Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models

Angela Lopez-Cardona [1,2] Carlos Segura [1] Alexandros Karatzoglou [3] Sergi Abadal [2] Ioannis Arapakis [1]

[1] Telefónica Scientific Research, Barcelona, Spain [2] Universitat Politècnica de Catalunya, Barcelona, Spain [3] Amazon, Barcelona, Spain

4/2025

Table of Contents

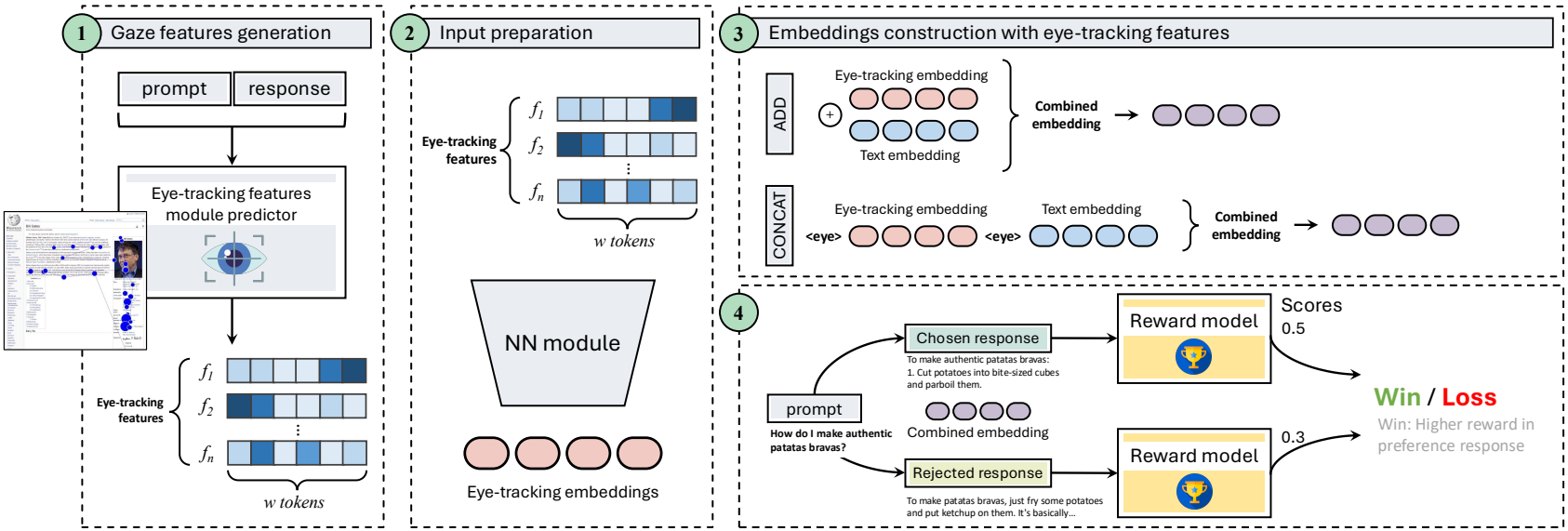

Figure: GazeReward Framework for using eye-tracking data for reward modelling. We use a generator model to compute ET features on a preference dataset D and we train the human preference by combining both text and ET embeddings

Eye-tracking (ET) technology

- Eye-tracking devices gather **raw data** about scanpaths, frequency and duration of fixations, as well as pupillary dilation variations.
- Algorithms to detect **fixations sequence**.
- From this **fixations sequence** $\rightarrow$ reading measures.

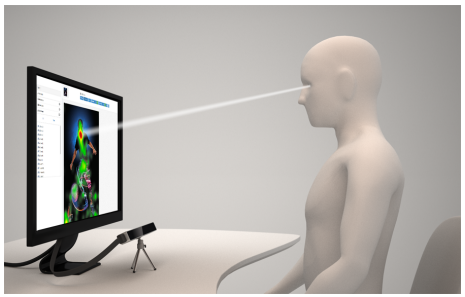

Figure: Eye-tracker

here's a limerick about cucumbers: there once
was a green cucumber, that was really quite
a late bloomer, it grew on the vine, and was
oh so fine, until it became someone's consumer.

Figure: TRT per word. Deeper colour represents longer fixation.

Table: ET reading measures per word [10]

Acronym	Measure	Definition
FFD	First Fixation Duration	Time spent on the initial fixation
GPT	Go-Past Time	Cumulative fixation time before moving to the right
TRT	Total Reading Time	Overall time spent fixating on a word
nFix	Number of Fixations	Number of fixations on each word
fixProp	Proportion of participants	Proportion of participants that fixated on the word

Training Large Language Models (LLMs)

- Unsupervised pretraining from raw text → **Language understanding**
- Large scale instruction tuning and reinforcement learning → **Human alignment**

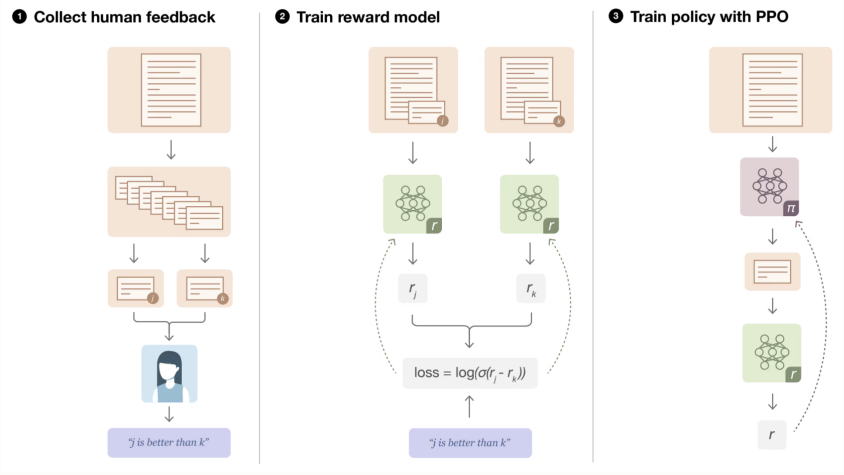

Figure: RLHF diagram [1]

RLHF phases:

- (1) collecting feedback
- (2) training a RM based on that feedback
- (3) optimising the LLMs using RL techniques

Other methods:

- Direct Preference Optimization (DPO) [25]
- Reinforcement Learning from AI Feedback (RLAIF) [2]

Reward Model (RM)

- The most **popular approach** to reward modeling follows the framework introduced by [23].
- Pretrained LLM + regression head that outputs a scalar **reward** [27].
- Reward indicates the quality of the model generation $\Rightarrow$ proxy for human judgment.
- **Dataset** D consists of human comparisons, where $r_\theta(x, y_w)$, $r_\theta(x, y_l)$ represents the RM θ scalar outputs for the preferred and less preferred completions, respectively [23].
- Loss function is defined in Equation 1, where y_w refers to the **preferred response** in a **pair** of completions y_w and y_l.

$$\text{loss}(\theta) = -E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right)\right] \quad (1)$$

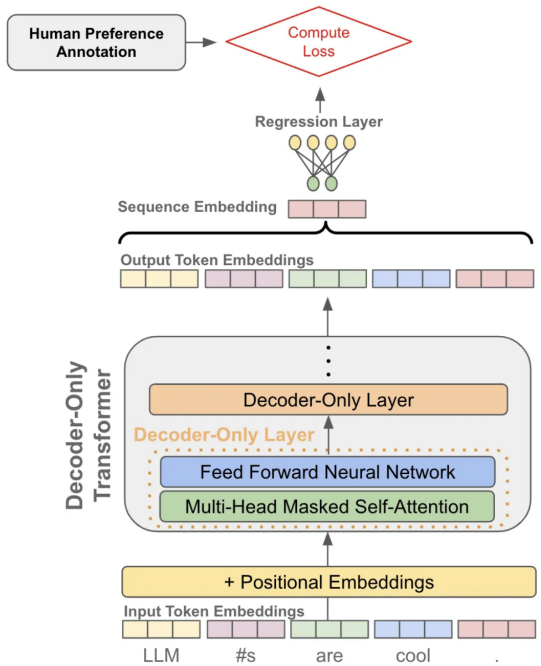

Figure: Reward model [1]

Beyond its role in training, RMs are also used in **inference**, as seen in best-of-N sampling [4]. Additionally, RMs play a key role in generating **synthetic data** for preference alignment [2].

Main research questions → how to combine this two elements?

State of the Art

- Several studies have examined **alternative** versions for refining **RMs**. More **fine-grained** reward structures [2, 32, 6, 30]. Another line of research has focused on Process Based Reward Models (PRMs) [19, 28]. Some works have leveraged **synthetic** preference data for reward modelling [4, 14, 31] and self-training [24].
- Optimal methods for gathering **feedback** to align LLMs with human goals remain an **open question** [3].
- **Explicit feedback** is collected after users have reviewed model outputs. However, human decision making is inherently complex and involves a comprehensive evaluation of diverse information types before taking action.
- **ET** data has been shown to add value in various **NLP tasks**, as demonstrated by prior work [12, 15, 9, 33, 16, 5, 21, 22].

Contribution

- We augment the traditional **Reward Model**, by incorporating (artificial) **implicit feedback** through ET.

Method

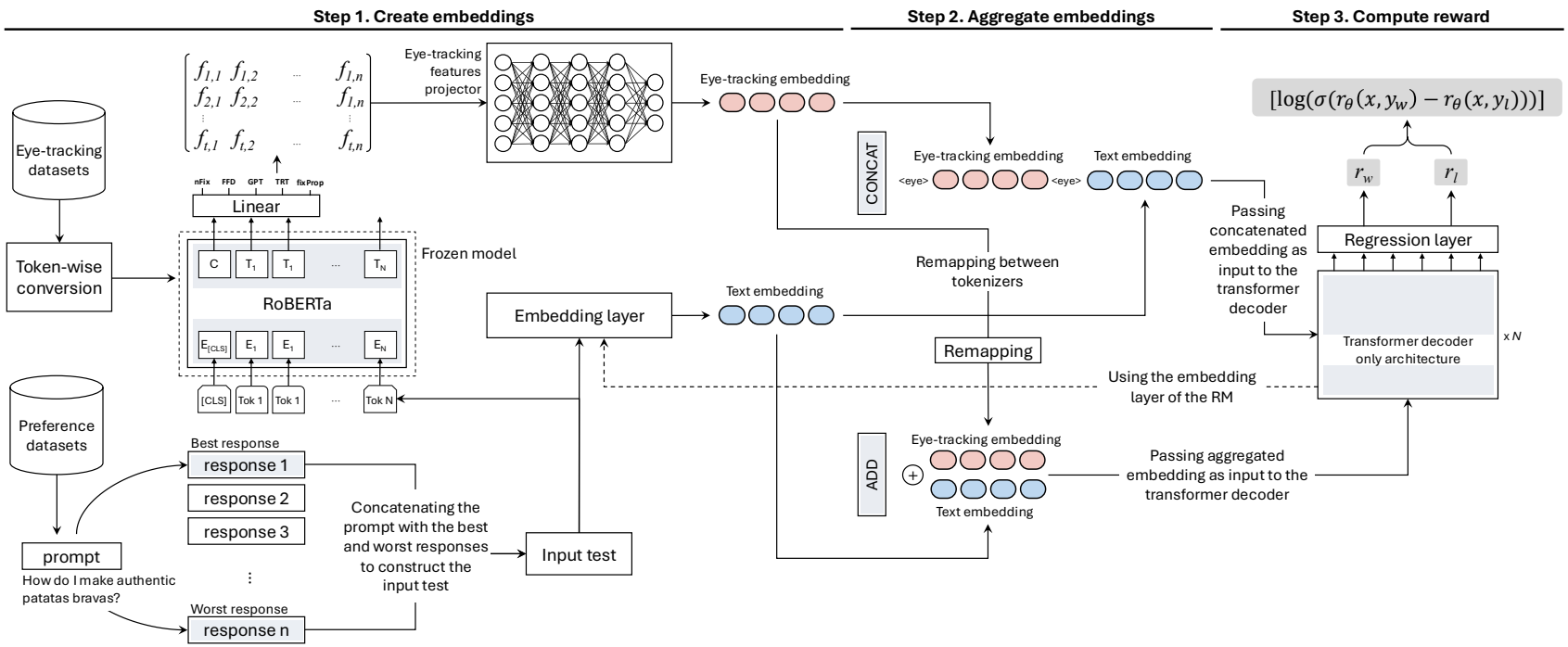

Figure: Overview of the **GazeReward** framework, incorporating eye-tracking features into the reward model. The architecture is illustrated in the figure using the second ET prediction model, but it would be identical if the first one were used instead

Experimental setup

2 Inclusion method:

- **GazeConcat:** The ET embedding, denoted as emb_{ETF}, is concatenated with the text embedding H to form the input for the RM: $(emb(<eye>) \circ emb_{ETF} \circ emb(</eye>) \circ H)$.
- **GazeAdd:** The input to the RM consists of the ET embedding emb_{ETF} and the text embedding H, which are added in an elementwise fashion: $(emb_{ETF} + H)$.

2 Datasets:

Table: Overview of different corpora used in the study to train the reward model.

Corpus	Train set	Val. set	Test set	Lang.	Reference
OASST1	6567	1160	416	EN*	[17]
HelpSteer2	5938	1049	364	EN	[29]

3 open source backbone models:
Llama 3 8B, Llama 3 8B-instruct [7] and Mistral 7B [13].

3 ET feature importance:

- $fcomb_1$ – TRT generated by [11].
- $fcomb_{2.5}$ – nFix, FFD, GPT, TRT, fixProp generated by [18].
- $fcomb_{2.2}$ – TRT and FFD generated by [18].

Experimental setup

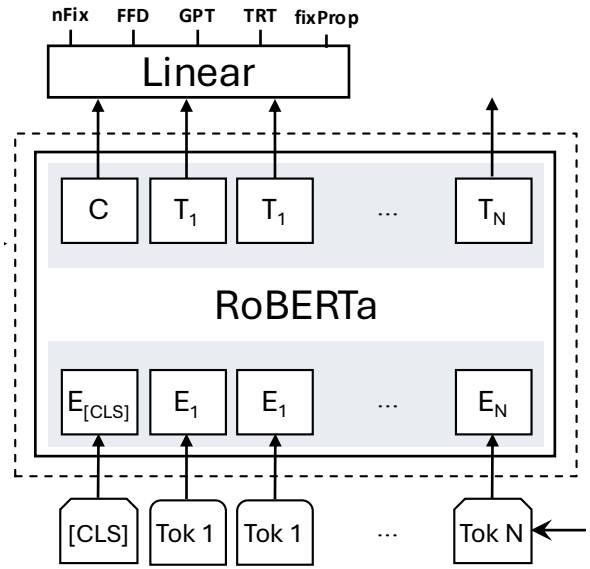

Figure: Generative model architecture [18].

2 Generative models:

- [18]: Based on RoBERTa [20] with a regression head on each token. This head is a linear layer that outputs five features: FFD, fixProp, GPT, TRT, and nFix.
- [11]: Based on T5 embedding layer [26], a two-layer BiLSTM [8], and a one-hidden-layer MLP. Predicts total reading time (TRT) per token.

Results

Table: Reward modeling accuracy (%) for OASST1 dataset. The highest results are in bold and the second highest are underlined.

		Llama-3-8B-Instruct		Llama-3-8B		Mistral-7B	
	baseline	65.9 ± 0.5	diff (%)	65.5 ± 2.1	diff (%)	66.3 ± 0.1	diff (%)
GazeConcat	$fcomb_1$	69.0 ± 0.4*	4.7	69.3 ± 0.6	5.9	67.6 ± 1.7	2.1
	$fcomb_{2.5}$	70.2 ± 0.3**	6.6	**71.5 ± 0.5**	9.2	70.2 ± 0.4*	5.9
	$fcomb_{2.2}$	70.0 ± 0.4**	6.3	71.2 ± 0.8	8.8	**71.0 ± 1.0**	7.1
GazeAdd	$fcomb_1$	68.9 ± 0.9	4.6	68.9 ± 1.0	5.3	-	
	$fcomb_{2.5}$	**70.2 ± 0.1***	6.6	69.5 ± 0.3	6.1	-	
	$fcomb_{2.2}$	69.0 ± 0.4*	4.7	68.3 ± 0.7	4.4	-	

Table: Reward modeling accuracy (%) for Helpsteer2 dataset. The highest results are in bold and the second highest are underlined.

		Llama-3-8B-Instruct		Llama-3-8B		Mistral-7B	
	baseline	54.7 ± 0.7	diff (%)	53.3 ± 0.8	diff (%)	54.1 ± 0.3	diff (%)
GazeConcat	$fcomb_1$	61.1 ± 1.2*	11.8	59.1 ± 0.2*	10.8	57.6 ± 2.3	6.5
	$fcomb_{2.5}$	58.5 ± 1.6	7.0	60.3 ± 0.5**	13.2	**58.7 ± 2.4**	8.5
	$fcomb_{2.2}$	60.6 ± 3.3	10.9	57.9 ± 2.0	8.6	56.0 ± 2.4	3.4
GazeAdd	$fcomb_1$	**62.3 ± 0.6****	13.9	**62.4 ± 1.0****	17.0	-	
	$fcomb_{2.5}$	59.6 ± 1.1*	9.0	58.6 ± 1.2*	10.0	-	
	$fcomb_{2.2}$	60.3 ± 0.5**	10.2	59.3 ± 0.1*	11.3	-	

Conclusions and future direction

- **Implicit Feedback for Reward Models:** The study presents a new framework integrating implicit feedback into the RM to enhance LLMs alignment and synthetic data generation.
- **Validation with Open-Source Models:** The approach was tested using Llama 3 and Mistral models, demonstrating consistent improvements in predicting user preferences in state-of-the-art models.
- **ET Features for Scalability:** The method leverages ET features generated by models, making it fully scalable and applicable to various human alignment techniques, including those using synthetic datasets.
- **Limitations in Data and Generalization:** The ET prediction models were trained on small, English-only datasets, raising questions about generalizability. Future work should focus on collecting task-specific ET data and exploring alternative feature integration methods.
- **Future Directions in Scaling and Application:** Further validation on larger models and datasets and integration into RLHF or rejection sampling for preference dataset generation could enhance alignment and overall LLM performance.

Contact information

Ángela López Cardona
angela.lopezcardona@telefonica.com

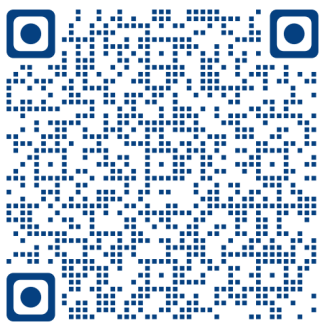

Figure: GitHub Respository

Acknowledgments

SYMBIOTIK Horizon Europe's European Innovation Council, under Pathfinder Grant 101071147. **Grant AGAUR** Generalitat de Catalunya, under Grant AGAUR 2023 DI060

Bibliography I

[1] Reinforcement learning with human feedback (RLHF) for LLMs, June 2023.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan.
Constitutional AI: Harmlessness from AI Feedback, December 2022.
arXiv:2212.08073 [cs].

Bibliography II

[3] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell.
Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, July 2023.
arXiv:2307.15217 [cs].

[4] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun.
UltraFeedback: Boosting Language Models with Scaled AI Feedback, July 2024.
arXiv:2310.01377 [cs].

[5] Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger.
Pre-Trained Language Models Augmented with Synthetic Scanpaths for Natural Language Understanding.
In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6500–6507, Singapore, December 2023. Association for Computational Linguistics.

Bibliography III

[6] Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev.
SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF, October 2023.
arXiv:2310.05344 [cs].

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer

Bibliography IV

van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas

Bibliography V

Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil,

Bibliography VI

Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal,

Bibliography VII

Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang,

Bibliography VIII

Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao.
The Llama 3 Herd of Models, August 2024.
arXiv:2407.21783 [cs].

[8] Sepp Hochreiter and Jürgen Schmidhuber.
Long Short-Term Memory.
Neural Computation, 9(8):1735–1780, November 1997.

[9] Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang.
Advancing NLP with Cognitive Language Processing Signals, April 2019.
arXiv:1904.02682 [cs].

[10] Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus.
CMCL 2021 Shared Task on Eye-Tracking Prediction.
In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online, June 2021. Association for Computational Linguistics.

Bibliography IX

[11] Xinting Huang and Nora Hollenstein.
Long-Range Language Modeling with Selective Cache.
In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4838–4858, Singapore, December 2023. Association for Computational Linguistics.

[12] Xinting Huang, Jiajing Wan, Ioannis Kritikos, and Nora Hollenstein.
Longer Fixations, More Computation: Gaze-Guided Recurrent Neural Networks, October 2023.
arXiv:2311.00159 [cs].

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed.
Mistral 7B, October 2023.
arXiv:2310.06825 [cs].

[14] Tianhao Wu * Hanlin Zhu and Jiantao Banghua Zhu * Jiao, Evan Frick *.
Starling-7B: Increasing LLM Helpfulness & Harmlessness with RLAIF, November 2023.

Bibliography X

[15] Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy.
Synthesizing Human Gaze Feedback for Improved NLP Performance.
pages 1895–1908, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[16] Samuel Kiegeland, David Robert Reich, Ryan Cotterell, Lena Ann Jäger, and Ethan Wilcox.
The Pupil Becomes the Master: Eye-Tracking Feedback for Tuning LLMs.
July 2024.

[17] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick.
OpenAssistant Conversations – Democratizing Large Language Model Alignment, April 2023.
arXiv:2304.07327 [cs].

Bibliography XI

[18] Bai Li and Frank Rudzicz.
TorontoCL at CMCL 2021 Shared Task: RoBERTa with Multi-Stage Fine-Tuning for Eye-Tracking Prediction.
In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online, June 2021. Association for Computational Linguistics.

[19] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe.
LET'S VERIFY STEP BY STEP.
2024.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019.
arXiv:1907.11692 [cs].

Bibliography XII

[21] Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya.
Eyes are the Windows to the Soul: Predicting the Rating of Text Quality Using Gaze Behaviour.
In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352–2362, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[22] Erik S. McGuire and Noriko Tomuro.
Sentiment Analysis with Cognitive Attention Supervision.
Proceedings of the Canadian Conference on Artificial Intelligence, June 2021.

[23] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe.
Training language models to follow instructions with human feedback, March 2022.
arXiv:2203.02155 [cs].

Bibliography XIII

[24] Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn.
West-of-N: Synthetic Preference Generation for Improved Reward Modeling, January 2024.
arXiv:2401.12086 [cs].

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn.
Direct Preference Optimization: Your Language Model is Secretly a Reward Model, December 2023.
arXiv:2305.18290 [cs].

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu.
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, July 2020.
arXiv:1910.10683 [cs, stat].

Bibliography XIV

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom.
Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023.
arXiv:2307.09288 [cs].

[28] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins.
Solving math word problems with process- and outcome-based feedback, November 2022.
arXiv:2211.14275 [cs].

Bibliography XV

[29] Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev.
HelpSteer2: Open-source dataset for training top-performing reward models, June 2024.
arXiv:2406.08673 [cs].

[30] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev.
HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM, November 2023.
arXiv:2311.09528 [cs].

[31] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar.
Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge, July 2024.
arXiv:2407.19594 [cs].

[32] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi.
Fine-Grained Human Feedback Gives Better Rewards for Language Model Training, October 2023.
arXiv:2306.01693 [cs].

Bibliography XVI

[33] Duo Yang and Nora Hollenstein.
PLM-AS: Pre-trained Language Models Augmented with Scanpaths for Sentiment Classification.
Proceedings of the Northern Lights Deep Learning Workshop, 4, January 2023.