![Meta](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/252294889_575082167077436_6034106545912333281_n.svg/meta-logo-primary_standardsize.svg?_nc_cat=1&ccb=1-7&_nc_sid=e280be&_nc_ohc=1qAUzKHJ4V4Q7kNvgFLFxjO&_nc_oc=AdgFEytplInj3jQKOS1-Qw7sbFh6vLK2zQQBxfA3pKkhwaZYtciuyk3Ko8Wokxp8lM_dzPctuHqjK3QyR840UG2O&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh=00_AYATN7SLgT5KZ2Yzg82ZCZEj8WY8pF84E3nT4EJuM_IZOQ&oe=67AC2839)

* Our approach
* Research
* Product experiences
* [Llama](https://llama.meta.com/)
* [Blog](/blog/)
* [Try Meta AI](https://www.meta.ai/?utm_source=ai_meta_site&utm_medium=web&utm_content=AI_nav&utm_campaign=April_moment)
* [](/)

CONVERSATIONAL AI

NLP

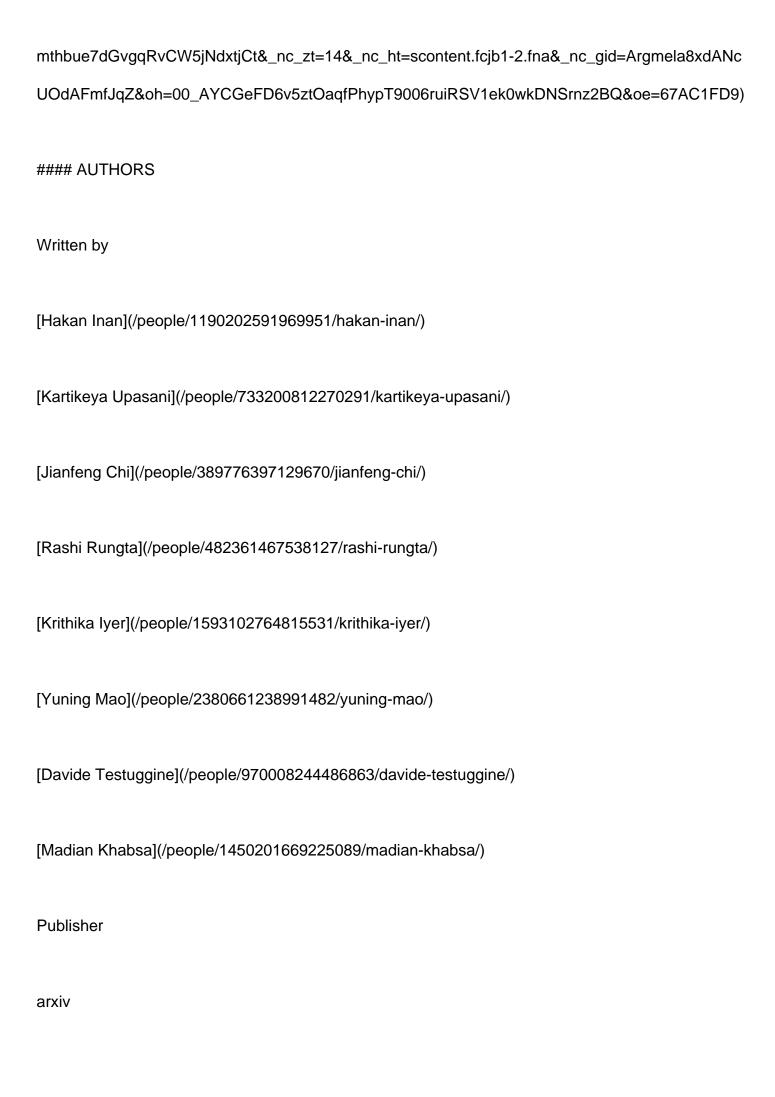# Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations

December 07, 2023

Abstract

We introduce Llama Guard, an LLM-based input-output safeguard model geared towards Human-AI conversation use cases. Our model incorporates a safety risk taxonomy, a valuable tool for categorizing a specific set of safety risks found in LLM prompts (i.e., prompt classification). This taxonomy is also instrumental in classifying the responses generated by LLMs to these prompts, a process we refer to as response classification. For the purpose of both prompt and response classification, we have meticulously gathered a dataset of high quality. Llama Guard, a Llama2-7b model that is instruction-tuned on our collected dataset, albeit low in volume, demonstrates strong performance on existing benchmarks such as the OpenAI Moderation Evaluation dataset and ToxicChat, where its performance matches or exceeds that of currently available content moderation tools. Llama Guard functions as a language model, carrying out multi-class classification and generating binary decision scores. Furthermore, the instruction fine-tuning of Llama Guard allows for the customization of tasks and the adaptation of output formats. This feature enhances the model's capabilities, such as enabling the adjustment of taxonomy categories to align with specific use cases, and facilitating zero-shot or few-shot prompting with diverse taxonomies at the input. We are making Llama Guard model weights available and we encourage researchers to further develop and adapt them to meet the evolving needs of the community for AI safety.

AUTHORS

Written by

[Hakan Inan](/people/1190202591969951/hakan-inan/)

[Kartikeya Upasani](/people/733200812270291/kartikeya-upasani/)

[Jianfeng Chi](/people/389776397129670/jianfeng-chi/)

[Rashi Rungta](/people/482361467538127/rashi-rungta/)

[Krithika Iyer](/people/1593102764815531/krithika-iyer/)

[Yuning Mao](/people/2380661238991482/yuning-mao/)

[Davide Testuggine](/people/970008244486863/davide-testuggine/)

[Madian Khabsa](/people/14502016692225089/madian-khabsa/)

Publisher

arxiv

Research Topics

[Conversational AI](/research/conversational-ai/)

[Natural Language Processing (NLP)](/research/nlp/)

Related Publications

February 07, 2025

NLP

BOUQuET: dataset, Benchmark and Open initiative for Universal Quality Evaluation in Translation

The Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussa, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Yiannis Tsiamas, Arina Turkatenko, Albert Ventayol, Shireen Yates

February 07, 2025

[Read the Paper __](/research/publications/bouquet-dataset-benchmark-and-open-initiative-for-universal-quality-evaluation-in-translation/)

February 06, 2025

RESEARCH

NLP

Brain-to-Text Decoding: A Non-invasive Approach via Typing

Jarod Levy, Mingfang (Lucy) Zhang, Svetlana Pinet, Jérémy Rapin, Hubert Jacob Banville, [Stéphane d'Ascoli](/people/7732427943457653/stephane-d-ascoli/), Jean Remi King

February 06, 2025

[Read the Paper __](/research/publications/brain-to-text-decoding-a-non-invasive-approach-via-typing/)

February 06, 2025

RESEARCH

NLP

From Thought to Action: How a Hierarchy of Neural Dynamics Supports Language Production

Mingfang (Lucy) Zhang, Jarod Levy, [Stéphane d'Ascoli](/people/7732427943457653/stephane-d-ascoli/), Jérémy Rapin, F.-Xavier Alario, Pierre Bourdillon, Svetlana Pinet, Jean Remi King

February 06, 2025

[Read the Paper __](/research/publications/from-thought-to-action-how-a-

hierarchy-of-neural-dynamics-supports-language-production/)

January 04, 2025

NLP

Transformers are Multi-State RNNs

Matanel Oren, Michael Hassid, [Yossef (Yossi)

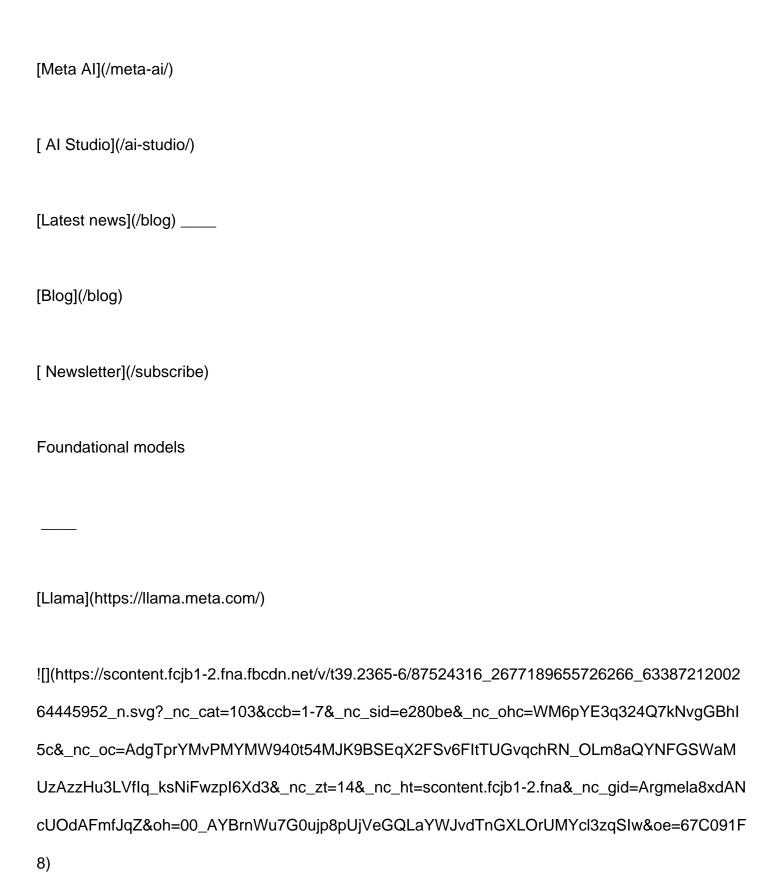Adi](/people/1930846650664668/yossi-adi/), Roy Schwartz

January 04, 2025

[Read the Paper __](/research/publications/transformers-are-multi-state-rnns/)

[See All Papers](/global_search/?content_types%5B0%5D=publication&page=1)

![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.2365-6/91900227_227456494975764_598856160477

2364288_n.jpg?_nc_cat=104&ccb=1-7&_nc_sid=e280be&_nc_ohc=5avdUWmTlwgQ7kNvgGh7PD

o&_nc_oc=AdjncsfdLeGXaxfZVcyN95wsymLH0jqTPLFI4LY-6sx_lEi2hC4hbjqeMuu_2KmWuegGcS

N7tvkOOGnVdcB8lWIe&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAF

mfJqZ&oh=00_AYDifibZqdcnrwLKvU1tD8BPif-5OaYWuBvBp97GmfLVpA&oe=67C0B197)

Help Us Pioneer The Future of AI

We share our open source frameworks, tools, libraries, and models for everything from research exploration to large-scale production deployment.

[Join our Team](/join-us/)

[Our approach](/about) _____

[About AI at Meta](/about)

[People](/results/?content_types%5B0%5D=person&sort_by=random)

[Careers](https://www.metacareers.com/jobs/?is_leadership=0&sub_teams\[0\]=Artificial%20Intelligence&is_in_page=0)

[Research](/research) _____

[Infrastructure](/infrastructure)

[Resources](/resources)

[Demos](https://aidemos.meta.com/)

Product experiences

_____

Foundational models

_____

[Llama](https://llama.meta.com/)

[![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/335682312_964107378293184_3093631164486164913_n.svg?_nc_cat=100&ccb=1-7&_nc_sid=e280be&_nc_ohc=sr5NyYN6SpkQ7kNvgFnsjYH&_nc_oc=Adi2vWDxynYndzx7cDVAFnH0xOvrxrMvng0BCZSfdCPwPSIGE-

EKK1uHCY5iptYdMQuXaOdZowZIZ5q78fsX8OEu&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gi
d=Argmela8xdANcUOdAFmfJqZ&oh=00_AYBQLHhaX4ZLGe64cm9EspP9ilcCz74ywWivB6jeEpVe
yA&oe=67AC1867)!][](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/335682312_964107378293
184_3093631164486164913_n.svg?_nc_cat=100&ccb=1-7&_nc_sid=e280be&_nc_ohc=sr5NyYN6
SpkQ7kNvgFnsjYH&_nc_oc=Adi2vWDxynYndzx7cDVAFnH0xOvrxrMvng0BCZSfdCPwPSIGE-
EKK1uHCY5iptYdMQuXaOdZowZIZ5q78fsX8OEu&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gi
d=Argmela8xdANcUOdAFmfJqZ&oh=00_AYBQLHhaX4ZLGe64cm9EspP9ilcCz74ywWivB6jeEpVe
yA&oe=67AC1867)](https://www.facebook.com/aiatmeta/)

[!][](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336009607_1870102080040414_6753977241
281150924_n.svg?_nc_cat=103&ccb=1-7&_nc_sid=e280be&_nc_ohc=d_oTT67GRxEQ7kNvgHm7
M1_&_nc_oc=Adi9vkmKuKRT27vdR1RFnup7ODKKZtVPGOcPLK7QQtidGF50vCMx4gHEMR0M2S
E-
gBNpwwZRM1JpU-
xkMDhu48IX&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh
=00_AYAQF-
Js2nOX0V2kSWVp4PRiUQnFsJoZDOZTf5TJP5XzUg&oe=67AC10A2)!][](https://scontent.fcjb1-2.fn
a.fbcdn.net/v/t39.8562-6/336009607_1870102080040414_6753977241281150924_n.svg?_nc_cat=
103&ccb=1-7&_nc_sid=e280be&_nc_ohc=d_oTT67GRxEQ7kNvgHm7M1_&_nc_oc=Adi9vkmKuKR
T27vdR1RFnup7ODKKZtVPGOcPLK7QQtidGF50vCMx4gHEMR0M2SE-
gBNpwwZRM1JpU-
xkMDhu48IX&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh
=00_AYAQF-
Js2nOX0V2kSWVp4PRiUQnFsJoZDOZTf5TJP5XzUg&oe=67AC10A2)](https://twitter.com/aiatmeta/
)

[!][](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336289415_1541032296405649_2165099305

308791297_n.svg?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=K9l2gxTnPiIQ7kNvgF-eUuK&_nc_oc=AdhCKrm90eNYDcq6A6m_58etIKG9oB7yOJUwePApgTJ8iCUzVE_47GanUFvZQUQEOR6urbH7XETqZ8FqyHHfCPfC&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh=00_AYAUw8ncM4Pmz1i8V_2R42iQIEuLWS8nRuWdsAZwhNh6Kg&oe=67AC3C7B)](https://www.linkedin.com/showcase/aiatmeta)

[](https://www.youtube.com/@aiatmeta)

Our approach

[Our approach](/about)[About AI at Meta](/about)[People](/results/?content_types%5B0%5D=person&sort_by=random)[Careers](https://www.metacareers.com/jobs/?is_leadership=0&sub_teams\[0\]=Artificial%20Intelligence&is_in_page=0)

Research

[Research](/research)[Infrastructure](/infrastructure)[Resources](/resources)[Demos](https://aidemos.meta.com/)

Product experiences

[Meta AI](/meta-ai/)[AI Studio](/ai-studio/)

Latest news

[Latest news](/blog)[Blog](/blog)[Newsletter](/subscribe)

Foundational models

[Llama](https://llama.meta.com/)

[](https://www.facebook.com/aiatmeta/)

[![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336009607_1870102080040414_6753977241281150924_n.svg?_nc_cat=103&ccb=1-7&_nc_sid=e280be&_nc_ohc=d_oTT67GRxEQ7kNvgHm7M1_&_nc_oc=Adi9vkmKuKR

T27vdR1RFnup7ODKKZtVPGOcPLK7QQtidGF50vCMx4gHEMR0M2SE-

gBNpwwZRM1JpU-

xkMDhu48IX&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh

=00_AYAQF-

Js2nOX0V2kSWVp4PRiUQnFsJoZDOZTf5TJP5XzUg&oe=67AC10A2)](https://twitter.com/aiatmeta/

)

[![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336289415_1541032296405649_2165099305

308791297_n.svg?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=K9l2gxTnPiIQ7kNvgF-

eUuK&_nc_oc=AdhCKrm90eNYDcq6A6m_58etIKG9oB7yOJUwePApgTJ8iCUzVE_47GanUFvZQU

QEOR6urbH7XETqZ8FqyHHfCPfC&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xd

ANcUOdAFmfJqZ&oh=00_AYAUw8ncM4Pmz1i8V_2R42iQIEuLWS8nRuWdsAZwhNh6Kg&oe=67A

C3C7B)!][](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336289415_1541032296405649_2165

099305308791297_n.svg?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=K9l2gxTnPiIQ7kNvg

F-

eUuK&_nc_oc=AdhCKrm90eNYDcq6A6m_58etIKG9oB7yOJUwePApgTJ8iCUzVE_47GanUFvZQU

QEOR6urbH7XETqZ8FqyHHfCPfC&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xd

ANcUOdAFmfJqZ&oh=00_AYAUw8ncM4Pmz1i8V_2R42iQIEuLWS8nRuWdsAZwhNh6Kg&oe=67A

C3C7B)](https://www.linkedin.com/showcase/aiatmeta)

[![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/335648731_142576991793348_77868191898

43639239_n.svg?_nc_cat=108&ccb=1-7&_nc_sid=e280be&_nc_ohc=NXAhgKVNrPYQ7kNvgEOYz

zs&_nc_oc=AdilaERCemBTcYmSXPieZi_uus3IHvTUwmLWOhJOLM2VRqGiPm7LAu5PJvtGqWirJy

nl4GUE4rj-

fs2YzoUls_NA&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&o

h=00_AYBiP_kBJNG9vEAJClBodfc5tyfbASiXzo__PiOwjrp21g&oe=67AC1DAE)![](https://scontent.fc

jb1-2.fna.fbcdn.net/v/t39.8562-6/335648731_142576991793348_7786819189843639239_n.svg?_n

c_cat=108&ccb=1-7&_nc_sid=e280be&_nc_ohc=NXAhgKVNrPYQ7kNvgEOYzzs&_nc_oc=AdilaER CemBTcYmSXPieZi_uus3IHvTUwmLWOhJOLM2VRqGiPm7LAu5PJvtGqWirJynl4GUE4rj- fs2YzoUls_NA&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&o h=00_AYBiP_kBJNG9vEAJClBodfc5tyfbASiXzo__PiOwjrp21g&oe=67AC1DAE)](https://www.youtub e.com/@aiatmeta)

[Privacy Policy](https://www.facebook.com/about/privacy/)

[Terms](https://www.facebook.com/policies/)

[Cookies](https://www.facebook.com/policies/cookies/)

[](https://www.facebook.com/aiatmeta/)

[![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336009607_1870102080040414_6753977241

281150924_n.svg?_nc_cat=103&ccb=1-7&_nc_sid=e280be&_nc_ohc=d_oTT67GRxEQ7kNvgHm7

M1_&_nc_oc=Adi9vkmKuKRT27vdR1RFnup7ODKKZtVPGOcPLK7QQtidGF50vCMx4gHEMR0M2S

E-

gBNpwwZRM1JpU-

xkMDhu48IX&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh

=00_AYAQF-

Js2nOX0V2kSWVp4PRiUQnFsJoZDOZTf5TJP5XzUg&oe=67AC10A2)![](https://scontent.fcjb1-2.fn

a.fbcdn.net/v/t39.8562-6/336009607_1870102080040414_6753977241281150924_n.svg?_nc_cat=

103&ccb=1-7&_nc_sid=e280be&_nc_ohc=d_oTT67GRxEQ7kNvgHm7M1_&_nc_oc=Adi9vkmKuKR

T27vdR1RFnup7ODKKZtVPGOcPLK7QQtidGF50vCMx4gHEMR0M2SE-

gBNpwwZRM1JpU-

xkMDhu48IX&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh

=00_AYAQF-

Js2nOX0V2kSWVp4PRiUQnFsJoZDOZTf5TJP5XzUg&oe=67AC10A2)](https://twitter.com/aiatmeta/

)

[![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336289415_1541032296405649_2165099305

308791297_n.svg?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=K9l2gxTnPiIQ7kNvgF-

eUuK&_nc_oc=AdhCKrm90eNYDcq6A6m_58etIKG9oB7yOJUwePApgTJ8iCUzVE_47GanUFvZQU

QEOR6urbH7XETqZ8FqyHHfCPfC&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xd

ANcUOdAFmfJqZ&oh=00_AYAUw8ncM4Pmz1i8V_2R42iQIEuLWS8nRuWdsAZwhNh6Kg&oe=67A

C3C7B)![](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/336289415_1541032296405649_2165

099305308791297_n.svg?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=K9l2gxTnPiIQ7kNvg

F-

eUuK&_nc_oc=AdhCKrm90eNYDcq6A6m_58etIKG9oB7yOJUwePApgTJ8iCUzVE_47GanUFvZQU

QEOR6urbH7XETqZ8FqyHHfCPfC&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xd

ANcUOdAFmfJqZ&oh=00_AYAUw8ncM4Pmz1i8V_2R42iQIEuLWS8nRuWdsAZwhNh6Kg&oe=67A

C3C7B)](https://www.linkedin.com/showcase/aiatmeta)

[![][](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/335648731_142576991793348_7786819189843639239_n.svg?_nc_cat=108&ccb=1-7&_nc_sid=e280be&_nc_ohc=NXAhgKVNrPYQ7kNvgEOYzzs&_nc_oc=AdilaERCemBTcYmSXPieZi_uus3IHvTUwmLWOhJOLM2VRqGiPm7LAu5PJvtGqWirJynl4GUE4rj-fs2YzoUls_NA&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh=00_AYBiP_kBJNG9vEAJClBodfc5tyfbASiXzo__PiOwjrp21g&oe=67AC1DAE)![][](https://scontent.fcjb1-2.fna.fbcdn.net/v/t39.8562-6/335648731_142576991793348_7786819189843639239_n.svg?_nc_cat=108&ccb=1-7&_nc_sid=e280be&_nc_ohc=NXAhgKVNrPYQ7kNvgEOYzzs&_nc_oc=AdilaERCemBTcYmSXPieZi_uus3IHvTUwmLWOhJOLM2VRqGiPm7LAu5PJvtGqWirJynl4GUE4rj-fs2YzoUls_NA&_nc_zt=14&_nc_ht=scontent.fcjb1-2.fna&_nc_gid=Argmela8xdANcUOdAFmfJqZ&oh=00_AYBiP_kBJNG9vEAJClBodfc5tyfbASiXzo__PiOwjrp21g&oe=67AC1DAE)](https://www.youtube.com/@aiatmeta)