reduction in time taken to annotate the same number of samples when using an LLM for pre-annotation compared to human-only annotation. Even though this appears to contradict the findings of Schroeder et al. (2025), it goes in line with previous research on human curation of LLM-generated labels (e.g. Choi et al., 2024).

4.4. Annotator Observations

Annotations in D_h^1 were often influenced by annotator world knowledge and personal background, several examples of such instances are presented below. To preserve anonymity, all named entities in provided comment examples have been redacted to obfuscate location.

Although not explicitly mentioned in the text, all annotators added the label 'COVID' for the comment in (**X**). The annotators explained that they used the date of the comment as context. The LLM did not attach the 'COVID' label to this comment, which shows that this real-world contextual knowledge was not incorporated into the LLM output.

Annotator	Labels		Effect Size $\mathbf{D_h^1}$	Effect Size $\mathbf{D_h^2}$
A	NIMBYism	**Availability**	small	medium
B	NIMBYism	Government Policy (I)	medium	N/A
	Relocation	Cost of Living	N/A	Medium
C	**Building**	**Quality**	medium	N/A
	Building	**Availability**	medium	N/A
	Foreign Investment	**Government Policy (E)**	small	medium

Table 5: Changes in co-occurrence between the human-only annotation and the LLM pre-annotation setting. Labels in bold are those that showed a statistically significant change for that annotator between annotation rounds. N/A denotes that the co-occurrence was not statistically significant in that setting.

Annotator	$\mathbf{D_h^1}$	$\mathbf{D_h^2}$
A	15.5	9.2
B	10.4	6.7
C	6	2

Table 6: Annotation times per annotator in hours. D_h^1 refers to the human-only annotated dataset, and D_h^2 the LLM pre-annotated dataset.

> **X**: If this gets truly bad, the government will need to put mortgages and even rent on hold (as I think other countries have already done in response to people not being able to work).

In another comment (**Y**), annotator *C* gave the label 'concern' and the others (including the LLM) 'off topic'. While the comment does not explicitly mention housing, annotator *C* based their interpretation on personal experience discussing the housing situation with peers living in the city.

> **Y**: I would love to live somewhere else, if [...] cities/towns hadn't f*** themselves up to cater to driving everywhere. At least in [...] I can walk and bike to places, even if it's not that safe. Go elsewhere and you're trudging through parking lots and stroads.

Annotators also had very different interpretations of sarcasm. For the comment in (**Z**), annotators *B* and *C* gave the label 'off topic' while annotator *A* and the LLM gave the label 'concern'. Annotator *A* only applied the 'cost of living' for *aspect-exacerbate*, while the LLM used both 'cost of living' for *aspect-exacerbate* and 'bootstraps' for *aspect-improve*. Annotator *A*, local to the city the comment discusses, used personal experience from conversations with peers mocking the 'bootstraps' argument and did not judge it as a serious suggestion by the comment's author.

> **Z**: No kids, dual income. We eat spaghetti noodles with butter 21 meals a week and our favourite pastime is sleeping 12 hours so we can save on [...] bills. We have sex on a plastic sheet to cut down on laundry. I haven't smiled since 2012. Waste of calories. Just ten more years of this and we will be able to retire early in a paid off townhouse just outside [...]. The five years before my unfortunate heart attack are going to be epic.

Overall, annotator background and experience played a large role in their interpretation. Annotators also reported they were more confident in annotations for comments from locales in which they had lived or spent some time in.

5. Discussion

Despite the potential resource reduction for descriptive annotation paradigms, there are major disadvantages to using LLMs for pre-annotation in subjective tasks. Like Schroeder et al. (2025), Choi et al. (2024), and Gao et al. (2023), the results of our analysis show that LLM pre-annotation significantly influences human annotation in such a way that is detrimental to subjective tasks (Wan et al., 2023).

For all categories, we observe a large increase in inter-annotator agreement. This indicates that annotators were more likely to rely on LLM pre-annotations, regardless of whether labels follow a prescriptive or descriptive paradigm (Röttger et al., 2022).

We also observe significant differences in label frequency between the annotation rounds. *Factors* in particular are more frequent in the LLM curated output. *Factors* describe measures of housing affordability, and are arguably the most general labels in our annotation frameworks. A label such as 'real price' can be understood in a variety of ways (direct mention of a buying price, presence of adjective relating to costliness, direct mention of a price increase, etc.); the same can be said

about 'quality' (price-quality relationship, quality of the infrastructure, quality of life in the neighbourhood, etc.) and 'availability' (lack of housing, lack of affordable housing, direct mention of a number of units being built, etc.). *Factors* are already among the most common labels in the human-only annotation round; the sharp increase in their frequency in the curated output could indicate that the LLM has a tendency to 'over-label' with broad, general tags that can have a variety of interpretations. Human annotators, on the other hand, could have a finer interpretation of both the label and the comment at hand, and therefore chose not to use general labels as often as LLMs. These results are in line with Choi et al. (2024), who observe that LLMs tend to select broader topics than humans.

Such change in label frequency suggests that annotators are affected by a combination of anchoring and automation bias. As Klenner et al. (2020) argues, homogenization as a result of LLM pre-annotation risk erasing valuable perspective in subjective annotation tasks. As exemplified in Section 4.4, there were several instances in which either the annotators disagreed with each other or the LLM as an effect of real world knowledge or annotator experience. These insights may be sacrificed as a consequence of using LLMs to pre-annotate data.

These effects are not trivial. There has been a push in recent years in NLP to acknowledge the importance that annotator subjectivity plays, both in terms of representation and in terms of modelling (Cabitza et al., 2023). Even though the focus has often laid on non-aggregation of labels, there is also a risk of annotator subjectivity erosion by other means (e.g. anchoring and automation biases). Our results confirm that label convergence in LLM-assisted annotation is a significant problem for subjective tasks that greatly impacts both prescriptive and descriptive annotation paradigms.

While previous explorations into the use of LLMs for augmenting human annotation have shown mixed results concerning resource benefit (Schroeder et al., 2025; Choi et al., 2024), our results indicate that they do have the potential to reduce annotation time.

However, time reduction from LLM pre-annotation may depend on the type of task and annotation setup. Schroeder et al. (2025) used prescriptive labels in a multiple choice environment and reported no meaningful difference in time between annotations with and without LLM pre-annotation. In contrast, our annotation setup utilized a mix of prescriptive and descriptive annotation in an open environment wherein annotators could freely add and adjust labels as necessary. Potentially, time reduction from LLM pre-annotation is greater for descriptive annotation paradigms.

6. Conclusion

In this paper, we have presented an investigation into the use of LLMs for pre-annotation in a subjective task, concern regarding a crisis of housing affordability in several major cities in the Anglosphere. We design a three step experimental process to compare inter-annotator agreement and patterns in label distribution with and without the use of LLMs for pre-annotation. In the first step, annotators worked collaboratively to develop labels as they annotated a dataset of 500 comments without the use of LLMs. In the next step, we pre-annotated an unseen set of 500 comments plus the original 500 comments with an LLM. In the final step, annotators curate the output of the LLM pre-annotations. Annotators reported a large reduction in time taken to complete the same number of comments when provided with pre-annotations. However, results also showed a significant difference in both inter-annotator agreement and label distribution with the use of LLM pre-annotations. Our findings show that annotations tend to converge when exposed to LLM labels, indicating that annotators are more likely to accept the LLM output as is. This loss of subjectivity may have serious consequences for downstream tasks.

While providing annotators with LLM augmented data has the potential to speed up annotation time, it comes at the cost of averaging human perspective. When presented with LLM augmented data, annotations become more homogeneous and result in an erasure of annotator perspective. Therefore, we conclude that use of LLMs for pre-annotation is detrimental to tasks where the nuance and subjectivity of human diversity are valuable.

Future studies could explore the presence of automation bias by conducting blind curation, containing a mixture of human and LLM annotations. If the curators' attitude does not change between human and LLM annotations, it implies that anonymising the source of the annotation is an effective preventer of automation bias. Conversely, if the curators correct human annotations more often than the LLM's when knowing the source, it could imply that curators experience automation bias. Another follow-up study could test the influence of automation bias and anchoring bias, by presenting a choice of ordered annotations to the curators; if the curators tend to choose the first option presented, regardless of the source of the annotations, then anchoring bias is the predominant factor. If the curators tend to choose the LLM annotations, even when they are not presented first, then automation bias is more prevalent.

Limitations

This paper has a few limitations in the data and annotations. First, although the issue of housing affordability affects many regions globally, our dataset and annotations are focused only on the Anglosphere. As a consequence, observations on housing affordability are limited to perspectives from the English speaking global North and may not be generalizable to other locations. In addition, as our data is sourced from a single website, the comments reflect only the Reddit community perspective for each locale. Additionally, the number of annotators is limited as we were not able to secure an annotator local to each region included in the dataset. Additionally, our dataset is limited to only 500 comments for each test case. Finally, our experiment lacks a control group. Although we have attempted to mitigate this by reporting odd ratio, there may be differences between the 500 comments in the human-only annotation set and the 500 LLM pre-annotation set that are not isolated from other potentially influencing effects.

Ethical Considerations

First, all annotators involved in the project were given authorship on the paper as compensation for their contributions.

There are two main considerations that come into play regarding the data used for these experiments:

1) Although we have taken steps to anonymize comments by removing clear PII, it is incredibly challenging to control for all information that has the potential to identify individuals. Data de-identification is still an open problem, with multiple considerations that must be taken in mind. See Volodina et al. (2025) for a deeper discussion on this topic.

2) Even though all comments are freely viewable on public facing forums, the comment authors did not give informed consent for their data to be used for research purposes. Because of this, we opted not to re-publish raw text from the resulting dataset in a public forum.

Bibliographical References

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 122–145. Association for the Advancement of Artificial Intelligence (AAAI).

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR workshop proceedings*, volume 2776, pages 31–40. CEUR-WS.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv pre-print*, 2509.08825.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Alexander S. Choi, Syeda Sabrina Akter, J. P. Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22032–22054. Association for Computational Linguistics (ACL).

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.

Ola H Diserud and Frode Ødegaard. 2007. A multiple-site similarity measure. *Biology Letters*, 3:20–22.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718.

Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40:35–42.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka Wei Lee, Simon Perrault, Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka Wei Lee, and Simon Perrault. 2023. CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *arXiv pre-print*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2024. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55.

Md Monjurul Karim, Sangeen Khan, Dong Hoang Van, Xinyue Liu, Chunhui Wang, and Qiang Qu. 2025. Transforming data annotation with AI agents: A review of architectures, reasoning, applications, and impact. *Future Internet*, 17.

Shahedul Huq Khandkar. 2009. Open coding. *University of Calgary, Technical report*, 23.

Manfred Klenner, Anne Göhring, Michael Amsler, Sarah Ebling, Don Tuggener, Manuela Hürlimann, and Martin Volk. 2020. Harmonization sometimes harms. In *SwissText/KONVENS*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10443–10461. Association for Computational Linguistics (ACL).

Anna Lindahl. 2025. LLMs as annotators of argumentation. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 242–252. Association for Computational Linguistics (ACL).

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 2472–2495. Association for Computational Linguistics.

A.E. Magurran. 2003. *Measuring Biological Diversity*. John Wiley & Sons.

OECD. 2021. OECD affordable housing database - indicator HC 1.5 overview of affordable housing indicators. Technical report, Organisation for Economic Co-operation and Development (OECD).

OECD. 2025a. OECD affordable housing database - indicator hc 1.2. house prices. Technical report, Organisation for Economic Co-operation and Development (OECD).

OECD. 2025b. OECD affordable housing database - indicator HC 1.4. subjective measures on housing. Technical report, Organisation for Economic Co-operation and Development (OECD).

Nicholas Pangakis and Samuel Wolken. 2024. Keeping humans in the loop: Human-centered automated annotation with generative AI. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1471–1492.

Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? Investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795. Association for Computational Linguistics (ACL).

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, Greta Lisa Södergård, Xuan-Son Vu, and T Attendee Lindström Tiedemann. 2025. Towards shared standards for psuedonymization of research data. In *Huminfra conference 2025*, pages 101–114. Huminfra.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:14523–14530.

Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. In *CHI Conference on Human Factors in Computing Systems*, volume 1, pages 1–7. Association for Computing Machinery.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205. Association for Computational Linguistics (ACL).

Lauren Wilcox. 2023. No Humans in the Loop: Killer Robots, Race, and AI. In Jude Browne, Stephen Cave, Eleanor Drage, and Kerry McInerney, editors, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford University Press.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50:237–291.

Language Resource References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 830–839.

A. Labels and their definitions

This appendix details the different labels we used for the different categories we considered.

Label	Definition
Housing to Income	The comment mentions factors such as housing price or rent price in comparison to income. Can include comments on the percentage/ratio of one's income spent on rent, mortgage, purchasing property, etc.
Quality	The comment mentions factors used to assess housing quality, such as number of bedrooms, location, bathrooms, pet friendliness, overcrowding, etc.
Availability	The comment mentions housing availability or lack of. Can include comments talking about lack of low-income designated housing, being unable to find a place to rent, etc.
Real Price	The comment talks about actual prices, fees, increases, or decreases in the price of property and rent. May be compared to other countries, cities, or other periods of time.

Table 7: Labels for the "factor" category.

Label	Definition
Government Policy (I)	The comment mentions or makes suggestions of government actions and policies that contribute to an improvement in housing affordability.
Building	The comment mentions that building more housing or increasing density contributes to improved housing affordability.
Bootstraps	The comment mentions that acquiring money, such as through working hard, saving, donation from family, etc., improves one's housing situation.
Group Action	The comment mentions that group action, such as unions or protesting, may contribute to improvement in housing affordability.
Relocation	The comment mentions relocating to another area as a method to secure housing in general or more affordable housing.
Money	The comment mentions that simply making more money or spending more money will fix the problem.

Table 8: Labels for the "aspect – improve" category.

Label	Definition
Foreign Investment	The comment mentions ‘foreign investment’ or ‘foreign buyers’ as a cause contributing to concerns about housing affordability.
Underbuilding	The comment mentions under-building of housing as a cause contributing to concerns about housing affordability. Either in terms of quantity alone, or quantity of quality units.
Government Policy (E)	The comment mentions or implies that the ‘government’ (local or national) has not done enough/should do more to improve the housing situation, or has directly contributed to housing affordability concerns through policy.
NIMBYism	The comment mentions that private owners/nimbys blocking development or policies that would improve housing affordability contribute to concerns about housing affordability.
The Rich	The comment mentions that the wealthy, landlords, or realtors contribute to concerns about housing affordability. This may be attributed to greed (buying many properties or charging excessive rent/fees), manipulation (influencing government policy), etc.
COVID	The comment mentions or implies that COVID-19 and lockdowns had an impact on housing affordability.
The Old	The comment suggests that older property owners/boomers/the elderly in general have an impact on housing affordability.
Wage Price Imbalance	The comment mentions that wages in general are not sufficient to buy a home. May mention wage stagnation or inflation.
Cost of Living	The comment suggests that rising cost of living contributes to struggles with housing and affordability. May mention rising prices, inflation, rising mortgages and insurance rates, etc.
AirBnB	The comment mentions that short term vacation rentals, such as AirBnB, contribute to challenges with housing.
Overcrowding	The comment mentions that high population density, overcrowding, or immigration contribute to difficulty in securing housing. Specifically mentions density as a reason why more houses are needed.
Demand	The comment mentions that high demand due to location desirability contributes to high housing/rental prices.

Table 9: Labels for the “aspect - exacerbate” category.

B. Label Counts

This appendix presents the label counts per annotator for both the human-only and the LLM pre-annotation experiments.

Label	Annotator A		Annotator B		Annotator C	
	D_h^1	D_h^2	D_h^1	D_h^2	D_h^1	D_h^2
Real Price	130	169	159	208	68	184
Quality	97	137	49	117	24	138
Availability	52	80	64	109	29	123
Housing to Income	43	50	66	85	25	73
Building	27	46	26	51	12	54
Bootstraps	39	49	15	39	64	40
Government Policy (I)	46	55	34	47	46	86
Group Action	4	1	1	1	2	1
Relocation	18	46	19	62	26	61
Wage Price Imbalance	24	34	44	67	14	68
Underbuilding	15	19	16	38	18	40
Government Policy (E)	39	52	33	56	47	71
Cost of Living	44	40	30	34	17	67
Foreign Investment	39	21	35	24	35	23
NIMBYism	23	15	20	17	30	26
Overcrowding	13	14	18	20	10	17
The Rich	61	61	52	61	46	65

Table 10: Frequency of labels per annotator in the human-only annotations (D_h^1) and the LLM-assisted annotations (D_h^2).

Label	LLM	
	D_l^1	D_l^2
Real Price	176	193
Quality	109	128
Availability	132	129
Housing to Income	81	78
Building	57	56
Bootstraps	50	36
Government Policy (I)	102	92
Group Action	7	1
Relocation	36	57
Wage Price Imbalance	70	72
Underbuilding	38	44
Government Policy (E)	70	73
Cost of Living	66	67
Foreign Investment	34	22
NIMBYism	33	24
Overcrowding	17	17
The Rich	75	62

Table 11: Frequency of labels between the first half of the dataset (D^1) and second half of the dataset (D^2), as annotated by the LLM.