A 360º View for Large Language Models: Early Detection of Amblyopia in Children using Multi-View Eye Movement Recordings

Dipak P. Upadhyaya[1], Aasef G. Shaikh[2], Gokce Busra Cakir[3], Katrina Prantzalos[1], Pedram Golnari[1], Fatema F. Ghasia[3], and Satya S. Sahoo[1(✉)]

[1]Department of Population & Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

{dipak.upadhyaya, katrina.prantzalos, pedram.golnari, aasef.shaikh, satya.sahoo}@case.edu

[2]National VA Parkinson's Consortium Center, Louis Stokes Cleveland VA Medical Center Cleveland, OH, USA

[3]Visual Neuroscience Laboratory, Cole Eye Institute, Cleveland Clinic, Cleveland, OH, USA

{ghasiaf, cakrig}@ccf.org

Abstract. Amblyopia is a neurodevelopmental visual disorder that affects approximately 3-5% of children globally and it can lead to monocular vision loss if it is not diagnosed and treated early. Traditional diagnostic methods, which rely on subjective assessments and expert interpretation of eye movement recordings presents challenges in resource-limited eye care clinics. This study introduces a new approach that integrates the Gemini large language model (LLM) with eye-tracking data to develop a classification tool for diagnosis of patients with amblyopia. The study demonstrates that, (i) LLMs can be used to analyze fixation eye movement data to diagnose patients with amblyopia; and (ii) integration of medical subject matter expertise improves the performance of LLMs in medical applications. Our LLM-based classification tool achieves an accuracy of 80% in diagnosing patients with amblyopia using a combination of few shot learning and multi-view prompting with expert input from pediatric ophthalmologist. The model classifies amblyopic patients with moderate or severe amblyopia from control subjects with an accuracy of 83% and mild or treated amblyopic patients from control subjects with an accuracy of 81%. Finally, the model achieves an accuracy of 85% for classifying amblyopic patients with nystagmus from control subjects. The results of this study demonstrate the feasibility of using LLMs in ophthalmology application and highlights the essential role of medical expert input in LLM-based medical applications.

Keywords: Transfer learning, large language models, in-context learning, fixation eye movement, amblyopia diagnosis.

1 Introduction

Amblyopia is a neurodevelopmental disorder of vision affecting 3-5% of the children's population, that is, an estimated 99.2 million children are affected by amblyopia worldwide [1]. This condition leads to decreased vision in one eye due to improper visual stimulation during early childhood [1]. If left untreated, amblyopia can result in irreversible visual impairment and contributes to lifelong consequences for the affected individuals. In the United States, amblyopia is the most common cause of monocular vision loss among children, adolescents, and middle-aged adults, underscoring the substantial public health challenge it poses across age groups [1]. The diagnosis of amblyopia has traditionally relied on subjective visual acuity assessments and the interpretation of eye movement recordings, which requires specialized equipment and expert clinical judgment. These diagnostic challenges are exacerbated in resource-limited settings, where access to specialized pediatric ophthalmology services is often unavailable. The increasing use of high-fidelity eye tracking instruments, such as video oculography tracker (EyeLink1000 plus), has enabled the use of eye tracking data for the detection of amblyopia. In this study, we leverage the powerful transfer learning capabilities of the Gemini large language model (LLM) (Google, Inc.) to analyze eye tracking data in a cohort of 135 subjects to develop a high accuracy classification tool that can be deployed in remote, resource scarce eye care clinics.

LLMs have demonstrated state of the art performance in a variety of applications; however, their performance in medical applications have highlighted multiple challenges due to the complexity of the medical domain [2-5]. Specifically, customized LLMs such as Med-PaLM, which extends the Pathway Language Model (PaLM) [6], require the development of fine-tuned prompting methods to answer medical domain questions (e.g., questions in the MultiMedQA multiple-choice benchmark dataset). Gemini is a state-of-the-art language model that leverages transfer learning together with attention-based transformer deep learning architecture to achieve high accuracy in a range of natural language processing tasks [7, 8]. However, our initial approach to analyze eye tracking data represented as cartesian plots resulted in low accuracy with multiple errors. When using single eye images per subject without clinical contextual information, both mild or previously treated amblyopia and control subjects were misclassified. Further, amblyopia associated with nystagmus was also inaccurately classified, with binary classification accuracy barely exceeding random chance of 50%.

To address these limitations, we developed a new multi-view imaging prompt framework for ophthalmology applications that uses eye tracking data under multiple viewing conditions together with medical expert input as part of ophthalmology focused prompt tuning to enhance the performance of LLMs [7]. Using this new prompting framework this study demonstrates:

1. The feasibility of using LLMs together with medical expertise to classify patients with amblyopia using eye tracking data.
2. The effectiveness of increased contextual information, represented as multi-view fixation eye movement plots in this study, in prompt tuning for ophthalmology applications.

In contrast to fine tuning LLMs, which requires retraining the LLM that leads to modified gradient weights, prompt tuning is a low-cost approach to improve the accuracy of results generated by LLMs with significance to the broader domain of medical applications [2]. Further, the use of detailed instructions created by medical experts (pediatric ophthalmologist in this study) enables the LLM to successfully interpret the medical query through natural language understanding and perform accurate classification through analysis of the plots representing eye tracking data.

1.1 Background

Eye movement abnormalities are being increasingly used as diagnostic markers in amblyopia research [9] to non-invasively capture eye movements and analyze the recordings to arrive at a diagnosis. Individuals with amblyopia demonstrate notable fixation instability, particularly in the amblyopic eye as compared to its fellow eye (Figure 1), which highlights the close link between fixation eye movement (FEM) abnormalities and the sensory deficits in amblyopia [10, 11]. The use of modern eye trackers provides an objective dataset crucial for assessing visual function and understanding the nuance of impairments in eye movement control associated with the disorder [12]. In parallel, the development of LLMs with capabilities to analyze multimodal data including text and images, presents new avenues for ophthalmology applications.

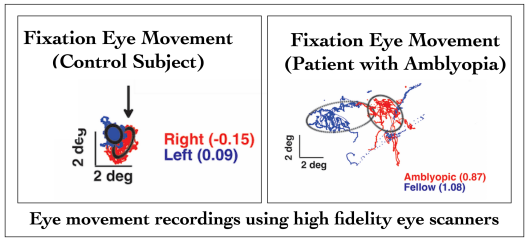


Figure 1: The visualization of eye movement recordings to distinguish children with amblyopia from control subjects with dense, focused saccades seen in control subjects (left); and dispersed, wide-ranging saccades seen in amblyopic patient (right)

1.2. Related Work

The application of machine learning in ophthalmology, particularly for amblyopia detection, has evolved over time [13]. Convolutional neural networks (CNNs) have been employed to analyze retinal images for amblyogenic risk factors and they have shown greater accuracy as compared to traditional methods [14]. Other machine learning methods have been explored for classification of visual field defects and assessing binocular vision, which is important for amblyopia diagnosis [15]. However, traditional approaches require extensive training data and their effectiveness have been affected due to the difficulty in identifying subtle features associated with amblyopia [16]. We have addressed some of these challenges in this study by leveraging the powerful capabilities of LLMs including transfer learning capabilities that rely on few-shot based in-context learning for generating accurate results in a specific domain. In-context learning refers to the ability of language models to adapt to new tasks or domains without explicit retraining of the LLM through the use of one or more examples [8], thereby

helping alleviate the challenge of inadequate training data in the medical domain application such as the diagnosis of amblyopia [8, 17].

2. Method

Figure 2 gives an overview of the method used in this study that consists of a multi-stage workflow, including: (i) acquisition of eye movement data using high fidelity eye tracking instrument; (ii) signal pre-processing methods to enhance data quality by reducing noise in the recordings; (iii) visualization of eye tracking recordings with appropriate scaling measures; and (iv) the use of multi-view prompting with input from pediatric ophthalmologist.

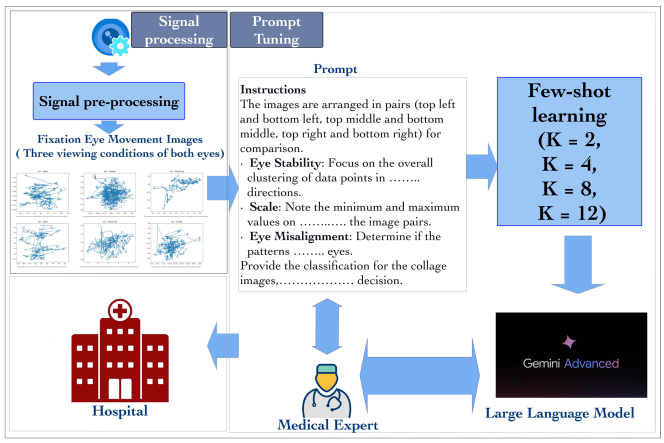


Figure 2: An overview of this study that introduced a new Multiview imaging LLM prompting techniques for analyzing pediatric ophthalmology data.

2.1 Study Cohort and Data Collection.

This study uses eye movement recordings from 135 participants collected at the Cleveland Clinic Eye Institute (the study was reviewed and approved by the Cleveland Clinic Institutional Review Board). The study cohort consisted of 95 patients with amblyopia and 40 control subjects. The age of the participants ranged between 3 years and 72 years. Control subjects were selected based on the absence of ocular or systemic abnormalities affecting visual acuity, except for refractive errors. The 95 patients with amblyopia included 51 females and 44 males. The 40 control subjects included 23 females and 17 males. The control group consisted of 24 White Caucasian, 6 African American, 1 Hispanic, 5 Asians, and 4 multi-cultural patients. The amblyopic subjects included 61 Whites Caucasians, 12 African American, 10 Hispanics, 3 Asians, and 9 multi-cultural patients.

The EyeLink1000 plus video-based eye tracker was used with a sampling rate of 500 Hz with participants wearing corrective lenses. In a dimly lit room, participants faced an LCD screen displaying various stimuli, with calibration involving a cartoon animal accompanied by sound cues to focus attention. Calibration adhered to a 5-point manufacturer guideline. The fixation measurements were taken as subjects focused on a white dot against a black screen, employing animal sounds to redirect attention if

needed. Each session was recorded with both eyes open (labeled as binocular vision, BV) or with left/right eye viewing while the other was covered (labeled as monocular viewing, MV), in a sequence determined randomly. The MV sessions were recorded from the active viewing eye, distinguishing between fellow eye (FE) and amblyopic eye viewing (AE) conditions. The covered eye was monitored using an infrared filter to avoid exposure to visible light.

2.2 Visualization of Eye Tracking Data

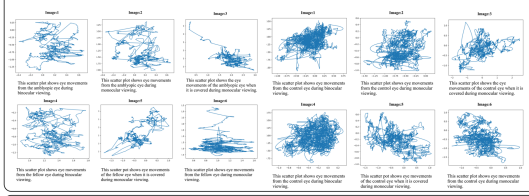
The raw eye-tracking data was plotted in a cartesian space to generate visualization of eye movements using the Pandas library for data manipulation and the Python Imaging Library (PIL). In the next step, we annotated these eye movement plots with text prepared by pediatric ophthalmologist to describe the viewing conditions (monocular or binocular, viewing or non-viewing). As discussed in the previous section, an initial approach that used a single view image, for example one image from an amblyopic eye and another from control subject resulted in low accuracy and errors. Therefore, we combined the eye movement plots corresponding to multiple viewing conditions for a patient to create a collage consisting of six images. Each collage consisted of three distinct viewing conditions: (i) monocular viewing (MV) with the amblyopic eye, (ii) monocular viewing with the fellow eye (MV), and (iii) binocular viewing (BV). The collage of eye movement plots

Context
The input images represent fixation eye movements used to diagnose amblyopia, a neurodevelopment disorder in ophthalmology.

Instructions
The images are arranged in pairs (top left and bottom left, top middle and bottom middle, top right and bottom right) for comparison.
- **Eye Stability**: Focus on the overall clustering of data points …… directions.
- **Scale**: Note the minimum and maximum values on…… the image pairs.
- **Eye Misalignment**: Determine if the patterns…….eyes.

Provide the classification for the collage images, ……. decision.

Examples as few shot learning (K= z)
Patient-X: Label: Amblyopia **Patient-Y: Label: Control**



Task
As a helpful assistant, analyze the eye movement recording images from a patient. Given a set of collage images, can you classify each image as 'Amblyopia' or 'Control'?

Image classification task
Output
Patient1: Classification-Abnormal, **Confidence-** 95%
Patient2: Classification-Normal, **Confidence-** 80%
Explanations
- The Patient 1 collage shows data from an amblyopic eye. The image of the …….. spread out.
- The Patient 2 collage exhibits a …… tightly clustered.

Figure 3: Overview of prompting template developed with medical expert feedback.

together with textual descriptions generate by clinical experts was used to construct the prompts for the LLM.

2.3 In-context Learning for Eye Movement Data

LLMs have demonstrated a remarkable capability to learn specific contexts using few examples (few-shot or in-context learning) that was highlighted in the Generative Pre-trained Transformer (GPT 3.5) model and many other subsequent studies [2, 6, 8]. In this study, we evaluated the performance of the Gemini LLM by combining the multi-view prompt expressions with medical expertise together with in-context learning using a different number of examples for in-context learning. Figure 3 shows an overview of the multi-view prompting approach developed and implemented in this study for the Gemini LLM. This prompting framework includes several essential elements, including a clear task definition for the LLM, annotations based on fine granularity medical expert input, illustrative examples for few-shot learning, and a structured format for image classification tasks. Using this new prompt tuning approach, the classification tool was evaluated using held out test dataset with 123 subjects (26 control subjects and 87 patients with amblyopia).

3. Results

3.1 Experimental Setup

This study used Gemini 1.0 version together with the multi-view prompting strategy described in the previous section. The evaluation was structured into four few-shot prompting scenarios with images from 2, 4, 8 or 12 subjects and the three viewing conditions. The subjects used in the few-shot prompting were selected to be representative of amblyopic patients and control subjects. This process was repeated for each of the viewing conditions (BV, MV, and BV + MV).

In the first phase of evaluation, the Gemini model was used to perform a binary classification of subjects into two groups, that is, patients with amblyopia or control subjects. To demonstrate the effectiveness of prompt expression created using fine granularity input from pediatric ophthalmologist, we performed a comparative evaluation of the LLM with and without the inclusion of medical expert generated text (Table 1). After establishing the effectiveness of medical expert mediated prompt expressions, all subsequent evaluations were performed using prompts incorporating the input of ophthalmologists only. In the second phase of the evaluation, the prompt-tuned Gemini model was used to classify patients with amblyopia based on the severity of amblyopia, that is, across the three categories of: (1) patients with severe or moderate amblyopia and control subjects; (2) patients with mild or treated amblyopia and control subjects; and (3) amblyopic patient with nystagmus and control subjects. A key feature of the multi-view prompt development was an iterative feedback mechanism that tuned the prompt to reflect the reasoning approach used by pediatric ophthalmologists in diagnosing patients with amblyopia.

The evaluation experiments were conducted on a server with 32 GB RAM and Intel® Core™ i7-13700K CPU, using Ubuntu 22.04.1 LTS (64-bit). The 123 subjects constituting the test dataset did not have any overlap with the 12 subjects used for in-context learning. Temperature is a tunable parameter of LLMs that corresponds to increased randomness in results with higher value and greater deterministic results with lower values [3]. In this study, a temperature parameter of 0.7 was found to be optimal for achieving the highest accuracy.

3.2. Evaluation Results

Table 1 enumerates the performance of the LLM in phase 1 using a single image corresponding to single eye movement plot of a subject. The results show that the performance of the Gemini LLM is low due to the limited context provided in the single eye movement plot and the lack of medical expert input in the prompt text. The Gemini model reports an accuracy of 0.54 when distinguishing between patients with amblyopia (irrespective of their severity) and control subjects; an accuracy of 0.58 to classify patients with mild or treated amblyopia versus control subjects; and an accuracy of 0.62 for classifying patients with moderate and severe amblyopia versus control subjects.

Table 1. Limited view prompting: Performance of the Gemini model using a single plot of eye movement recordings for classifying patients with amblyopia (together with subpopulations) and control subjects.

Model	Amblyopia versus control	Mild and treated amblyopia versus control	Moderate and severe amblyopia versus control
Few shot (K=12)	0.54	0.58	0.62

To address the limitations of our initial prompting approach, we used the multi-view prompting approach consisting of all 6-eye movement recording plots. Table 2 shows the evaluation results across three classification tasks, that is: all patients with amblyopia versus control subjects, and subgroups of amblyopic patients based on their severity. Table 2 also shows the performance of the LLM with and without input from a medical expert. The classification results show significant improvement with input from medical expert together with higher number of learning examples (few shot learning with k = 12) under BV viewing condition with eye tracking data recorded from both eyes.

Table 3 lists the performance of the models across different classification tasks based on the severity of amblyopia, including mild and treated amblyopia, moderate and severe amblyopia, and amblyopia with nystagmus as compared to the control group using the few-shot in-context learning approach. Table 3 also includes the results for combined binocular and monocular viewing (BV + MV), MV, and BV models, which demonstrates the improvement in accuracy of the model with an increased number of examples for in-context learning across all viewing conditions.

Table 2. Comparative evaluation of amblyopia classification using varied few-shot learning and viewing conditions with base and expert crafted prompts in collages.

Model	BV + MV model (Base prompt) Accuracy	BV + MV model (Prompt with medical expert input) Accuracy	BV + MV model (Base prompt) Accuracy	BV + MV model (Prompt with medical expert input) Accuracy	BV model (Base prompt) Accuracy	BV model (Prompt with medical expert input) Accuracy
Few shot (K=2)	0.55	0.62	0.55	0.63	0.6	0.65
Few shot (K=4)	0.55	0.64	0.55	0.67	0.6	0.6
Few shot (K=8)	0.57	0.74	0.57	0.79	0.62	0.79
Few shot (K=12)	0.57	0.78	0.57	0.78	0.62	0.8

BV: Binocular viewing (Two images per collage), MV: Monocular viewing (Four images per collage), BV+MV (6 images per collage)

Table 3. Comparative evaluation of amblyopia classification models using expert-crafted prompts and collages across various few-shot learning and viewing conditions.

Model	BV + MV model Accuracy	MV model Accuracy	BV model Accuracy
Mild and treated amblyopia versus control			
Few shot (K=2)	0.6	0.65	0.7
Few shot (K=4)	0.63	0.69	0.68
Few shot (K=8)	0.76	0.78	0.8
Few shot (K=12)	0.77	0.81	0.81
Moderate and severe amblyopia versus control			
Few shot (K=2)	0.68	0.69	0.7
Few shot (K=4)	0.71	0.74	0.77
Few shot (K=8)	0.78	0.8	0.81
Few shot (K=12)	0.8	0.82	0.83
Amblyopia with nystagmus versus control			
Few shot (K=2)	0.7	0.72	0.73
Few shot (K=4)	0.74	0.74	0.76
Few shot (K=8)	0.8	0.81	0.83
Few shot (K=12)	0.8	0.83	0.85

4. Discussion and Conclusion

Our study for the first time demonstrates that LLMs such as Gemini 1.0 can be successfully used to analyze eye tracking data for diagnosis of amblyopia in children. We demonstrate the effectiveness of a new multi-view prompting framework for the ophthalmology domain that enhances the contextual information available to an LLM to

perform classification tasks with higher accuracy. This study also validates the results from previous studies that LLMs are capable few-shot learners, and their performance improves with higher number of examples used for in-context learning, including imaging data used in ophthalmology applications. Finally, the results of this study demonstrate the critical value of input from medical experts such as pediatric ophthalmologists in significantly enhancing the performance of LLMs. Although many studies have identified the limitations of LLMs in accurately addressing medical domain tasks due to the lack of specific domain training; however, the powerful transfer learning capabilities of LLMs is clearly demonstrated in this study that leverages the high-quality input from a medical expert to generate high accuracy results.

.

The integration of expert feedback together with reinforcement learning method or reinforcement learning with human feedback (RHLF) has been successfully used to improve instruction following in LLMs [18]. Our approach to incorporate expert feedback in prompt tuning results in significant improvement in the quality of results and this approach can also be integrated with other approaches as chain of though (CoT) prompting. Specifically, in our future work, we propose a multimodal CoT prompting approach to further improve the accuracy of the LLMs together with greater insight into the analysis approach used by LLMs for classification in ophthalmology applications.

Acknowledgement. This research was funded in part by grants from the US National Institutes of Health (NIH): U24EB029005, R01DA053028, the US Department of Defense (DoD) grant W81XWH2110859, and the Clinical and Translational Science Collaborative of Cleveland, which is funded by the NIH, National Center for Advancing Translational Sciences, Clinical and Translational Science Award grant, UL1TR002548. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Additional funding came from NEI T32: 5 T32 EY 24236-4, Case Western Reserve University Biomedical Research Fellowship-Hartwell Foundation (F.G.), Blind Children's Foundation (F.G.), Research to Prevent Blindness Disney Amblyopia Award (F.G.), and Cleveland Clinic RPC Grant (F.G.). Support also included a Cleveland VA Medical Center grant 5I21RX003878-02 (AS and FG), Research to Prevent Blindness, CCLCM Unrestricted Block Grant, NIH-NEI P30 Core Grant, and Cleveland Eye Bank.

Ethics and Institutional Governance Approvals: This study was reviewed and approved by the Cleveland Clinic Institutional Review Board (IRB). Written informed consent was obtained from each participant or parent/legal guardian for this study as mandated by the Declaration of Helsinki. The patient data with protected health information (PHI) elements was securely stored at Cleveland Clinic and only the deidentified data was subsequently shared with Case Western Reserve University for analysis by the transformer deep learning model.

References

1. McKean-Cowdin, R., et al., *Prevalence of amblyopia or strabismus in asian and non-Hispanic white preschool children: multi-ethnic pediatric eye disease study.* Ophthalmology, 2013. **120**(10): p. 2117-24.
2. Singhal, K., et al., *Large language models encode clinical knowledge.* Nature, 2023. **620**(7972): p. 172-180.

3. Sahoo, S.S., Plasek J.M., Xu H., Uzuner O, Cohen C., Yetisgen, M., Liu, H., Stéphane, M., Wang, Y., *Large Language Models for Biomedicine: Foundations, Opportunities, Challenges, and Best Practices.* Journal of the American Medical Informatics Association (JAMIA), 2024.
4. Shah, N.H., Entwistle, D., Pfeffer, M.A., *Creation and adoption of large language models in medicine.* Jama, 2023. **330**(9): p. 866-869.
5. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., *Large language models in medicine.* Nature medicine, 2023. **29**(8): p. 1930-1940.
6. Chowdhery, A., et al., *Palm: Scaling language modeling with pathways.* Journal of Machine Learning Research, 2023. **24**(240): p. 1-113.
7. Team, G., et al., *Gemini: a family of highly capable multimodal models.* arXiv preprint arXiv:2312.11805, 2023.
8. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
9. Ghasia, F. and J. Wang, *Amblyopia and fixation eye movements.* Journal of the Neurological Sciences, 2022: p. 120373.
10. Subramanian, V., R.M. Jost, and E.E. Birch, *A quantitative study of fixation stability in amblyopia.* Invest Ophthalmol Vis Sci, 2013. **54**(3): p. 1998-2003.
11. Shi, X.F., et al., *Fixational saccadic eye movements are altered in anisometropic amblyopia.* Restor Neurol Neurosci, 2012. **30**(6): p. 445-62.
12. Niechwiej-Szwedo, E., L. Colpa, and A.M. Wong, *Visuomotor behaviour in amblyopia: deficits and compensatory adaptations.* Neural Plasticity, 2019. **2019**.
13. Vadhera, R. and M. Sharma. *Review of Amblyopia and Artificial Intelligence Techniques Used for Its Detection*. in *Proceedings of the Second International Conference on Information Management and Machine Intelligence: ICIMMI 2020*. 2021. Springer.
14. Ting, D.S.W., et al., *Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes.* Jama, 2017. **318**(22): p. 2211-2223.
15. Holmes, J.M. and M.P. Clarke, *Amblyopia.* Lancet, 2006. **367**(9519): p. 1343-51.
16. Webber, A.L. and J. Wood, *Amblyopia: prevalence, natural history, functional effects and treatment.* Clin Exp Optom, 2005. **88**(6): p. 365-75.
17. Wang, Y., et al., *Generalizing from a few examples: A survey on few-shot learning.* ACM computing surveys (csur), 2020. **53**(3): p. 1-34.
18. Ouyang, L., et al., *Training language models to follow instructions with human feedback.* Advances in neural information processing systems, 2022. **35**: p. 27730-27744.