Can I Use Wood Instead? Using an LLM to Compile Substitution Suggestions from Online DIY Tutorial Comments

Marcel Lahaye
RWTH Aachen University
Aachen, Germany
lahaye@cs.rwth-aachen.de

Filiz Günal
RWTH Aachen University
Aachen, Germany
filiz.guenal@rwth-aachen.de

Adrian Wagner
RWTH Aachen University
Aachen, Germany
wagner@cs.rwth-aachen.de

Sarah Sahabi
RWTH Aachen University
Aachen, Germany
sahabi@cs.rwth-aachen.de

Ana Maria Nitulescu
RWTH Aachen University
Aachen, Germany
ana.nitulescu@rwth-aachen.de

Jan Borchers
RWTH Aachen University
Aachen, Germany
borchers@cs.rwth-aachen.de

Abstract

Makers regularly discuss substitution suggestions for materials, tools, and practices around DIY projects online. However, these suggestions are often lost in the comment sections of projects, which can lead to mistakes being repeated and slower iterative project improvements. To address this, we propose utilizing large language models to identify, collect, and structure substitution suggestions in users' comments. We prototyped such a workflow using OpenAI's GPT-4o model. To evaluate its performance, we labeled 4193 comments regarding whether they contain a substitution and what it is about. The workflow successfully identifies substitution suggestions, including the substitute, original, username, and comment ID, from DIY tutorials on Instructables and YouTube and outputs them in a JSON format for further processing. We report the quantitative performance metrics F_1, ROUGE-L, and BERTScore, qualitative insights into limitations and benefits, and solutions to reduce the adverse side effects of generative variability. The collected data is provided as supplements.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

DIY Substitutions, Large Language Models, Maker Culture, DIY Tutorials

ACM Reference Format:
Marcel Lahaye, Filiz Günal, Adrian Wagner, Sarah Sahabi, Ana Maria Nitulescu, and Jan Borchers. 2025. Can I Use Wood Instead? Using an LLM to Compile Substitution Suggestions from Online DIY Tutorial Comments. In *ACM Symposium on Computational Fabrication (SCF Adjunct '25), November 20–21, 2025, Cambridge, MA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3774746.3779253

SCF Adjunct '25, Cambridge, MA, USA

ACM ISBN 979-8-4007-2295-0/25/11
https://doi.org/10.1145/3774746.3779253

1 Introduction

Makers regularly exchange substitution suggestions of materials, tools, and practices [Oehlberg et al. 2015; Tseng and Resnick 2014]. However, apart from project-specific related work [Fernando and Kuznetsov 2020; Smyth and Helgason 2019] and platforms like Thingiverse[1] [Oehlberg et al. 2015], substitution suggestions are mostly limited to the comment section and easily lost among other comments [Kuznetsov and Paulos 2010; Tseng and Resnick 2014]. This can stifle project-related exchange and discourage hobbyists [Tseng and Resnick 2014].

To address this, we propose automatically identifying and compiling substitution suggestions in online DIY tutorial comments.

Natural language of comments suggests utilizing the promising natural language processing capabilities [Min et al. 2023] of large language models (LLMs), similar to other tutorial domain work [Chen et al. 2024; Croitoru et al. 2023]. To evaluate the feasibility and utility of this approach, we prototyped a pipeline using GPT-4o[2] to (1) collect tutorial and comment data, (2) identify substitution suggestions in the comments, and (3) extract the related substitute, original, author, and comment ID in JSON format.

We used this prototype to investigate the following two research questions as outlined in the methodology section (Section 2):

RQ1: Do LLMs provide a feasible way to identify substitution suggestions in online DIY tutorial comments?
RQ2: Do LLMs provide a feasible way to organize a substitution suggestion into a specified format, including the original, the substitute, and a reference to the underlying online DIY tutorial comment?

Additionally, we conducted a qualitative analysis to investigate potential influences of comment phrasing on the output performance.

The average performance metrics over all 24 tutorials with 4193 comments are promising, but vary with inconsistent results.

The detailed result tables, full prompts, labeled data, a workflow figure, and a summary video can be found in the supplementary materials.

[1] https://www.thingiverse.com (accessed, October 9th, 2025)
[2] https://platform.openai.com/docs/models/gpt-4o, (accessed, October 9th, 2025)

2 Methodology

2.1 Data Collection

We selected 24 tutorials (15 from Instructables, 9 from YouTube), with 4193 comments in total, with a range of materials, tools, and number of comments. We collected all user comments with their respective platform comment ID and author username to retain a reference to the comment and author.

The GPT-4o model does not accept video input. Therefore, to create the YouTube tutorial context, we utilized the shot boundary detector Transnet V2 [Soucek and Lokoc 2020] to identify transition frames between crafting steps and used the in-between middle frames representing a step as input images (adopted from [Chen et al. 2024]). If available, transcripts or captions[3] are used as text input.

For the Instructables tutorials, we use the tutorial text and images. The context is provided to enable the relation and inference of information.

2.2 Labeling

We created ground-truth data by hand-labeling the comments on substitution suggestion (*true* or *false*), replaced object or process (original), and substitute.

We define a substitution suggestion as follows:

> A substitution suggestion in the maker domain is a suggestion to replace, remove, or add a material, tool, or process in a project to change or improve at least one aspect of the project while keeping the overall result equal or similar.

We labeled substitution suggestions as *false* where the author appeared unsure about the substitution feasibility. However, other users can validate the suggestion, resulting in a *true* label for the validation. In cases of missing mention of the replaced object, we left the replaced object (original) label empty. Multiple substitutions in a comment were considered individually by duplicating the comment and applying individual labeling. Finally, when multiple originals were replaced, we listed everything that was replaced in the "original" label.

2.3 Prompting

We refined our prompts by reviewing different prompt engineering techniques [Ramesh et al. 2021; Schulhoff et al. 2025]. This results in an initial 0-shot prompt based on preliminary testing and a second few-shot prompt that utilizes the insights from our analysis (Section 3).

The combination of tutorial context and comments often exceeds the 128,000 tokens limit of the GPT-4o model. Therefore, data was sequenced between multiple API calls, saving the response for the final prompt. In the initial prompts, we ask the model to analyze the images and the instructions. Then, we prompt the model to identify the substitution suggestions in the comments, extract the requested information, and provide it in a JSON format.

Hallucinated substitutions, verbose explanations, or lists of materials and tools in the output have been prevented by prompting to exclude everything except the JSON array.

[3]https://support.google.com/youtube/topic/9257536 (accessed, October 9th, 2025)

2.4 Calculating Performance Metrics

To calculate precision, recall, and F_1, we combined the model output with our labeled data by comment ID matching. Output that did not match the ID of an existing comment was considered a hallucination and appended at the end. The resulting tables were then used to identify true positives, true negatives, false positives, and false negatives.

To evaluate the content similarity between comment-based substitution and model-identified substitution, we calculated ROUGE-L precision [Lin 2004] with word-stemming enabled. Additionally, we used the BERTscore precision [Devlin et al. 2019] for semantic similarity comparison for follow-up work and our few-shot testing.

Finally, the correctness of model-provided original and substitute was determined by hand. The *Correct* label was applied for output that was tutorial-related and substitution definition matching (Subsection 2.2). A low score means that the model inferred information incorrectly or used parts of a comment unrelated to the tutorial project. Correct original and substitute in a marked hallucination were still counted as such.

2.5 Thematic Analysis

We investigated potential comment-related output influence patterns, acknowledging the LLM black-box-like nature [Liu et al. 2023]. Therefore, we applied an inductive reflexive thematic analysis [Braun and Clarke 2006] that supports identifying themes closely based on the collected data.

3 Results & Discussion

3.1 Quantitative Analysis

3.1.1 Hallucinations. Notably, most hallucinations (24) occurred for Tutorial 11, one with low comment count (34), and no substitution suggestions. Triple the number of hallucinations, compared to Tutorial 4, which has the second most hallucinations (8). Otherwise, hallucinations appear low, with no hallucinations in 13 of the 24 tutorials.

3.1.2 Correct Original & Substitute. Extraction correctness appears promising. Incorrect extraction occurred mostly for unrelated suggestions. In Tutorial 18, users recommended reducing the video speed to improve comprehension, which was model-identified as "video"(original) and "slower video mode" (substitute). Surprisingly, hallucinated substitutions were often correct because they were generated but usable substitution suggestions.

3.1.3 Precision. The precision score for the substitution appears low overall. For 15 of the 24 tutorials, at least half of the positive substitution identifications were *false positives*, meaning the model identified a comment as containing a substitution suggestion that we labeled as *false*. A lot of the *false positives* were either unvalidated questions or incorrect question-answer substitution referencing (Labeling Rules 2.2).

3.1.4 Recall. The results for the recall score are similar to the precision score. For 14 of the 24 tutorials, less than half of the existing substitution suggestions were identified.

Overall, low precision and recall score results in a low F_1 score. Our qualitative analysis attempts to identify patterns that result in this overall performance (Section 3.2).

3.1.5 ROUGE-L precision. The ROUGE-L precision content similarity metric appeared more promising. For our task, 0 means no words match, and 1 means that all words match. Scores of 0 were tutorials with only hallucinated suggestions, meaning the reference phrase, the comment, was empty. Otherwise, for 20 of the 24 tutorials, at least half of the words matched. However, incorrect words were often left-out information that was inferred from the model context.

3.2 Qualitative Analysis

We identified phrasing-related themes like *Tone* (suggesting, assertive, implicit, explicit), *Anecdotal Style*, *Level of Information*, and *Substitution-Related* and expected either weaker or stronger model-based association to substitutions, due to, for example, phrasing that appears often in substitution suggestions or information obfuscation. However, no consistent output performance-based patterns supported this.

Substitution suggestions in the *Social Cues* theme are either unrelated, sarcastic, or meant as a joke, which appear challenging in LLM identification tasks [Chang et al. 2024]. This was validated in our analysis with a recurring pattern of *false positive* scorings for such comments.

Missing comment information, like the replaced object, was repeatedly correctly inferred from the context. Incorrect inference happened mostly when the substitution added something without replacing anything.

4 Few-Shot Prompting

For output improvisation, we added nine few-shot input-output exemplar pairs [Schulhoff et al. 2025], based on our qualitative analysis themes. For each, we collected matching comments from Instructables and YouTube tutorials outside our dataset.

The average F_1 score improved from 0.48 to 0.58, but individual F_1 scores varied. The average ROUGE-L Precision score decreased from 0.63 to 0.2, and the average BERTscore Precision decreased from 0.82 to 0.75. After review, no eminent improvements or changes were apparent in the data.

5 Summary and Future Work

Confirming RQ1, using an LLM to identify and compile substitution suggestions from comments on online DIY tutorials appears limited but promising, addressing the potentially tedious process of processing comments manually. The approach can preserve the comment-based interaction while reducing the workload to engage with substitution suggestions from comments.

While performance metrics were low, indicating an inconsistent extraction reliability, only partially confirming RQ2, comment ID and author username can reliably be used to identify hallucinations. Overall, the process can identify such comments and automate a potentially tedious process, but further investigation is required to improve the output.

References

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (mar 2024), 45 pages. doi:10.1145/3641289

Yuexi Chen, Vlad I Morariu, Anh Truong, and Zhicheng Liu. 2024. TutoAI: a cross-domain framework for AI-assisted mixed-media tutorial creation on physical tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 161, 17 pages. doi:10.1145/3613904.3642443

Ioana Croitoru, Simion-Vlad Bogolin, Samuel Albanie, Yang Liu, Zhaowen Wang, Seunghyun Yoon, Franck Dernoncourt, Hailin Jin, and Trung Bui. 2023. Moment Detection in Long Tutorial Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. The Computer Vision Foundation, Paris, 2594–2604.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

Piyum Fernando and Stacey Kuznetsov. 2020. OScH in the Wild: Dissemination of Open Science Hardware and Implications for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376659

Stacey Kuznetsov and Eric Paulos. 2010. Rise of the Expert Amateur: DIY Projects, Communities, and Cultures. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI '10)*. Association for Computing Machinery, New York, NY, USA, 295–304. doi:10.1145/1868914.1868950

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. ACL, Barcelona, Spain, 74–81.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. doi:10.1145/3560815

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* 56, 2, Article 30 (sep 2023), 40 pages. doi:10.1145/3605943

Lora Oehlberg, Wesley Willett, and Wendy E. Mackay. 2015. Patterns of Physical Design Remixing in Online Maker Communities. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 639–648. doi:10.1145/2702123.2702175

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Cambridge, MA, USA, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv:2406.06608 [cs.CL] https://arxiv.org/abs/2406.06608

Michael Smyth and Ingi Helgason. 2019. DIY Community WiFi Networks: Insights on Participatory Design. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3313073

Tomás Soucek and Jakub Lokoc. 2020. TransNet V2: An effective deep network architecture for fast shot transition detection. *CoRR* abs/2008.04838 (2020), 1. arXiv:2008.04838 https://arxiv.org/abs/2008.04838

Tiffany Tseng and Mitchel Resnick. 2014. Product versus Process: Representing and Appropriating DIY Projects Online. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. Association for Computing Machinery, New York, NY, USA, 425–428. doi:10.1145/2598510.2598540