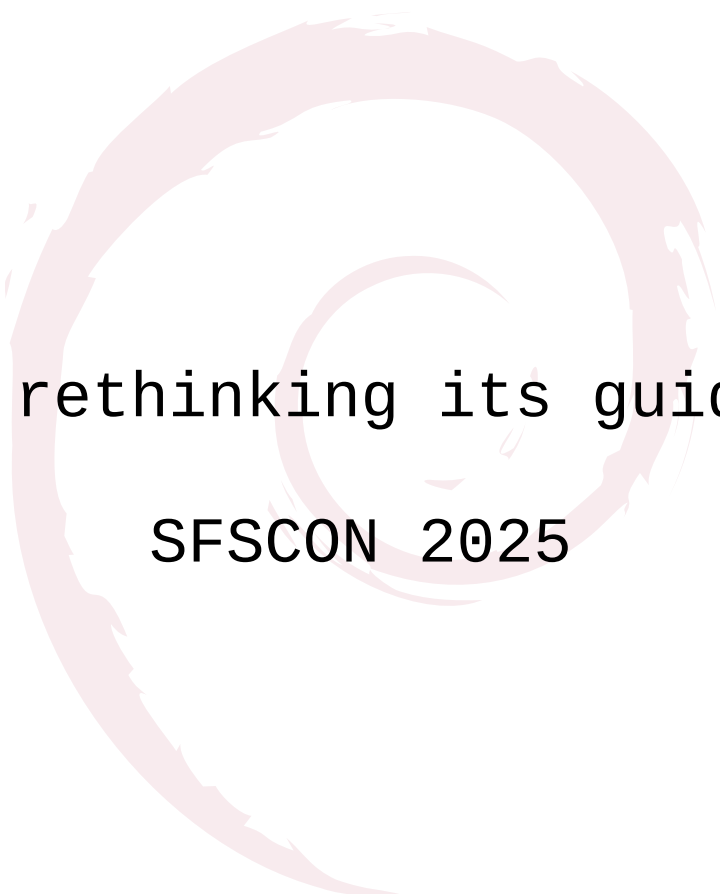


Why you cannot “apt install” an LLM (yet)



How Debian is rethinking its guidelines for AI

SFSCON 2025

About me

- Andrea Pappacoda <tachi>
- Debian contributor since 2021
- Debian Developer since 2024
- Computer science student specializing in cybersecurity (sometimes)

Intro

So, why cannot you
`apt-install` an LLM?

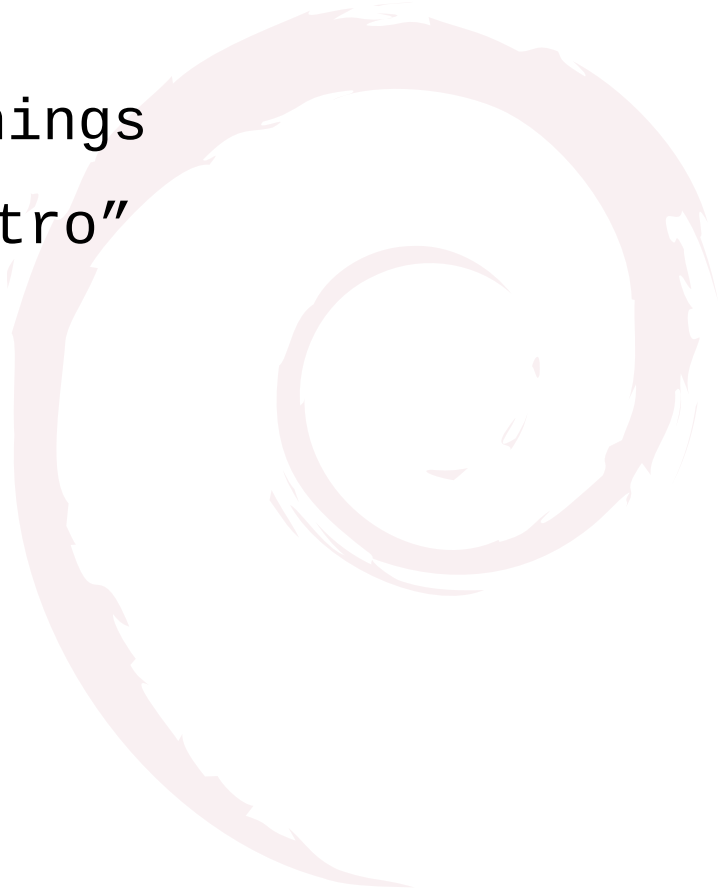
It's a bit of a mess!

- In the value chain
- Things change with time
- We need free datasets

About Debian

Debian is many things

- GNU/Linux “distro”



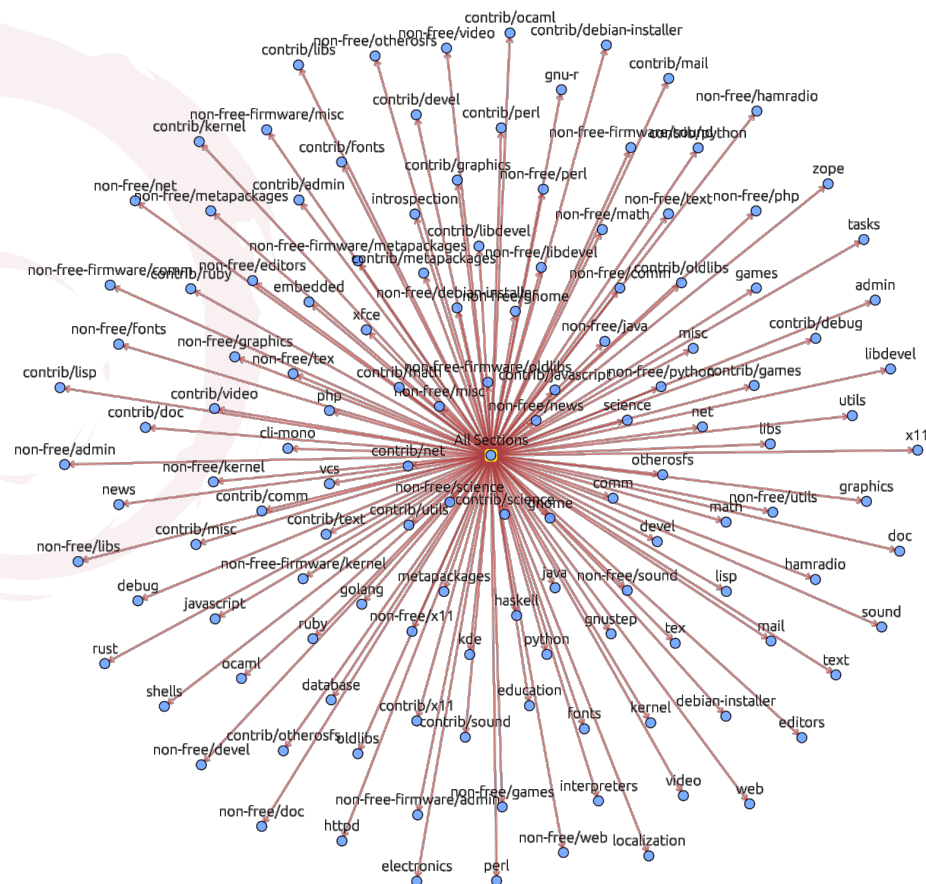
About Debian

Debian is many things

- GNU/Linux “distro”
- Distribution of Free Software

About Debian

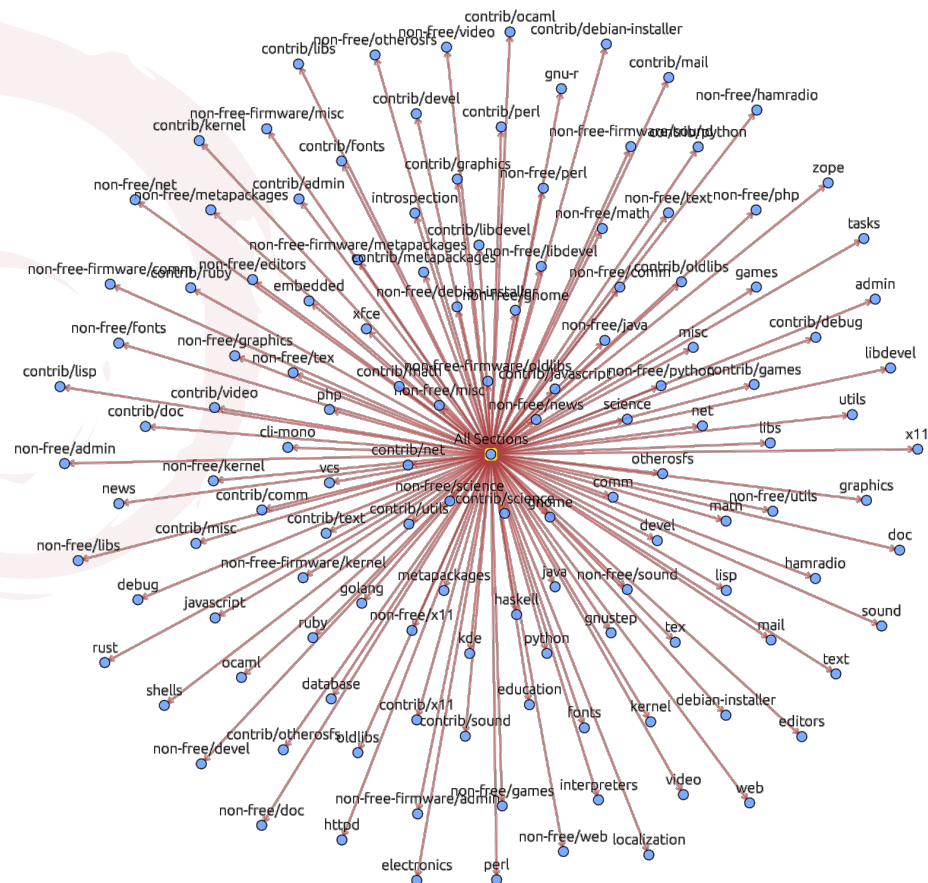
We distribute a lot of software.



About Debian

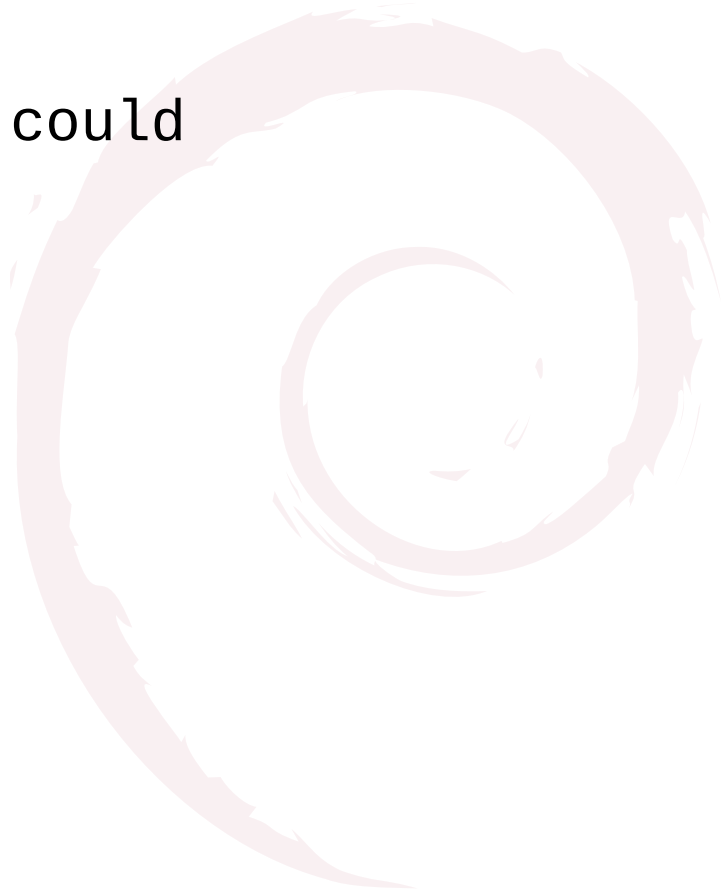
We distribute a lot of software.

If every dot is a category... how many packages are there?



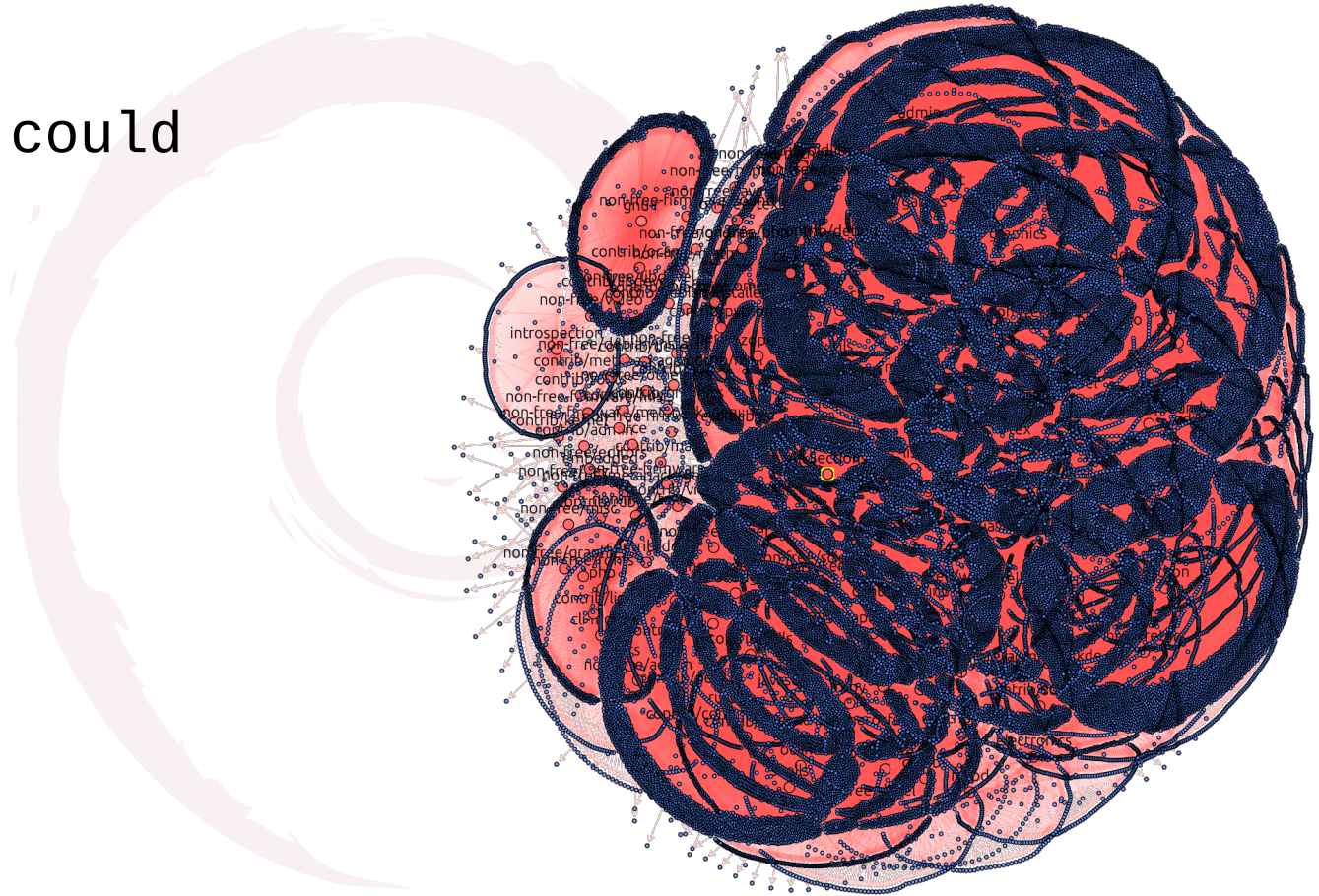
About Debian

...more than you could
imagine



About Debian

...more than you could
imagine



About Debian

It's a community of people

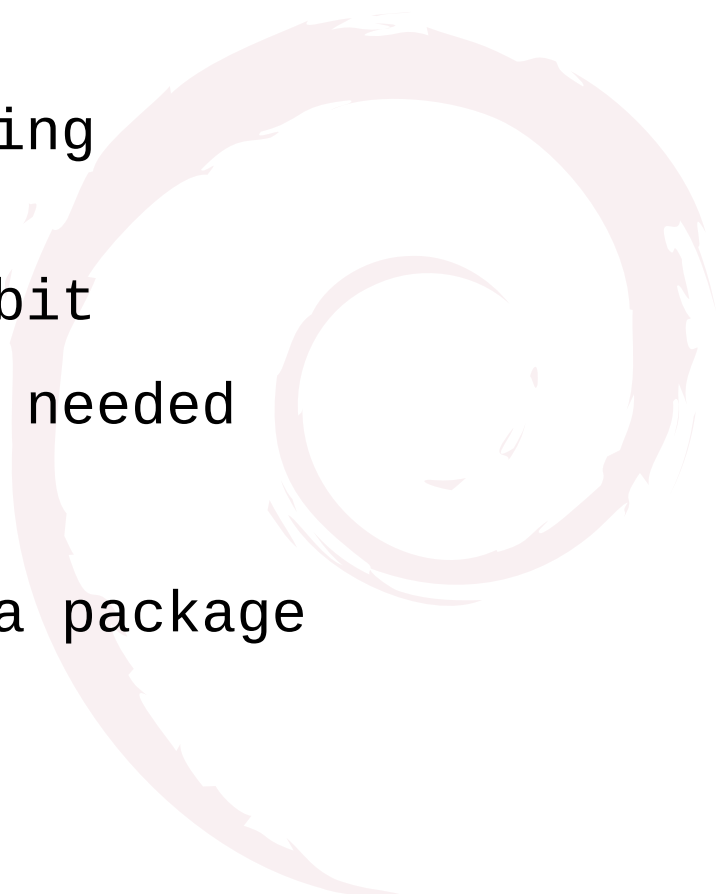


DebConf26
see you in
Santa Fe
Argentina

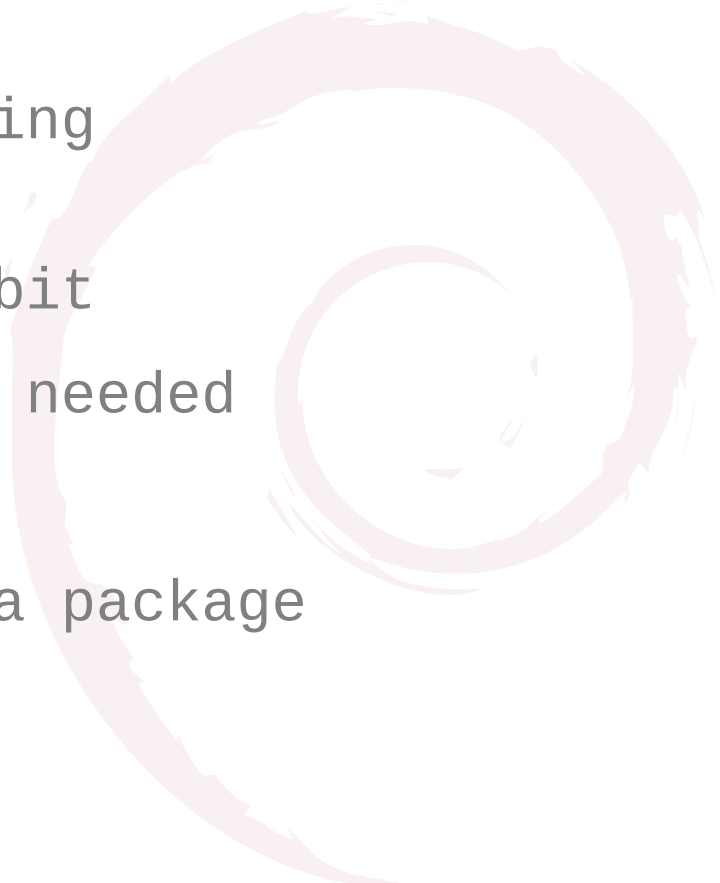


"Whoever takes without giving doesn't have any friend;
whoever takes and gives gets love in every country"
Breton proverb

Birth of a package

- We find something interesting
 - We study it a bit
 - We tweak it if needed
 - We build it
 - We wrap it in a package
- 

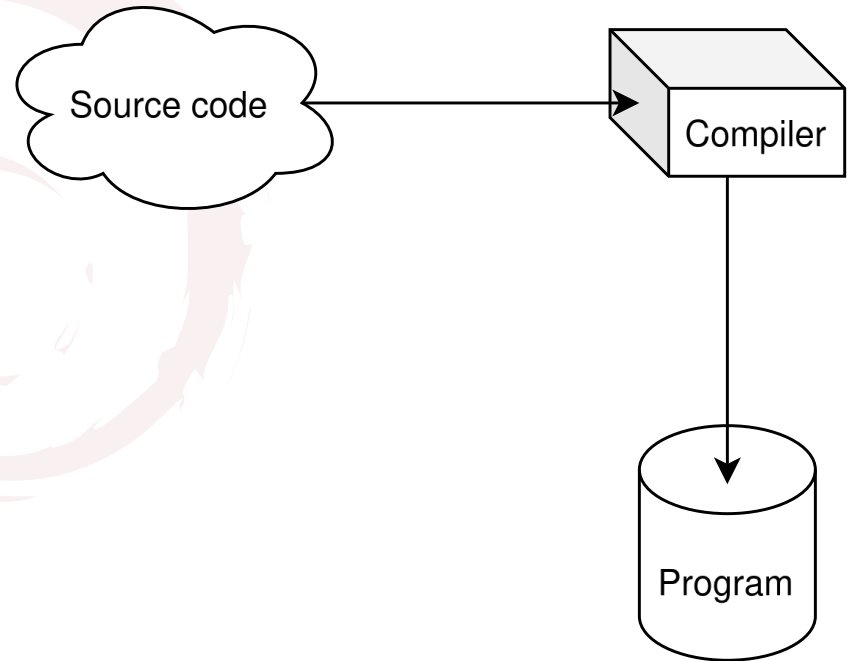
Birth of a package

- We find something interesting
 - We study it a bit
 - We tweak it if needed
 - We build it
 - We wrap it in a package
- 

Birth of a package

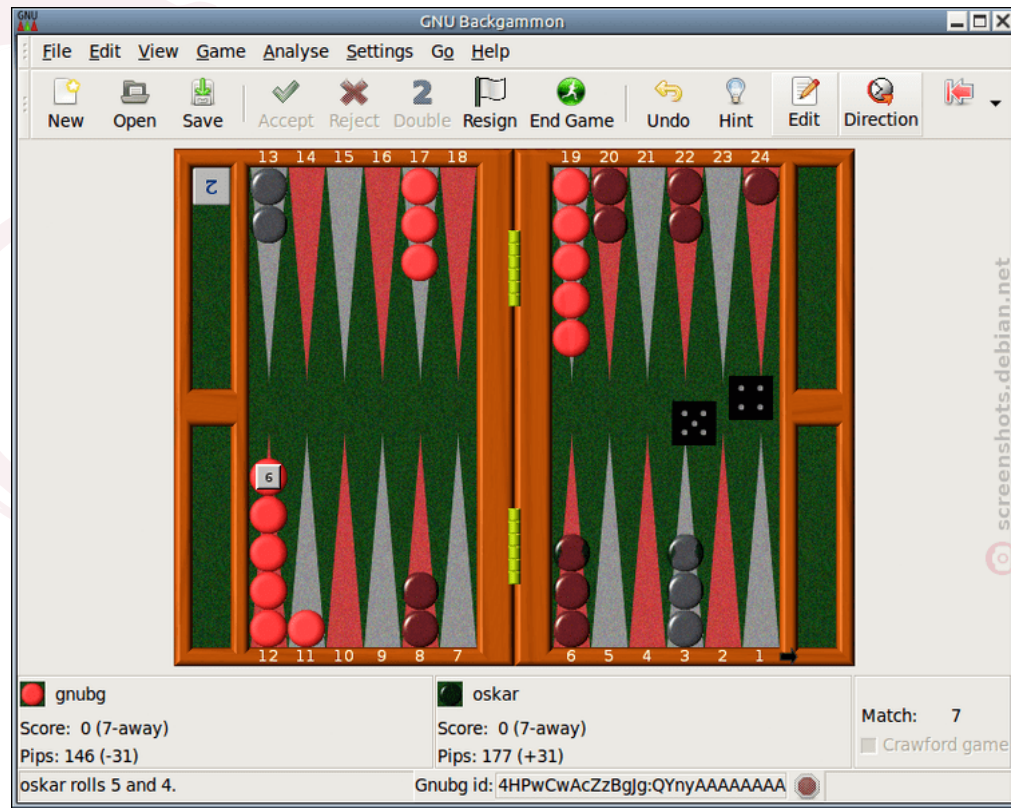
Building a software entails

- The source code
- A compiler
- The resulting program



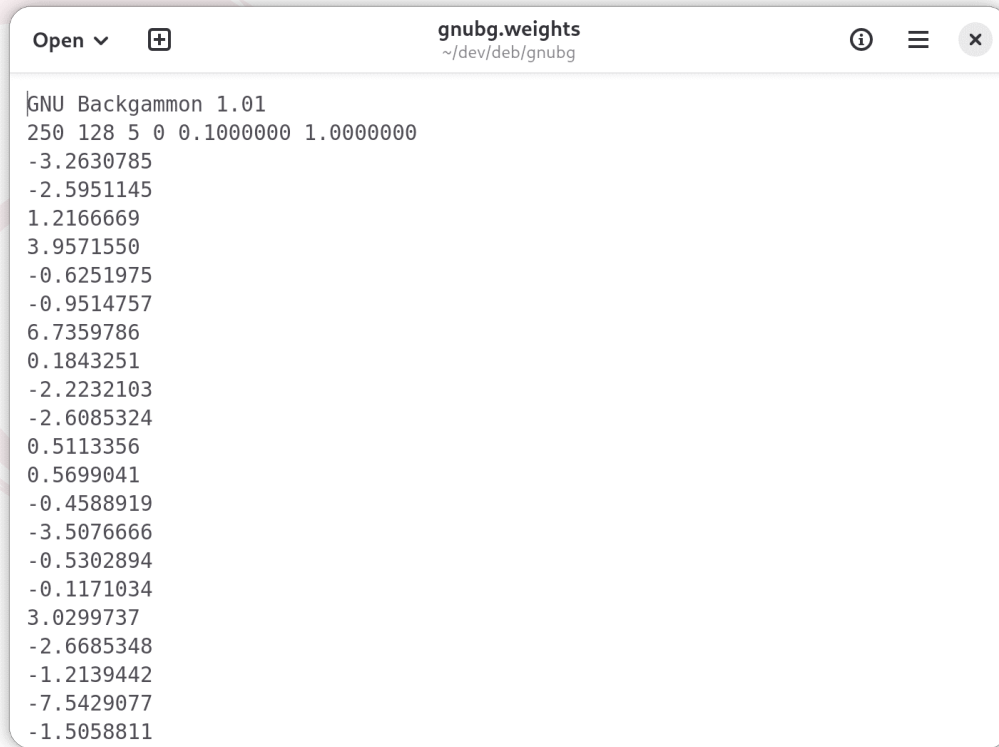
An example: gnubg

- A backgammon game and engine
- In Debian since 2001
- Can be used to play against the computer



An example: gnubg

- How can the computer play backgammon?
- What are all those numbers?
- Did someone write them all?



```
gnubg.weights
~/dev/deb/gnubg

GNU Backgammon 1.01
250 128 5 0 0.1000000 1.0000000
-3.2630785
-2.5951145
1.2166669
3.9571550
-0.6251975
-0.9514757
6.7359786
0.1843251
-2.2232103
-2.6085324
0.5113356
0.5699041
-0.4588919
-3.5076666
-0.5302894
-0.1171034
3.0299737
-2.6685348
-1.2139442
-7.5429077
-1.5058811
```

An example: gnubg

- For more than 20 years,
nobody cared
- With LLMs becoming more
relevant, old practice
is put under discussion

An example: gnubg

- For more than 20 years, nobody cared
- With LLMs becoming more relevant, old practice is put under discussion

gnubg, at least, comes with neural network weights that do not have source code under this definition. I have to admit that while I was maintaining the package I didn't give this a ton of a thought because it predates the whole LLM craze.

--

*Russ Alberry
<rra@debian.org>*

An example: gnubg

- For more than 20 years, nobody cared
- With LLMs becoming more relevant, old practice is put under discussion

*gnubg, at least, comes with neural network weights that do not have source code under this definition. I have to admit that while I was maintaining the package **I didn't give this a ton of a thought** because it predates the whole LLM craze.*

--

*Russ Alberry
<rra@debian.org>*

Definitions

- Boring way of spelling out our goals
- What's "source code" in a neural network?

Debian Social Contract

Version 1.2 ratified on October 1st, 2022.

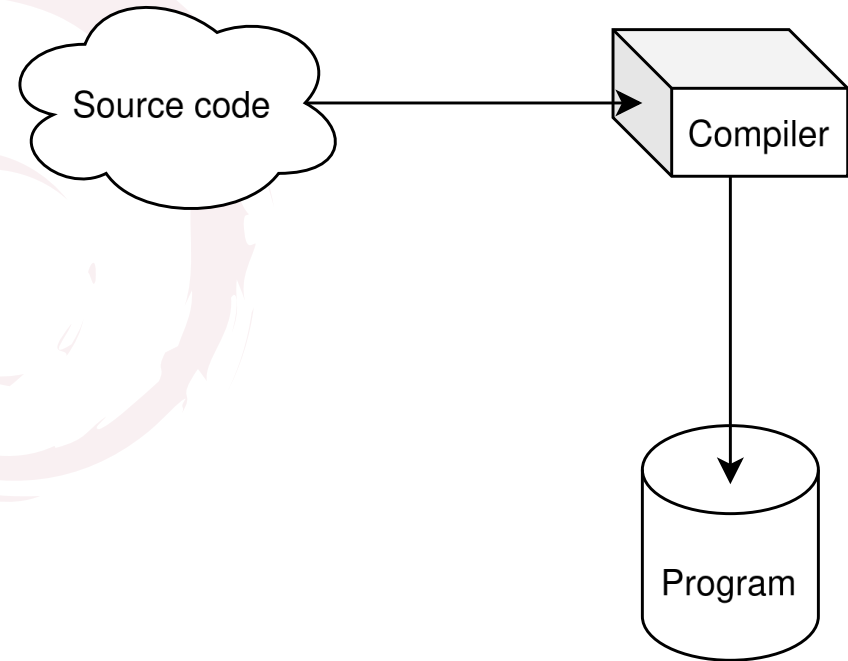
Supersedes [Version 1.1](#) ratified on April 26th, 2004, and [Version 1.0](#) ratified on July 5, 1997.

Debian, the producers of the Debian system, have created the **Debian Social Contract**. The [Debian Free Software Guidelines \(DFSG\)](#) part of the contract, initially designed as a set of commitments that we agree to abide by, has been adopted by the free software community as the basis of the [Open Source Definition](#).

Definitions

Building a software entails

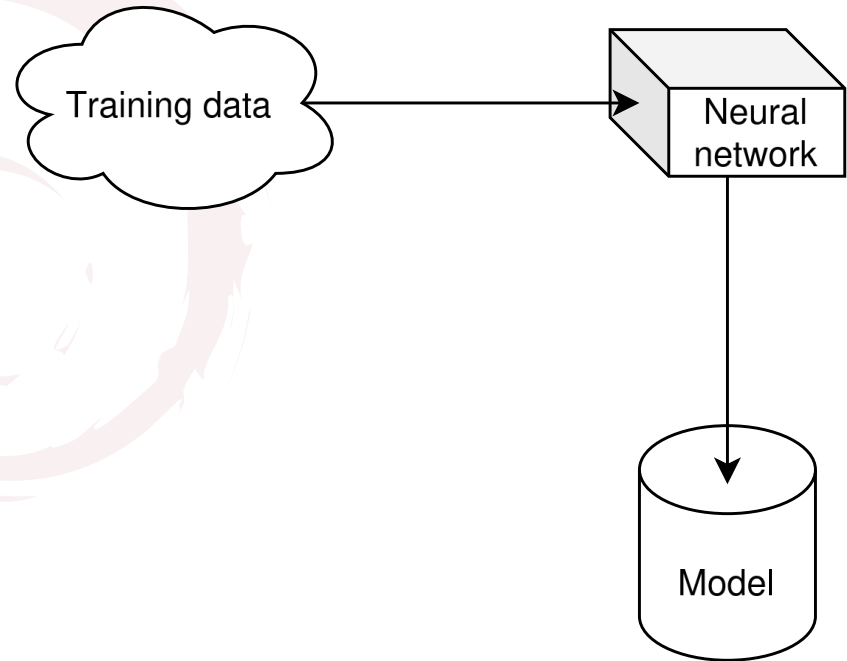
- The source code
- A compiler
- The resulting program



Definitions

Building an AI model entails

- Training data
- A neural network
- The resulting model



Definitions



Interpretation of DFSG on
Artificial Intelligence (AI)
Models

Definitions

Interpretation of DFSG on Artificial Intelligence (AI) Models

AI models released under DFSG-compatible license without original training data or program" are not seen as DFSG-compliant.

The "AI models released under DFSG-compatible license without original training data or program", a particular type of files as explained above, are not seen as DFSG-compliant. Hence, they can not be included in the "main" section of the Debian archive. This proposal does not specify whether the "non-free" section of Debian archive can include those files.

Definitions

The point is not the definition, but the thought behind it

Interpretation of DFSG on Artificial Intelligence (AI) Models

AI models released under DFSG-compatible license without original training data or program" are not seen as DFSG-compliant.

The "AI models released under DFSG-compatible license without original training data or program", a particular type of files as explained above, are not seen as DFSG-compliant. Hence, they can not be included in the "main" section of the Debian archive. This proposal does not specify whether the "non-free" section of Debian archive can include those files.

Consequences

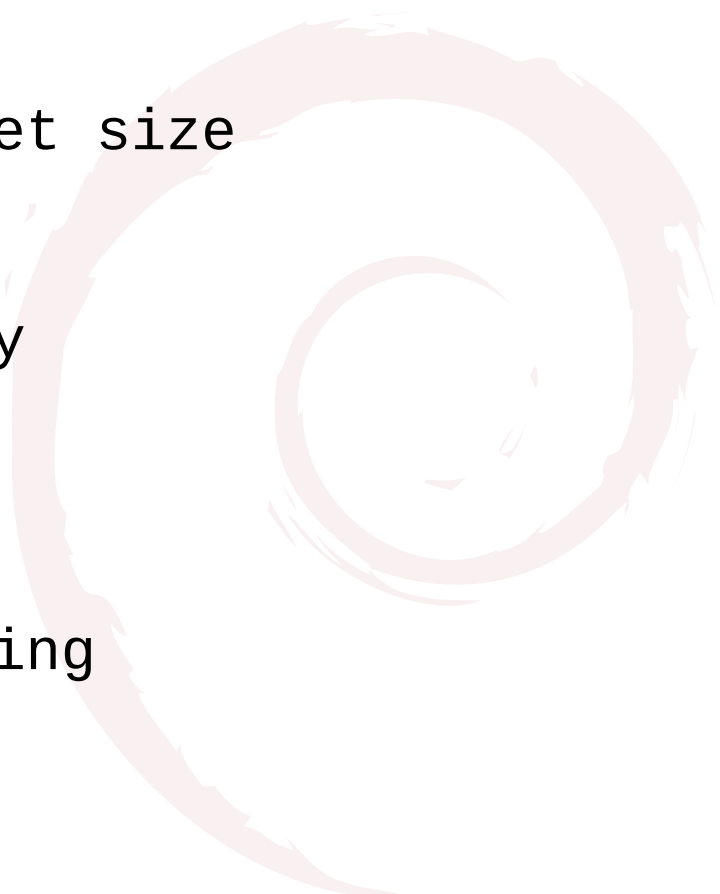
Requiring training data
has practical,
technical, and ethical
consequences

Interpretation of DFSG on Artificial Intelligence (AI) Models

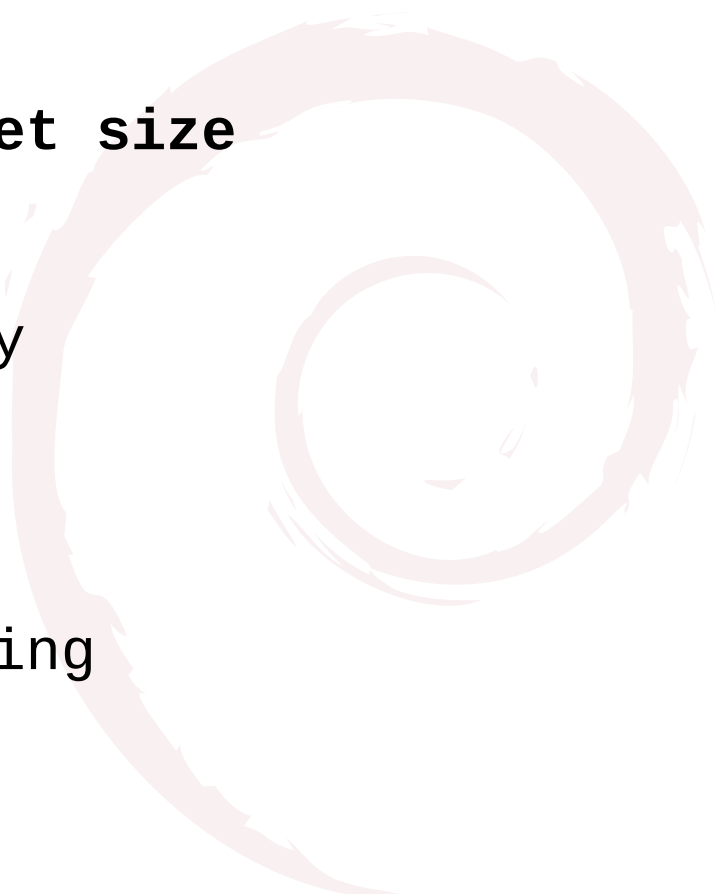
AI models released under DFSG-compatible license without original training data or program" are not seen as DFSG-compliant.

The "AI models released under DFSG-compatible license without original training data or program", a particular type of files as explained above, are not seen as DFSG-compliant. Hence, they can not be included in the "main" section of the Debian archive. This proposal does not specify whether the "non-free" section of Debian archive can include those files.

Technical consequences

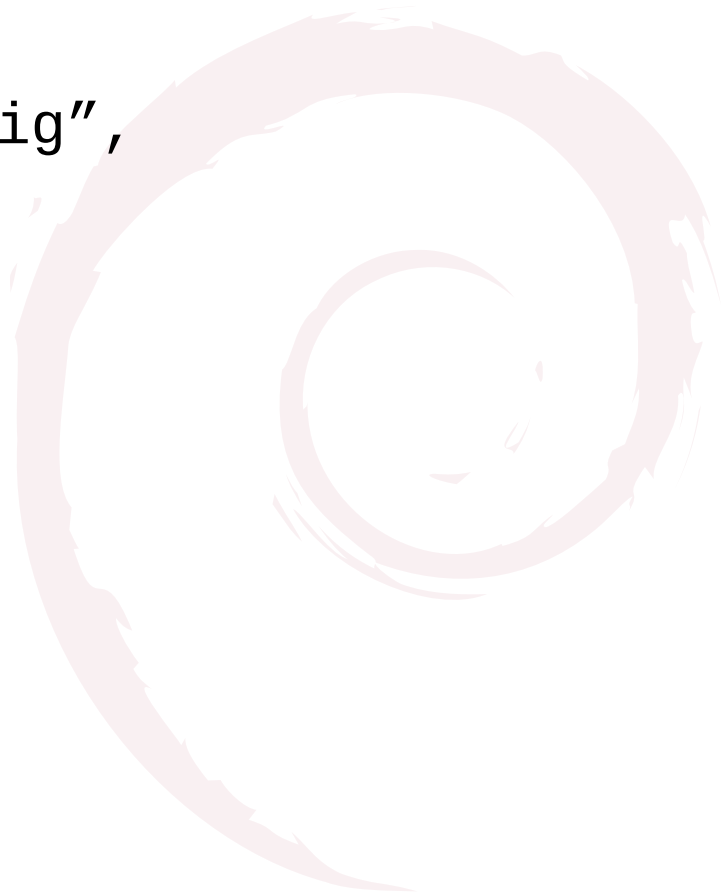
- Training dataset size
 - Training time
 - Reproducibility
 - Security
 - Transparency
 - Dataset licensing
- 

Technical consequences

- **Training dataset size**
 - **Training time**
 - Reproducibility
 - Security
 - Transparency
 - Dataset licensing
- 

Technical consequences

- Data is “too big”,
LLMs are out



Technical consequences

- Data is “too big”, LLMs are out ...for now?
- The llm.c project trained GPT-2 (124M) in 20 minutes
- The original dataset was 40 GB, the largest package is 1 GB
- TinyStories is ~2GB



Andrej Karpathy

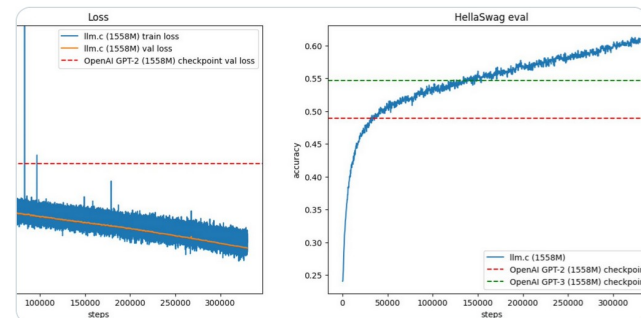
@karpathy · Follow



In 2019, OpenAI announced GPT-2 with this post:
[openai.com/index/better-l...](https://openai.com/index/better-l-...)

Today (~5 years later) you can train your own for ~\$672, running on one 8XH100 GPU node for 24 hours. Our latest llm.c post gives the walkthrough in some detail:
github.com/karpathy/llm.c...

Incredibly, the [Show more](#)



8:26 PM · Jul 11, 2024



6.3K



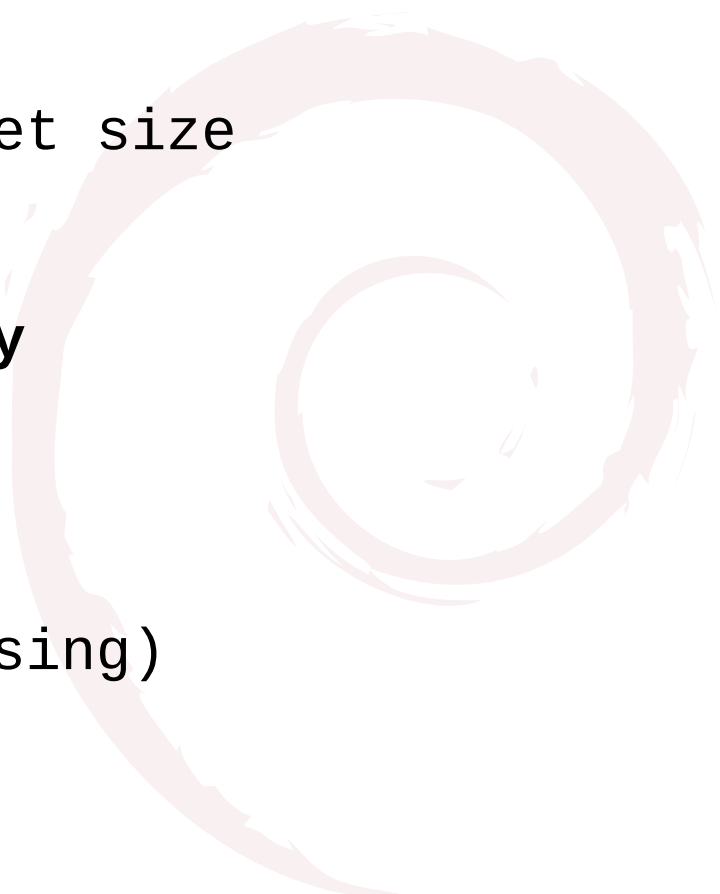
Reply



Copy link

[Read 125 replies](#)

Technical consequences

- Training dataset size
 - Training time
 - **Reproducibility**
 - **Security**
 - Transparency
 - (Dataset licensing)
- 

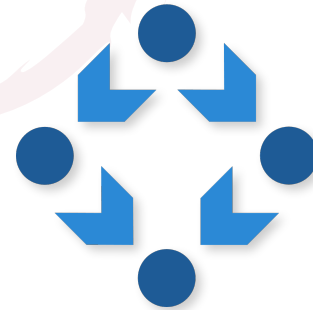
Technical consequences

- *Secure code delivery is the problem of getting software from its author to its users safely, with a healthy dose of mistrust towards the author and everything else in between*

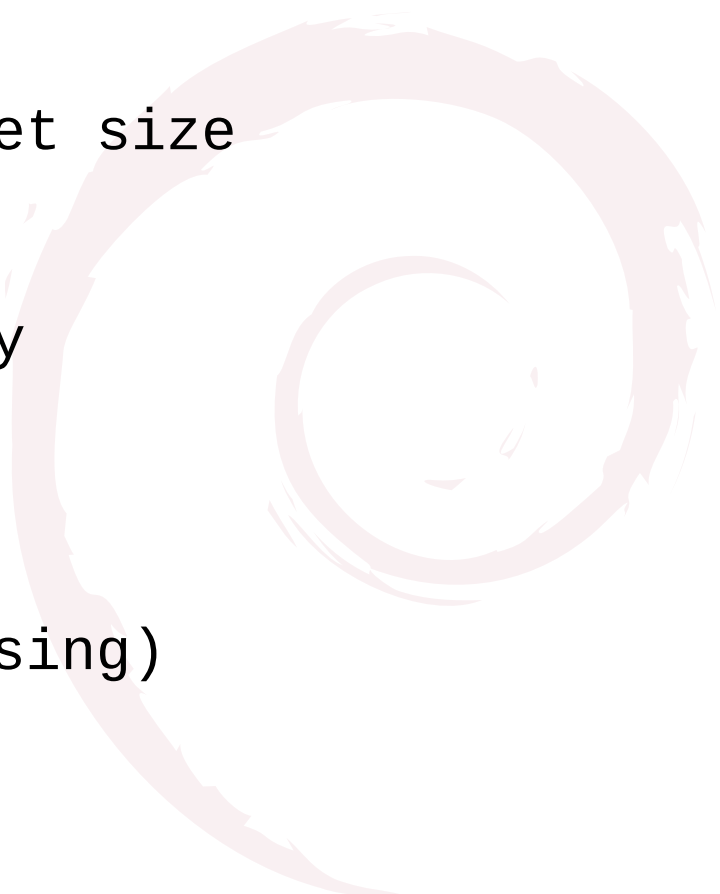
Technical consequences

- *Secure code delivery is the problem of getting software from its author to its users safely, with a healthy dose of mistrust towards the author and everything else in between*
- Build reproducibility is one vertex of the triangle

Reproducible Builds

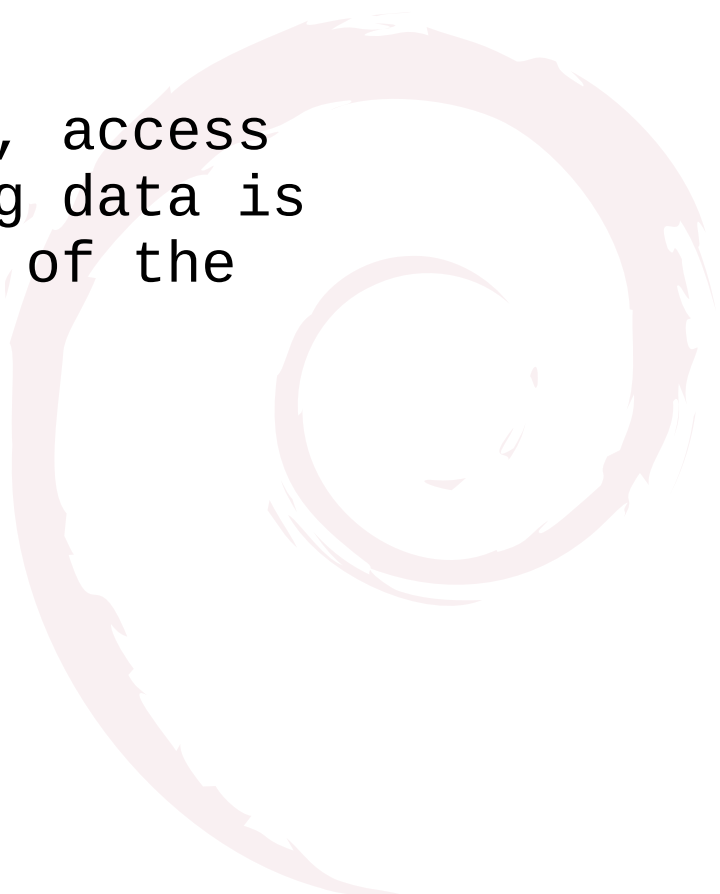


Technical consequences

- Training dataset size
 - Training time
 - Reproducibility
 - Security
 - **Transparency**
 - (Dataset licensing)
- 

Technical consequences

- In some fields, access to the training data is a crucial part of the model's value



Technical consequences

- In some fields, access to the training data is a crucial part of the model's value

I encountered an example wherein an AI model identified a previously unknown biomarker associated with cancer. This discovery was only possible because the researchers had access to the underlying dataset. Without that access, the model's findings would have been unverifiable.

--

*Arian Ott
<arian.ott@ieee.org>*

Technical consequences

- In some fields, access to the training data is a crucial part of the model's value

*I encountered an example wherein an AI model identified a previously unknown biomarker associated with cancer. This discovery was **only possible because the researchers had access to the underlying dataset**. Without that access, the model's findings would have been unverifiable.*

--

*Arian Ott
<arian.ott@ieee.org>*

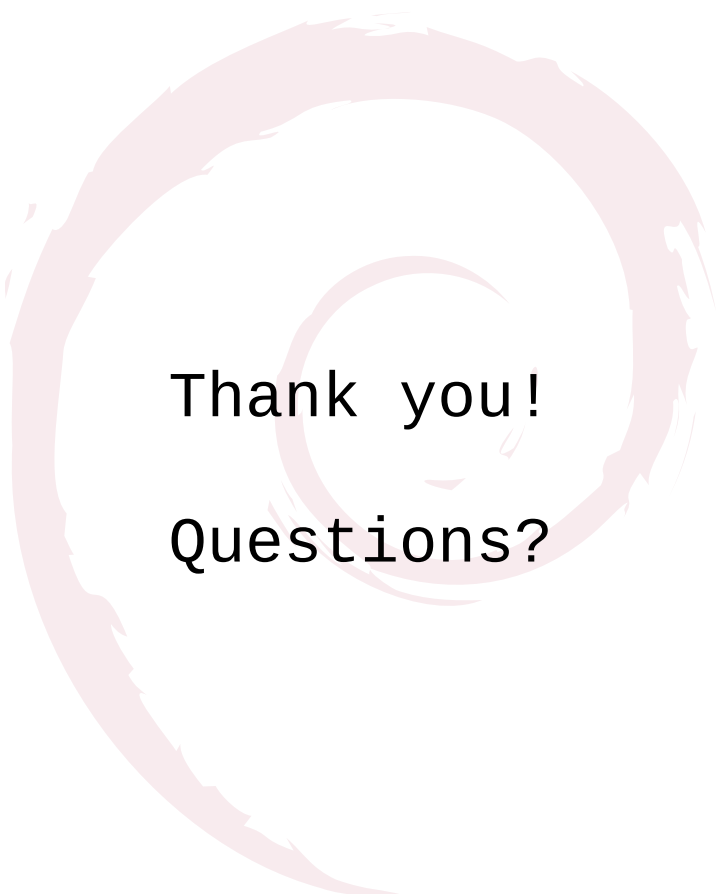
Ethical consequences

- Dataset licensing (!)
 - Is copyright too strict?
- Bias of the model
- Power dynamics
 - Is it fair for the author only to know the data?

Takeaways

- LLMs are currently too large to be re-trained by Debian
 - Not all AI models are LLMs!
- Our priorities are our users and free software, not new technology
- We should focus on distributing AI models which align with our values

Why you cannot “apt install” an LLM (yet)



Thank you!

Questions?