Groupe de lecture GDR TAL
8 novembre 2024

Léane Jourdan

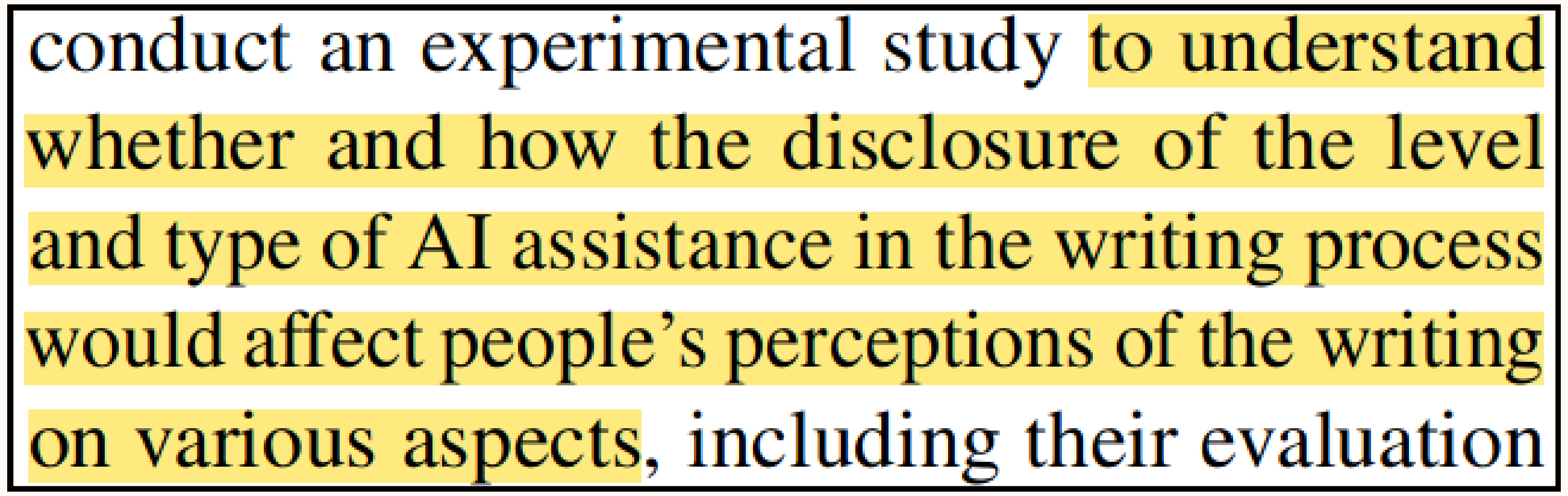
How Does the Disclosure of AI Assistance Affect the Perceptions of Writing?

Zhuoyan Li[1], Chen Liang[2], Jing Peng[2], Ming Yin[1]
[1] Purdue University
[2] University of Connecticut

TAL Traitement automatique des langues

Question générale

Est-ce que divulguer le fait qu'un assistant IA ait été utilisé modifie la perception du lecteur sur la qualité d'écriture ?



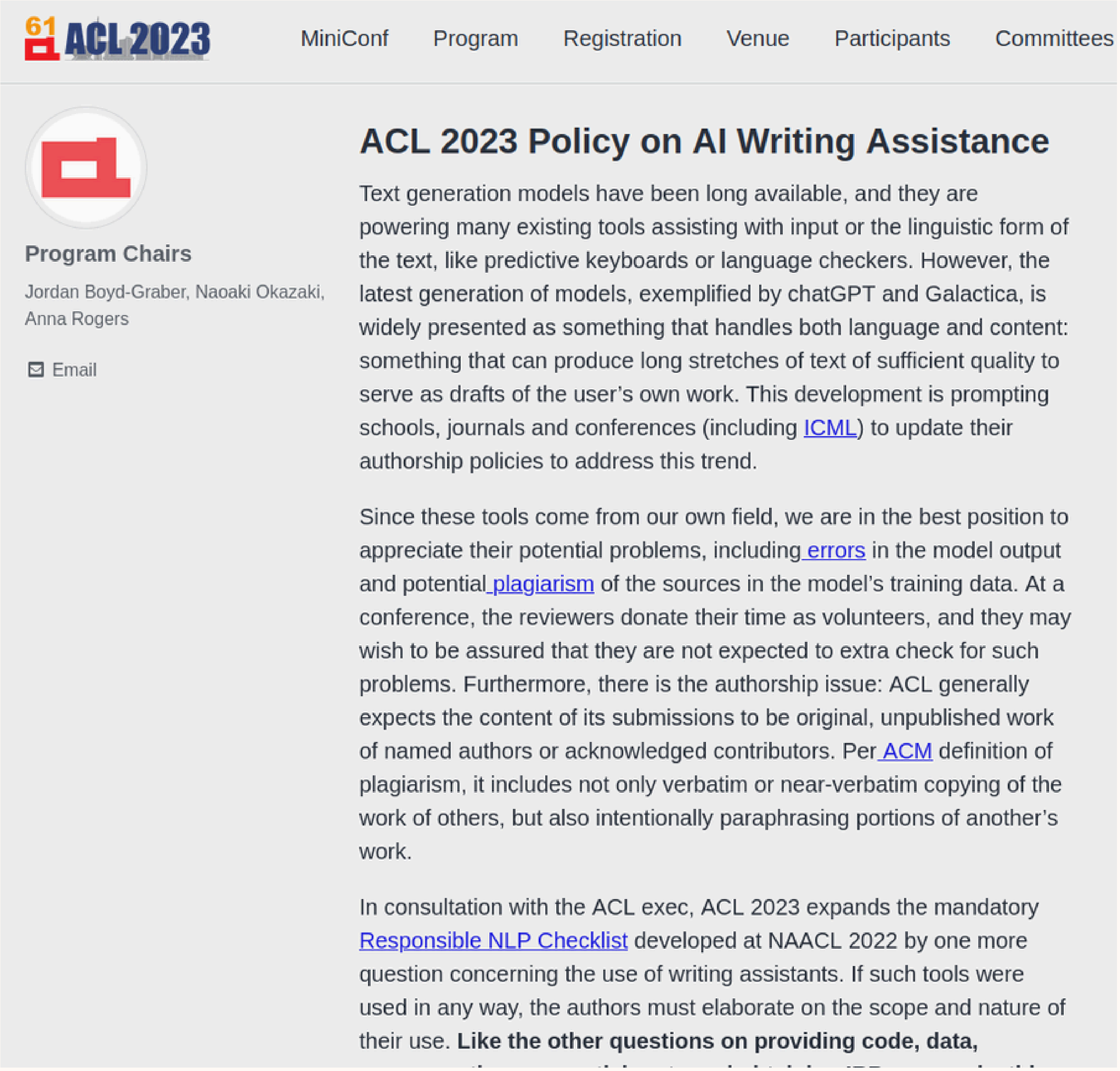
conduct an experimental study to understand whether and how the disclosure of the level and type of AI assistance in the writing process would affect people's perceptions of the writing on various aspects, including their evaluation

Motivations

Augmentation de l'utilisation des **LLM en assistant d'écriture** sur diverses tâches

Comment doit-on **présenter** ces textes ayant reçu l'assistance d'une IA **au lecteur** ?

Quel **impact** qu'aura cette information **sur le lecteur** ? Notamment sur sa **perception de la qualité de l'écriture** ?



Program Chairs

Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers

Email

ACL 2023 Policy on AI Writing Assistance

Text generation models have been long available, and they are powering many existing tools assisting with input or the linguistic form of the text, like predictive keyboards or language checkers. However, the latest generation of models, exemplified by chatGPT and Galactica, is widely presented as something that handles both language and content: something that can produce long stretches of text of sufficient quality to serve as drafts of the user's own work. This development is prompting schools, journals and conferences (including ICML) to update their authorship policies to address this trend.

Since these tools come from our own field, we are in the best position to appreciate their potential problems, including errors in the model output and potential plagiarism of the sources in the model's training data. At a conference, the reviewers donate their time as volunteers, and they may wish to be assured that they are not expected to extra check for such problems. Furthermore, there is the authorship issue: ACL generally expects the content of its submissions to be original, unpublished work of named authors or acknowledged contributors. Per ACM definition of plagiarism, it includes not only verbatim or near-verbatim copying of the work of others, but also intentionally paraphrasing portions of another's work.

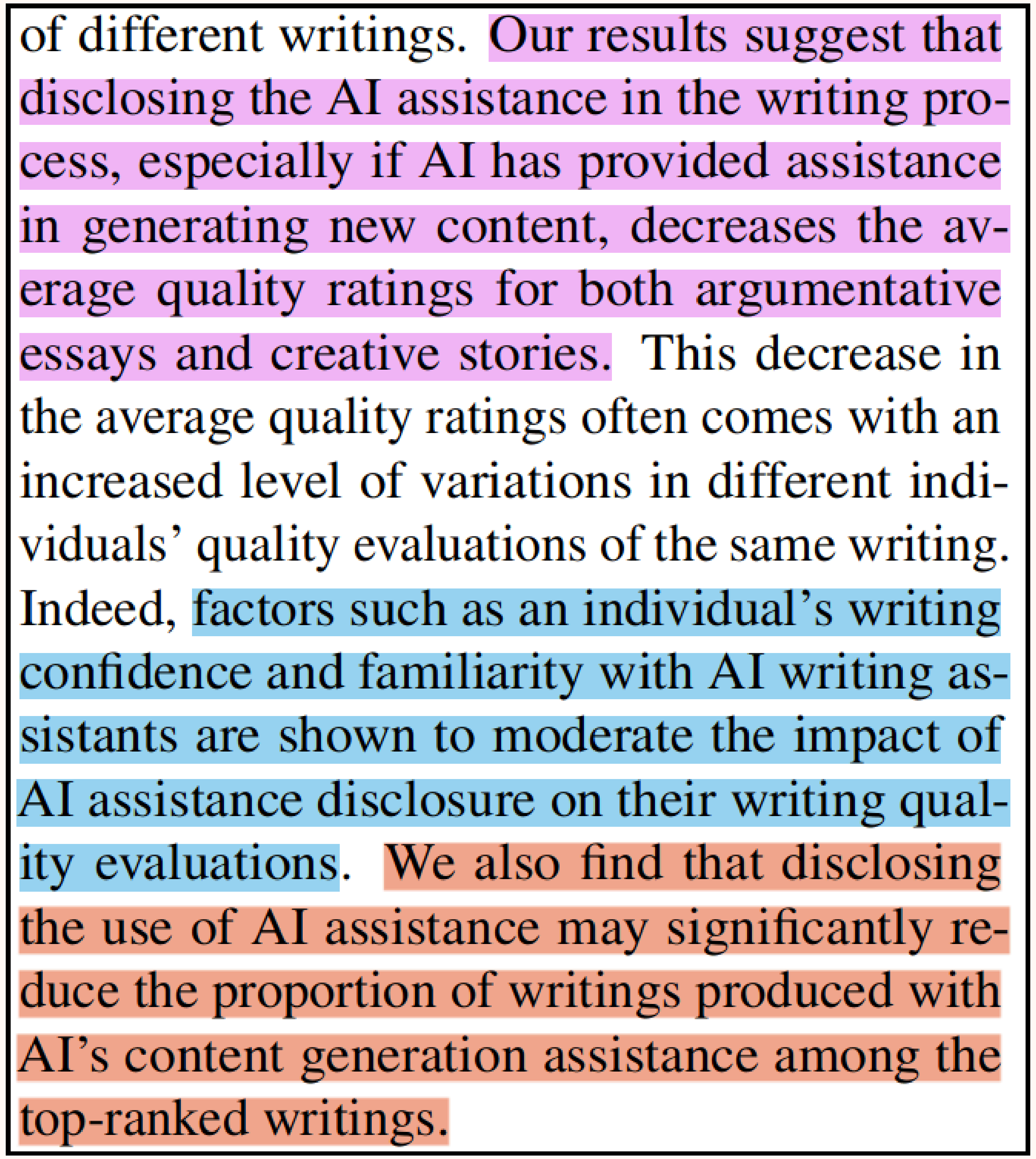
In consultation with the ACL exec, ACL 2023 expands the mandatory Responsible NLP Checklist developed at NAACL 2022 by one more question concerning the use of writing assistants. If such tools were used in any way, the authors must elaborate on the scope and nature of their use. **Like the other questions on providing code, data,**

Résultats clés

Divulguer le fait qu'il y ait eu une assistance par IA :

1. **diminue** significativement **les notes** sur qualité du texte
2. la note devient **dépendante de l'évaluateur**
3. **diminue** la part de **contenu avec IA dans le top k** des meilleurs textes

Ces effets **dépendent** en partie de la **confiance de l'évaluateur dans sa propre écriture** et sa **familiarité avec chatGPT**



of different writings. Our results suggest that disclosing the AI assistance in the writing process, especially if AI has provided assistance in generating new content, decreases the average quality ratings for both argumentative essays and creative stories. This decrease in the average quality ratings often comes with an increased level of variations in different individuals' quality evaluations of the same writing. Indeed, factors such as an individual's writing confidence and familiarity with AI writing assistants are shown to moderate the impact of AI assistance disclosure on their writing quality evaluations. We also find that disclosing the use of AI assistance may significantly reduce the proportion of writings produced with AI's content generation assistance among the top-ranked writings.

Phase 1 : Collecte des écrits

Collecte de textes avec différent degrés d'assistance IA (GPT) lors de l'écriture

Tâches d'écriture

200-250 mots

45 min

Argumentative essay

TOEFL writing exam

"Nowadays it is easier to maintain good health than it was in the past."

Creative story

Reedsy's Short Story Contest

Someone saying "Let's go for a walk."

Modes d'écriture

Independent

sans assistance

AI editing:

Brouillon: Humain

Revision: chatGPT

AI (content) generation:

Brouillon: chatGPT

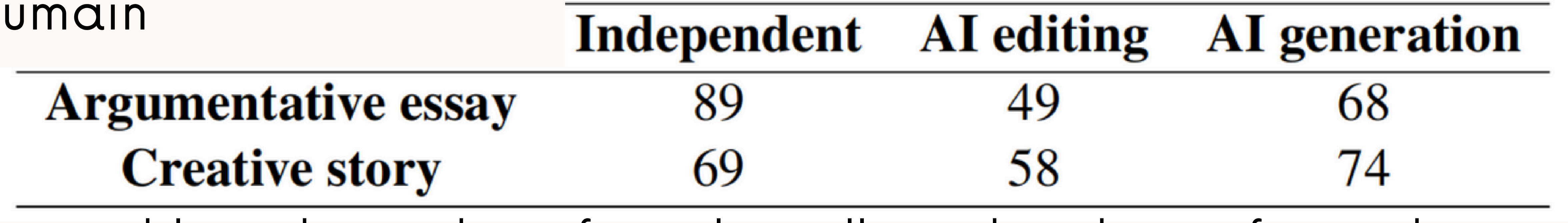
Revision: chatGPT sur consignes de l'humain

Procedure

Participants

"U.S. workers whose primary language is English"

1. "Independent vs. AI editing"
2. "Independent vs. AI generation."



	Independent	AI editing	AI generation
Argumentative essay	89	49	68
Creative story	69	58	74

Table 1: The number of articles collected in Phase 1 for each writing mode across the two types of writing tasks.

Phase 2: Evaluation de la perception des lecteurs

Participants différents de la phase 1
Chacun relit 6 textes

Groupes

1. **Non-Disclose**
2. **Disclose**

"the draft of this article was generated by ChatGPT, and the crowd worker had prompted ChatGPT to revise and improve the content"

Critères d'évaluation

- **Overall quality** : 1 à 5 paliers de 0.5
- **5 critères** :
  - Grammaire et vocabulaire
  - Organisation
  - Originalité
  - Créativité
  - Authenticité émotionnelle

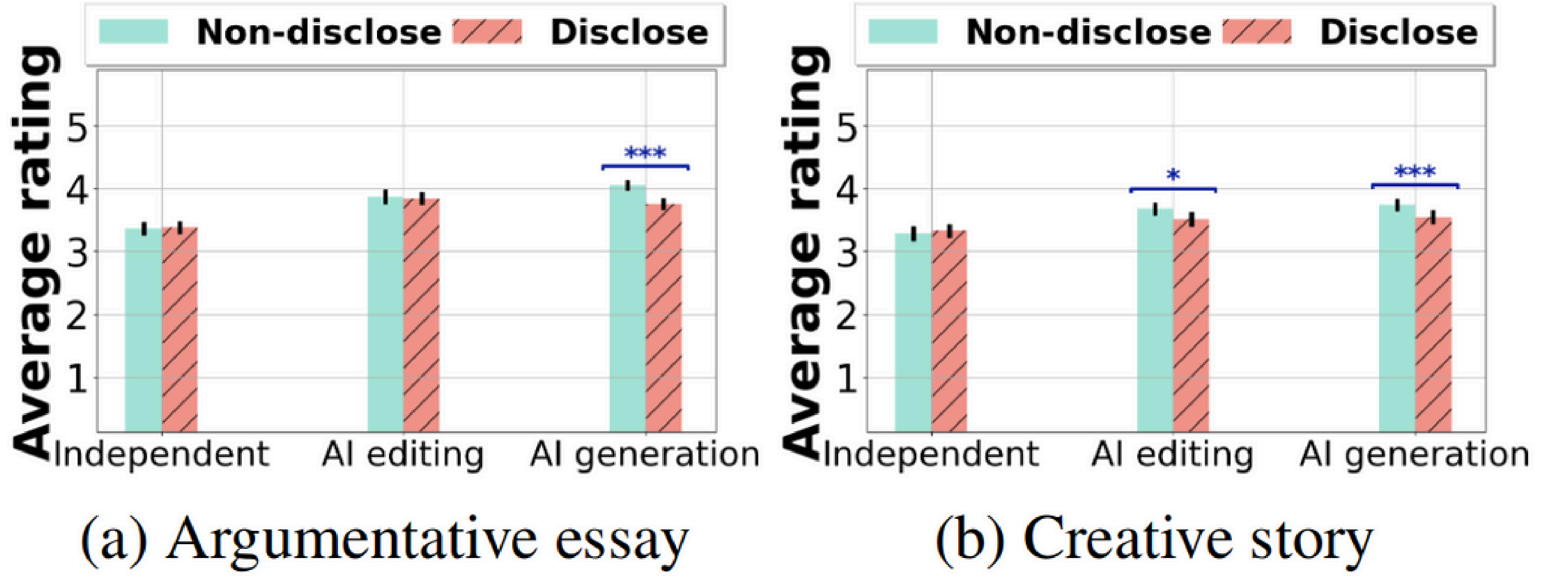
Questionnaire

- Confiance en leur compétence d'écriture
- Familiarité avec chatgpt

Résultats

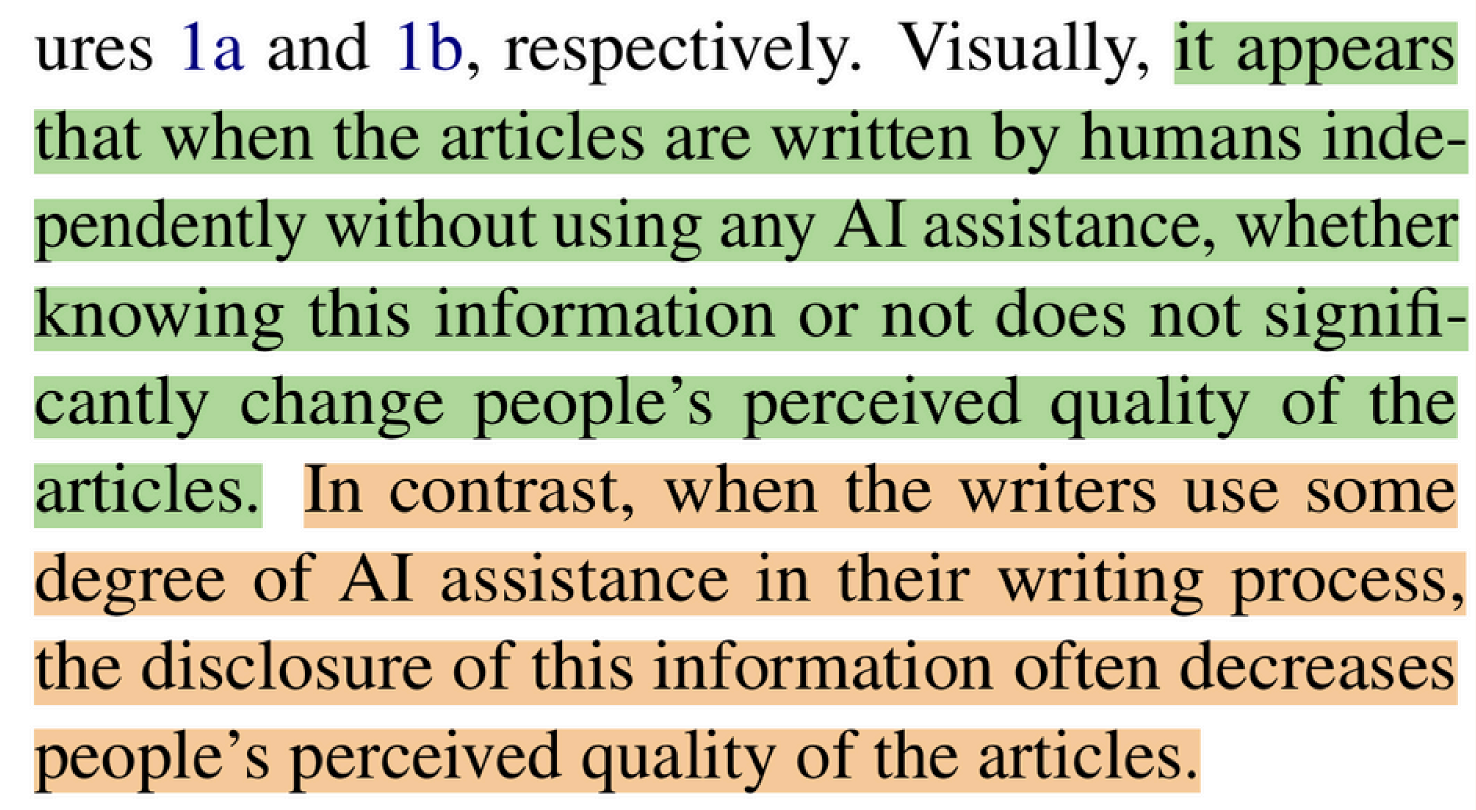
	Non-disclose	Disclose	Total
Participants	380	406	786
# évaluations / article	5.6	5.9	

Impact sur l’évaluation de la qualité d’écriture

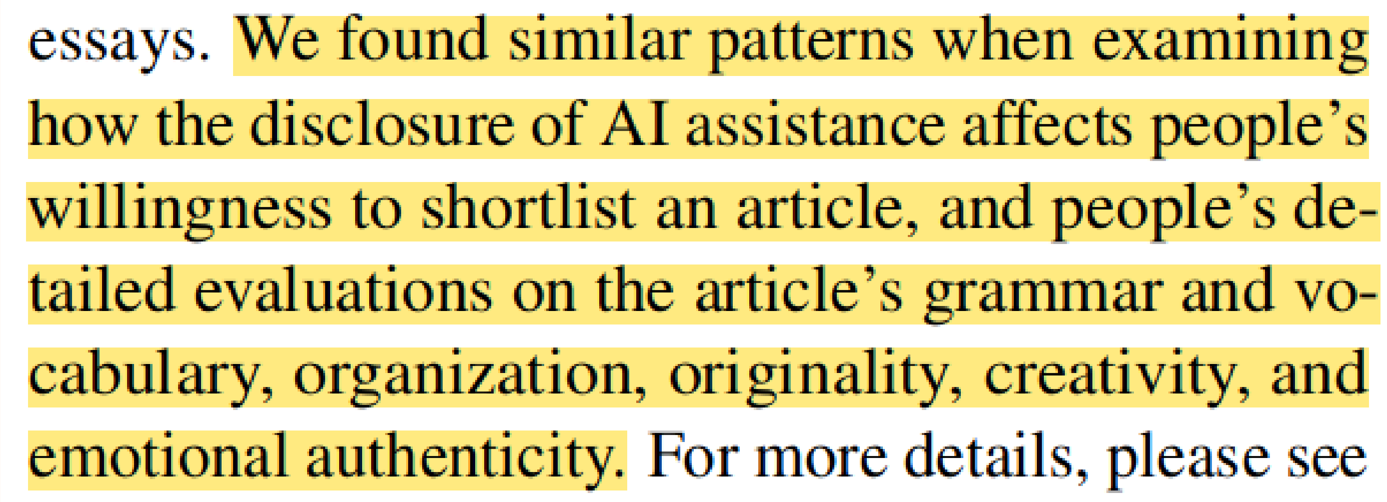


(a) Argumentative essay (b) Creative story

Figure 1: Comparing *average* ratings of the overall quality of articles generated under the independent, AI editing, or AI generation writing modes, with and without disclosure of the use and type of AI assistance during the writing process. Error bars represent the 95% confidence intervals of the mean values. * and *** denote significance levels of 0.05 and 0.001, respectively.



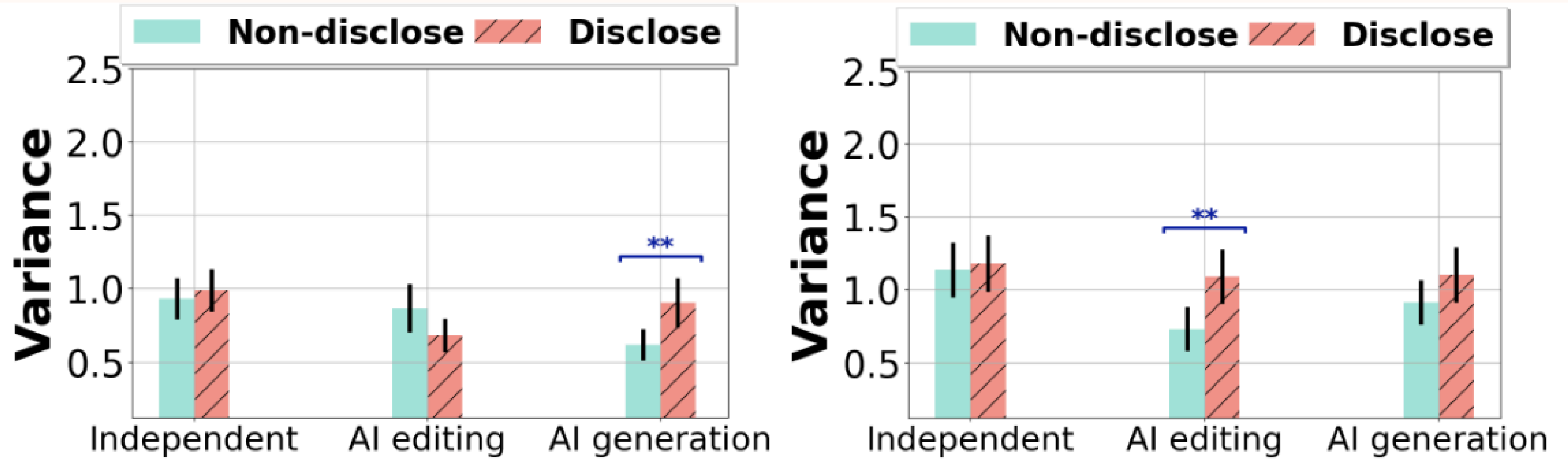
ures 1a and 1b, respectively. Visually, it appears that when the articles are written by humans independently without using any AI assistance, whether knowing this information or not does not significantly change people’s perceived quality of the articles. In contrast, when the writers use some degree of AI assistance in their writing process, the disclosure of this information often decreases people’s perceived quality of the articles.



essays. We found similar patterns when examining how the disclosure of AI assistance affects people’s willingness to shortlist an article, and people’s detailed evaluations on the article’s grammar and vocabulary, organization, originality, creativity, and emotional authenticity. For more details, please see

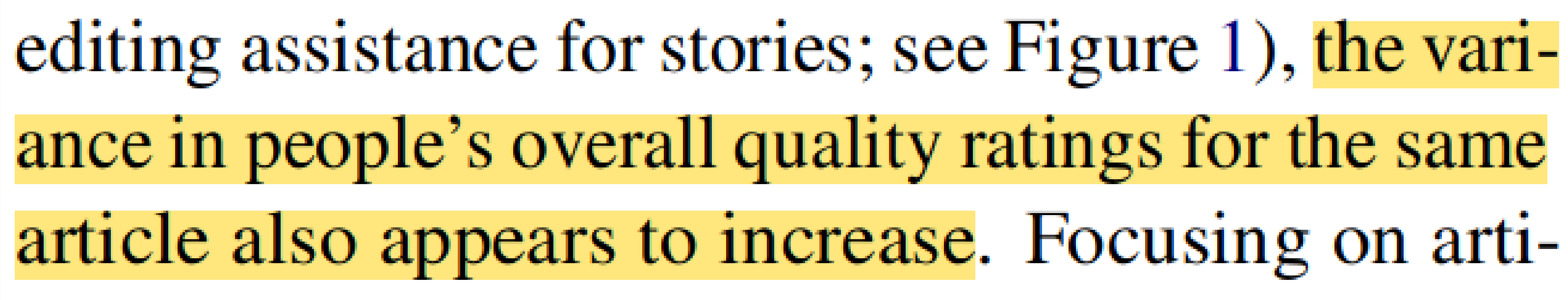
Résultats

Impacts sur la variation l'évaluation de la qualité

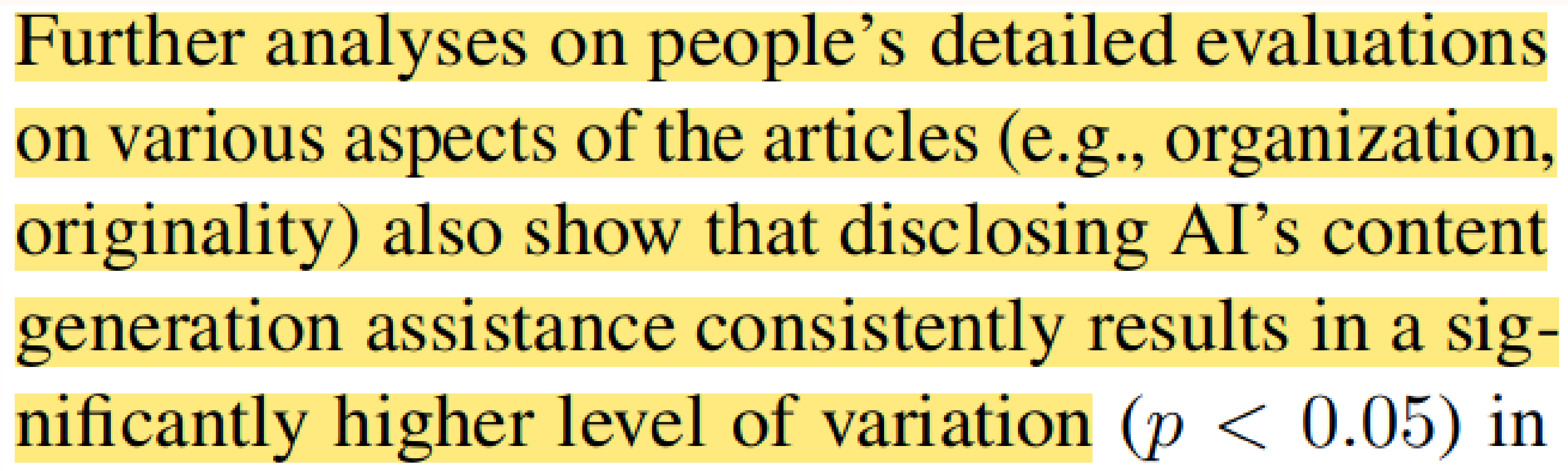


(a) Argumentative essay (b) Creative story

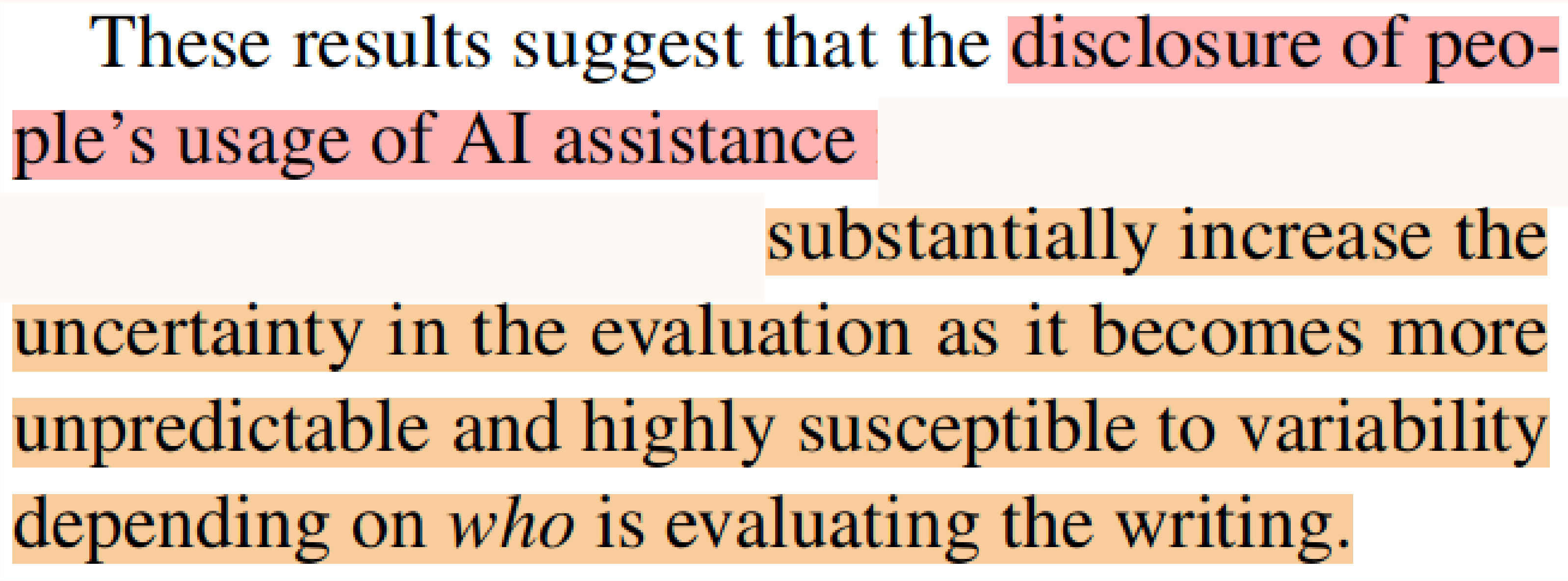
Figure 2: Comparing the *variance* in the overall quality ratings of articles generated under the independent, AI editing, or AI generation writing modes, with and without disclosure of the use and type of AI assistance during the writing process. Error bars represent the 95% confidence intervals of the variance. ** denotes the significance level of 0.01.



editing assistance for stories; see Figure 1), the variance in people's overall quality ratings for the same article also appears to increase. Focusing on arti-



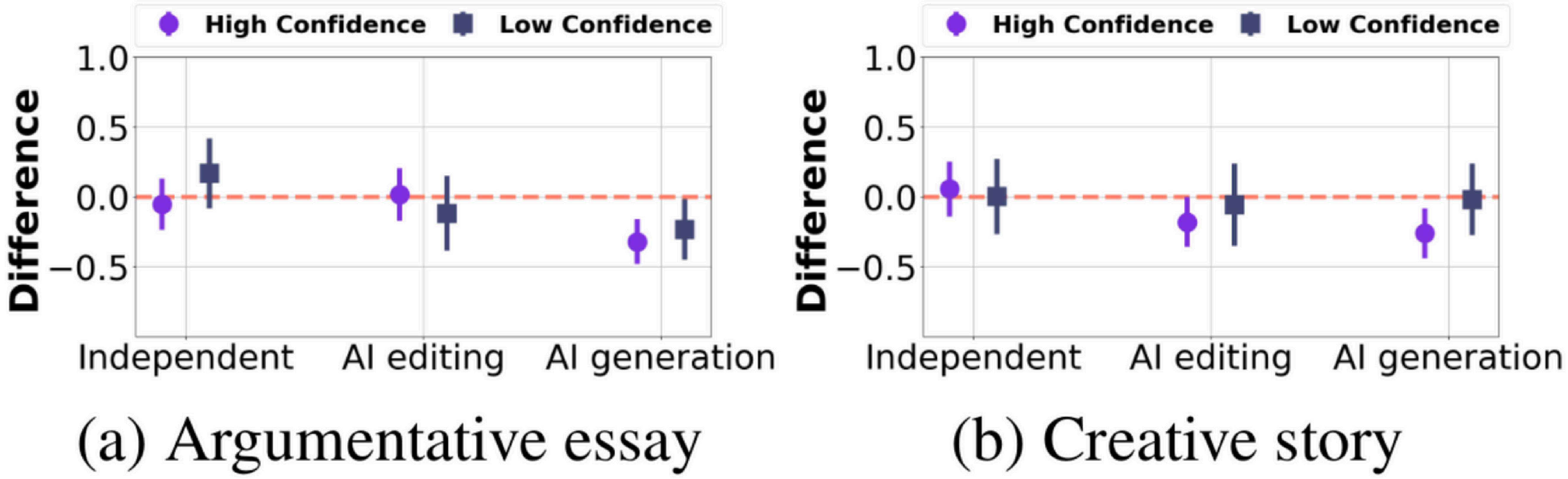
Further analyses on people's detailed evaluations on various aspects of the articles (e.g., organization, originality) also show that disclosing AI's content generation assistance consistently results in a significantly higher level of variation ($p < 0.05$) in



These results suggest that the disclosure of people's usage of AI assistance substantially increase the uncertainty in the evaluation as it becomes more unpredictable and highly susceptible to variability depending on *who* is evaluating the writing.

Résultats

Impact des différences individuelles sur l’évaluation de la qualité



(a) Argumentative essay

(b) Creative story

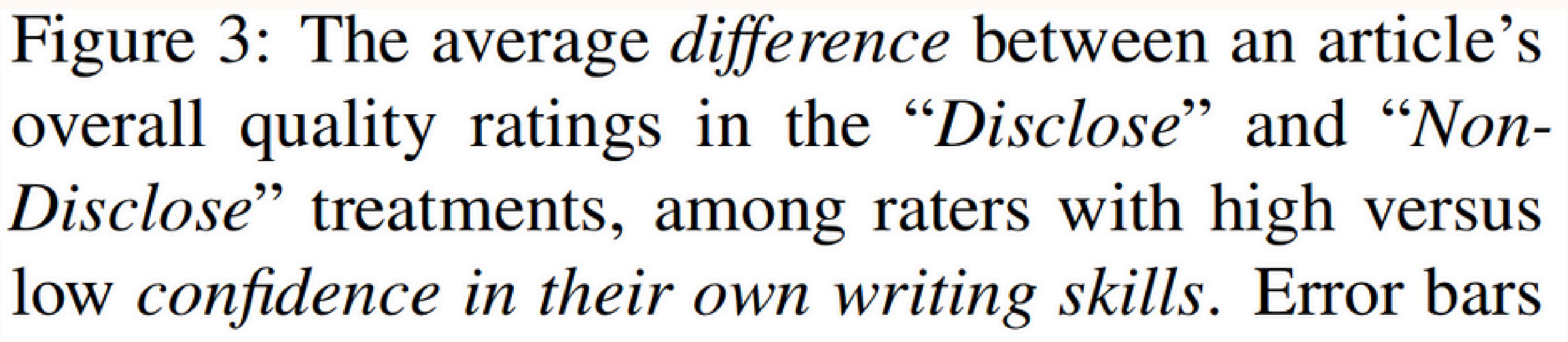
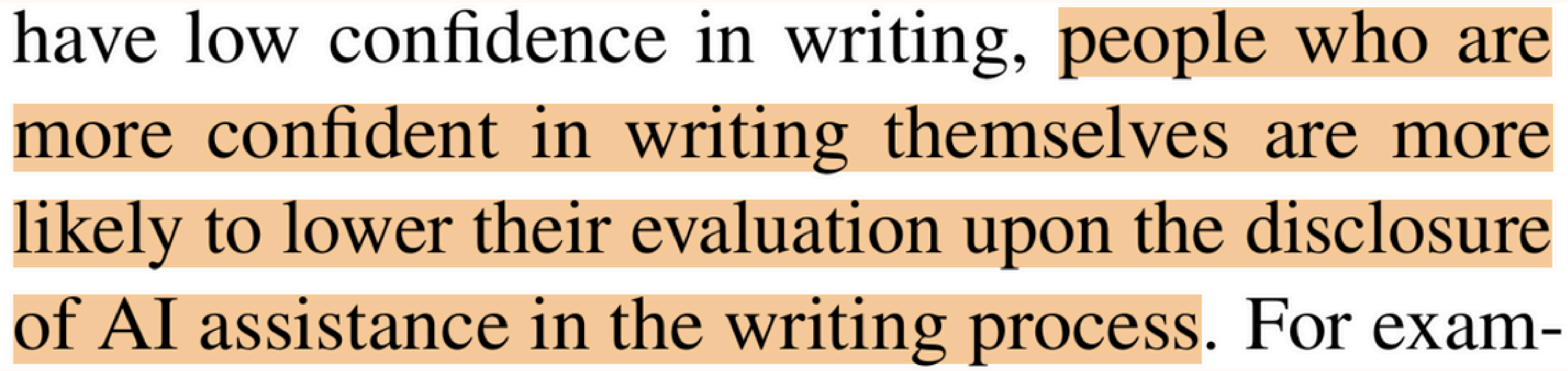
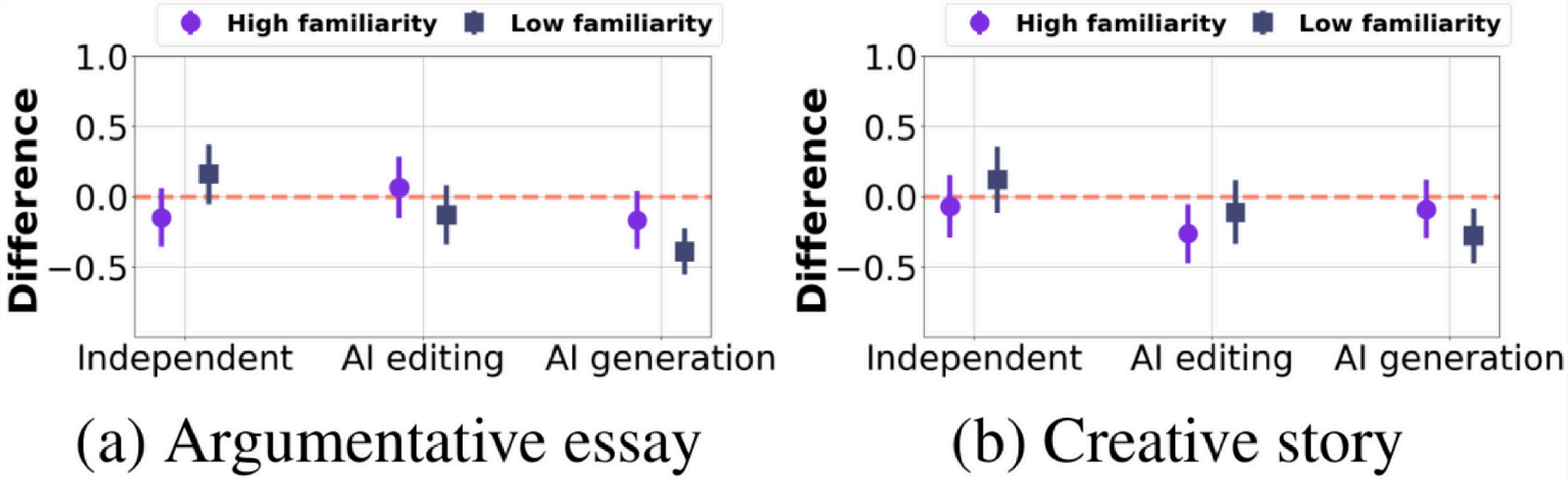


Figure 3: The average *difference* between an article’s overall quality ratings in the “*Disclose*” and “*Non-Disclose*” treatments, among raters with high versus low *confidence in their own writing skills*. Error bars


have low confidence in writing, people who are more confident in writing themselves are more likely to lower their evaluation upon the disclosure of AI assistance in the writing process. For exam-

Résultats

Impact des différences individuelles sur l’évaluation de la qualité



(a) Argumentative essay

(b) Creative story

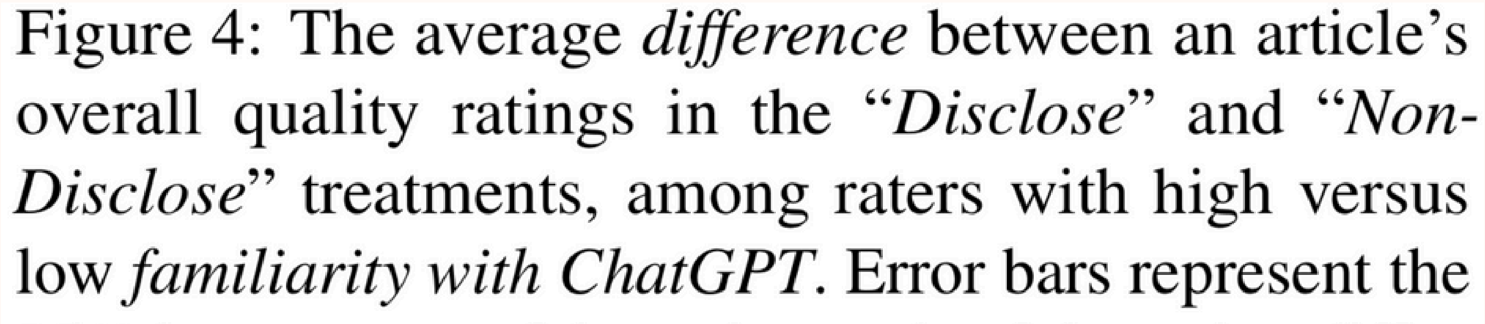
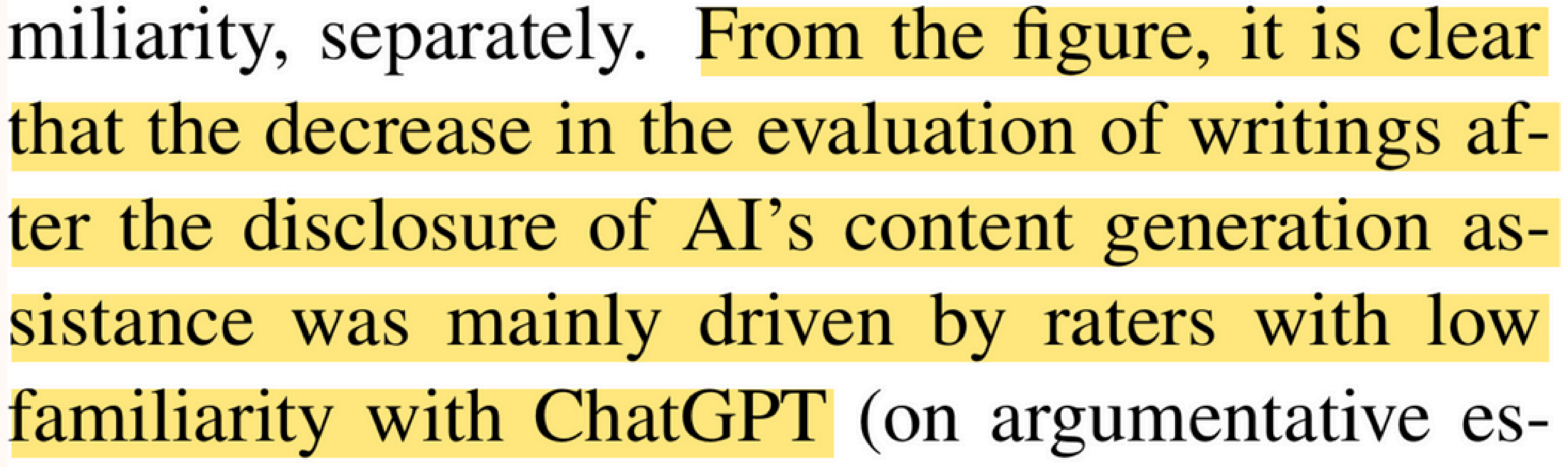
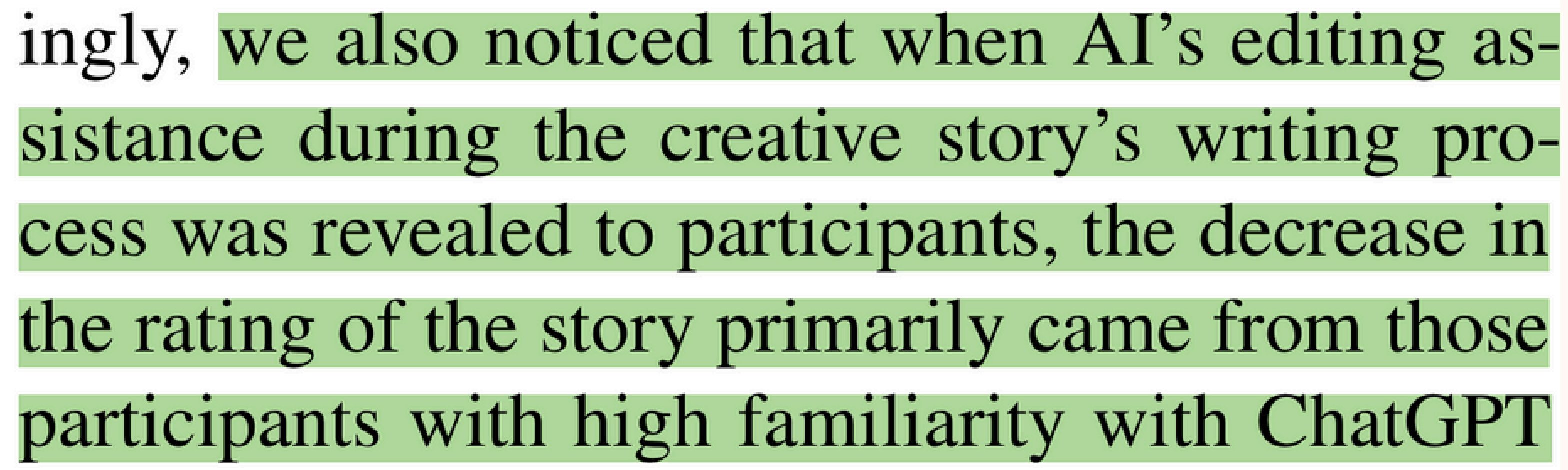


Figure 4: The average *difference* between an article’s overall quality ratings in the “*Disclose*” and “*Non-Disclose*” treatments, among raters with high versus low *familiarity with ChatGPT*. Error bars represent the



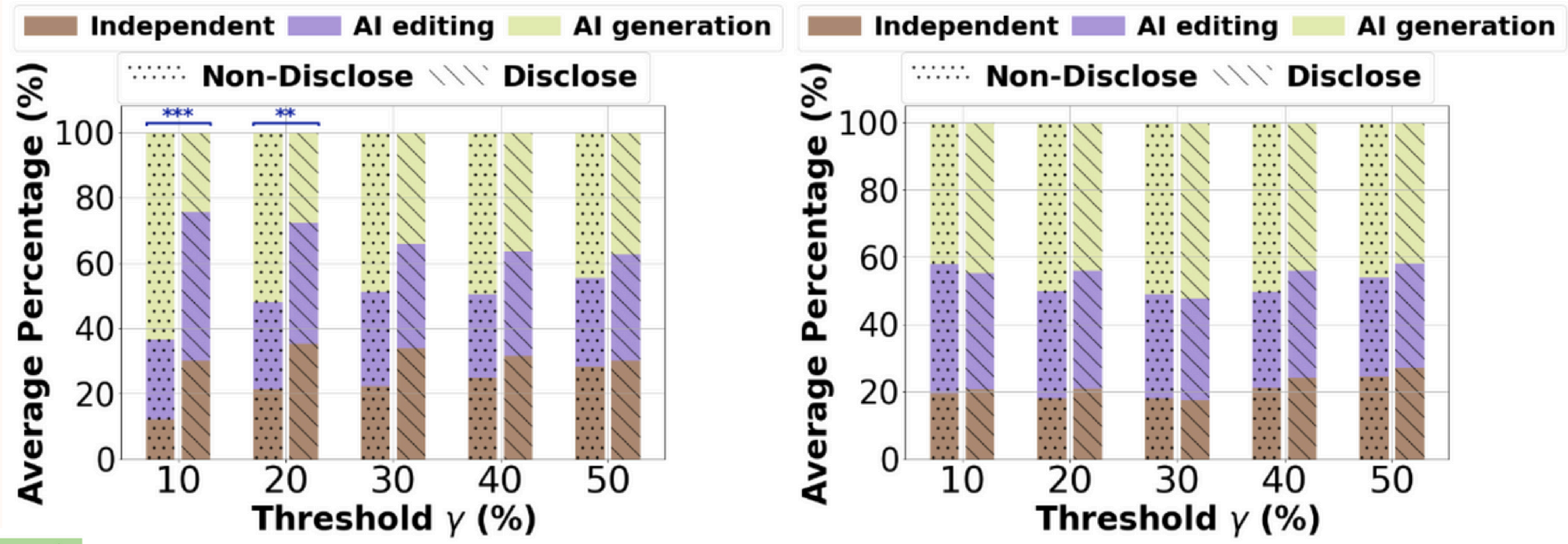
miliarity, separately. From the figure, it is clear that the decrease in the evaluation of writings after the disclosure of AI’s content generation assistance was mainly driven by raters with low familiarity with ChatGPT (on argumentative es-



ingly, we also noticed that when AI’s editing assistance during the creative story’s writing process was revealed to participants, the decrease in the rating of the story primarily came from those participants with high familiarity with ChatGPT

Résultats

Impact sur le classement des écrits



(a) Argumentative essay

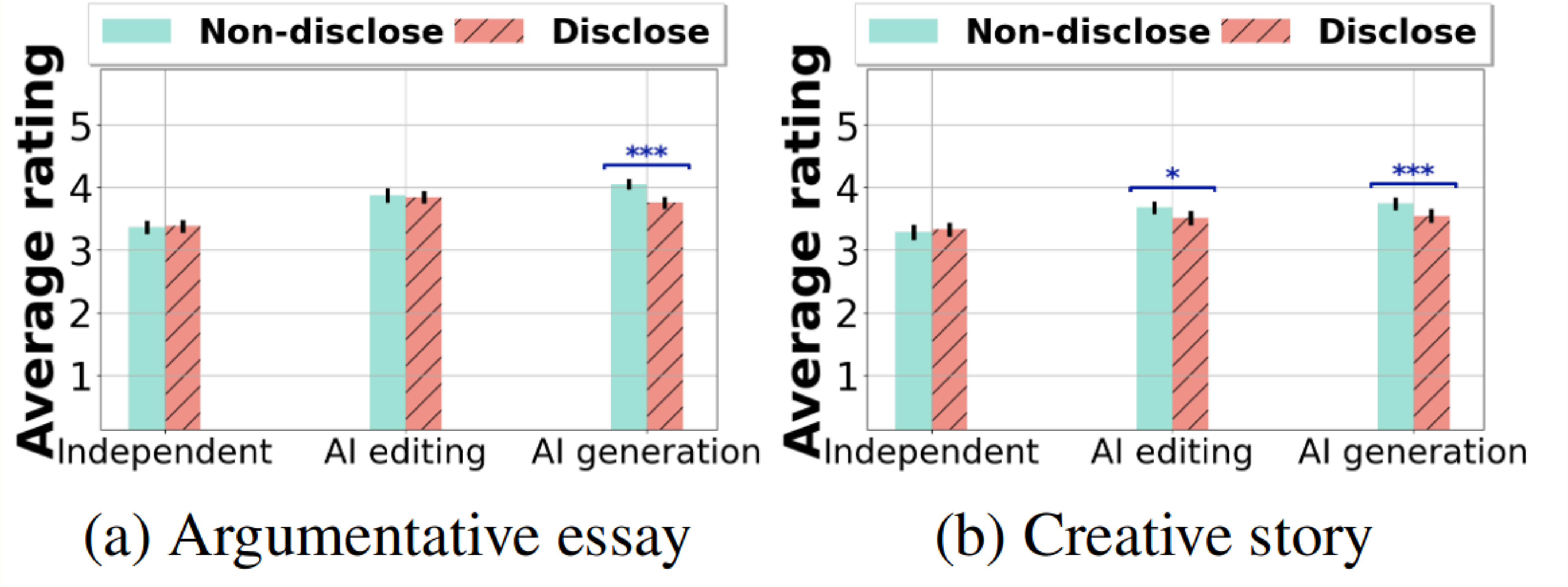
(b) Creative story

Figure 5: Within the top $\gamma\%$ of articles for the same writing task (ranked by articles' average overall quality ratings), the percentages of articles that were written in each of the three writing modes, with and without disclosing the use and type of AI assistance. ** and *** denote the significance level of 0.01 and 0.001, respec-

Suppose the articles' average overall quality ratings are used to determine their rankings. Given all the articles written on the same topic, we look into that within the top $\gamma\%$ of the articles for this topic, what proportions of the articles are written in the *independent*, *AI editing*, and *AI generation* writing modes, respectively, and how these proportions change after we informed participants about the use and type of AI assistance during the writing process of these articles.

Discussion

- Hypothèse : L'**effort humain** mis dans l'écriture du texte
- "we found that writers under the AI generation writing mode spend **significantly less time** than writers under other modes"
- Différence **accentuée** dans le cas de **génération**, hypothèse : Le crédit est donné à l'assistant IA
- Hypothèse : Les **attentes sont plus élevées**



(a) Argumentative essay (b) Creative story

"even after the AI assistance is disclosed, **participants who used the AI's content generation assistance may still receive a higher average rating** of their writing quality than participants who wrote independently."

"Phase 1 study (i.e., participants self-selected into their preferred writing mode), a comparison across the three writing modes **does not allow for a causal interpretation**."