Les LLMs pour les nuls

(Aka faire mieux que
 GPT4 avec des modèles
In-house)

crisp

Crisp: le support client du futur

- 600 000 entreprises
- 100M+ de personnes communiquent chaque mois
- Plus de 100 pays
- Créé en 2015

Crisp: le support client du futur

Centralisation
des messages

Collaboratif

Data Unifiée

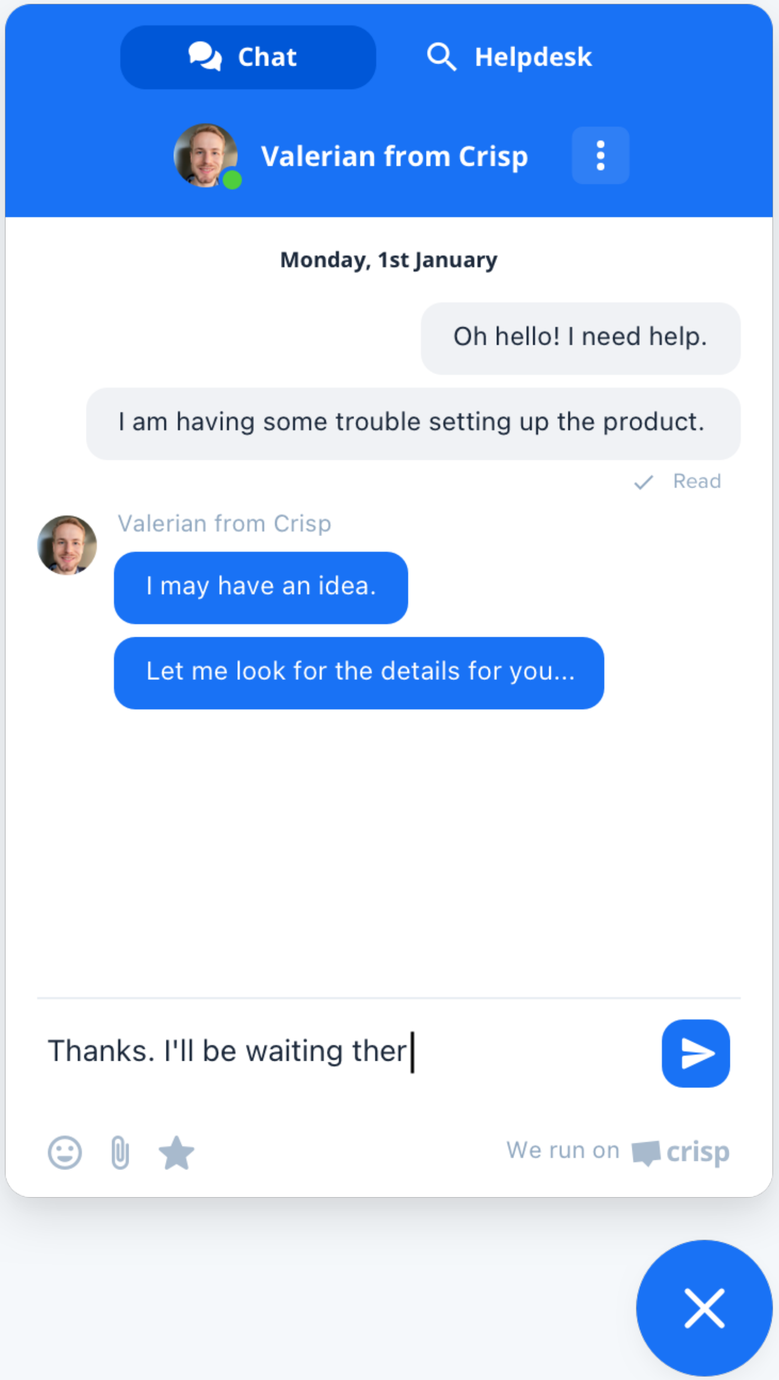

Rakuten


DECATHLON


N26


RENAULT

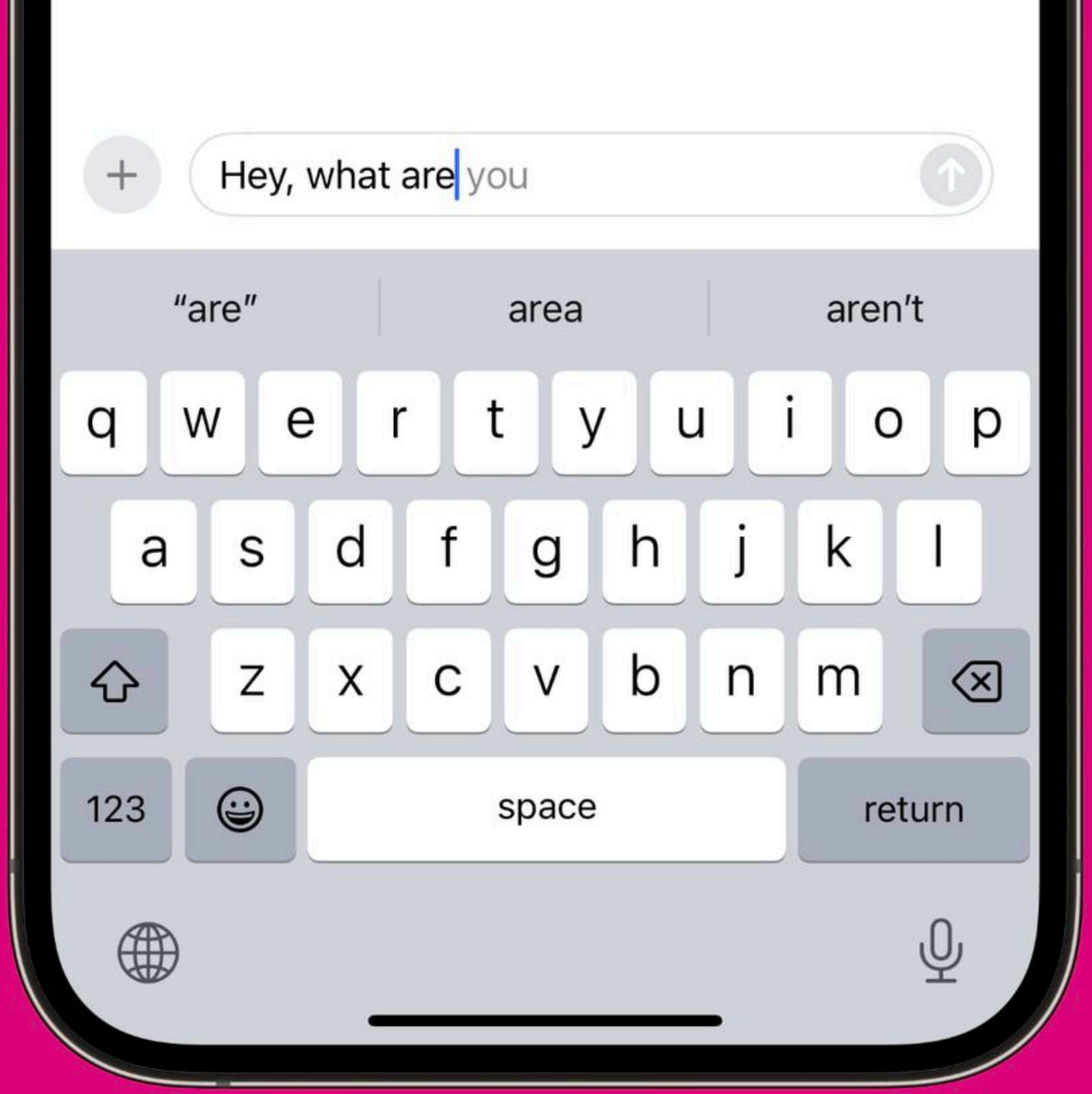
Livechat


Chat
Helpdesk
Valerian from Crisp
Monday, 1st January
Oh hello! I need help.
I am having some trouble setting up the product.
Read
Valerian from Crisp
I may have an idea.
Let me look for the details for you...
Thanks. I'll be waiting ther
We run on crisp

Multicanal

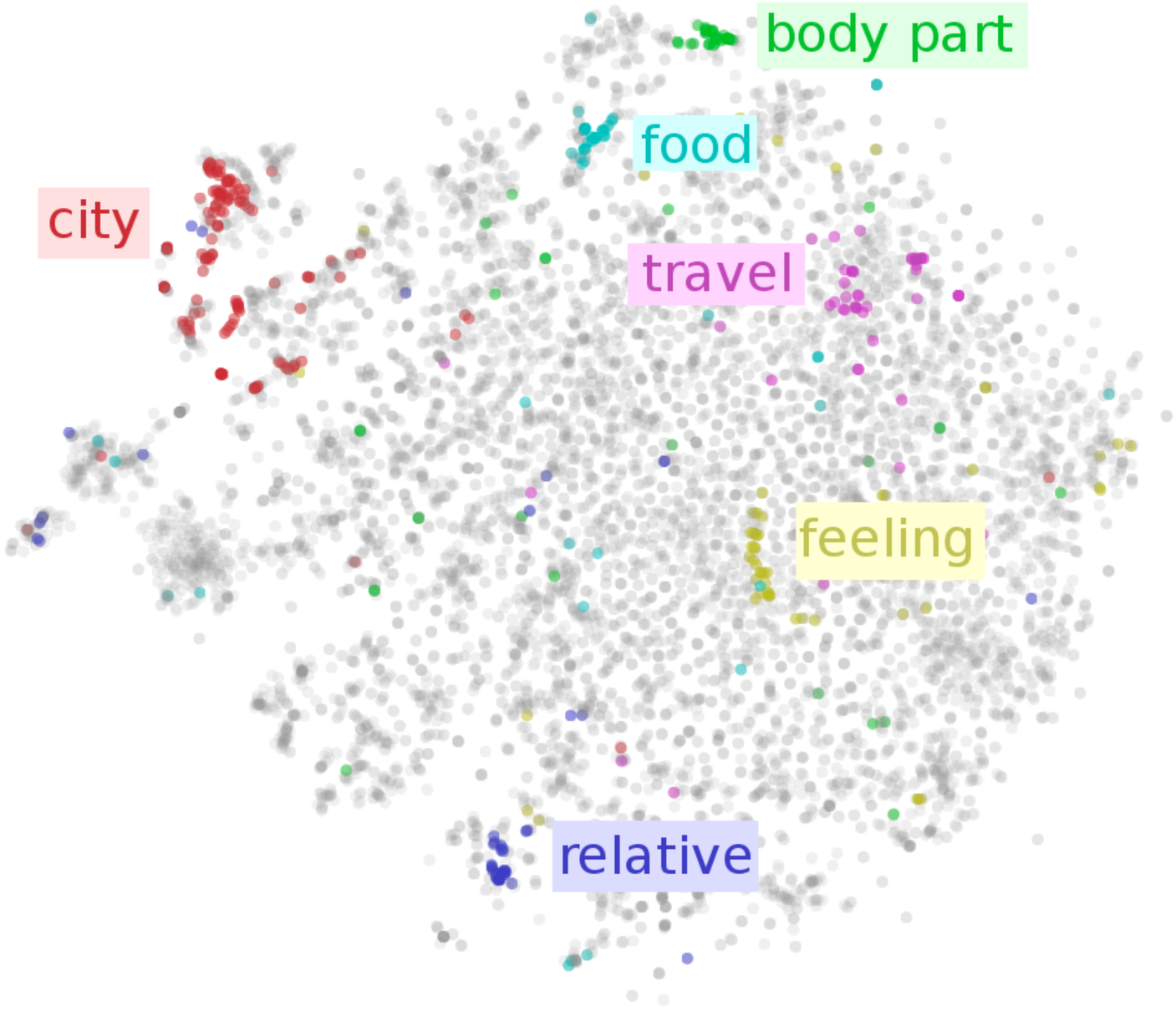

All
Unread
Unresolved
Filters
Walter › Antoine
21h
Hey Walter, It's Antoine from Crisp. Following back on what
Charlotte › Antoine
Yesterday
Sure, let's go for it
Antoine
19 Sep
Perfect! Someone in our team will be back to you.
visitor59 › Antoine
Yesterday
Hang on a minute
Emma › Antoine
Yesterday
Sure, Let me tell you about what we offer
Leo › Antoine
29 Aug
I will Have a nice day!
Edwin › Antoine
27 Aug
Hello ! Our shipping delays are between two and three days.
Hello! I think I fou
Hello ! I'm sorry what happens?
Sure. Taking a sc
MEN'S RUNNING SHOES
Hey Walter, It's
Reply
Edit
Note
Send your message to

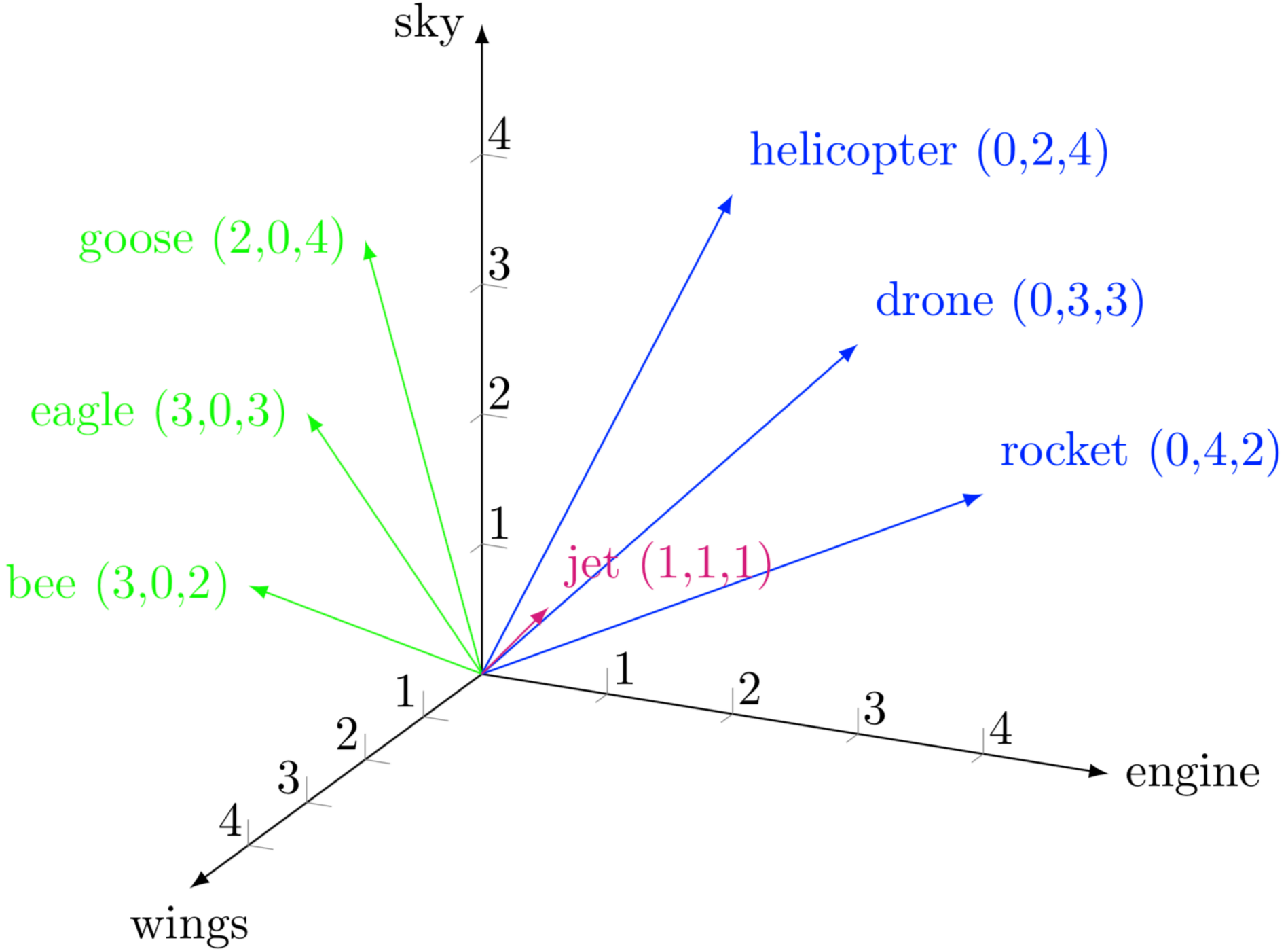
C’est quoi un LLM


Hey, what are you
“are”
area
aren’t
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
space
return

Années 2000

- Word Embeddings
- Convertir des idées en vecteurs
- Permet de rechercher et regrouper des corpus ayant des significations proches


body part
food
city
travel
feeling
relative


sky
4
3
2
1
helicopter (0,2,4)
drone (0,3,3)
rocket (0,4,2)
goose (2,0,4)
eagle (3,0,3)
bee (3,0,2)
jet (1,1,1)
1
2
3
4
engine
1
2
3
4
wings

Années 2010

- Embedding sur des phrases/corpus
- Transformers
- Attention is all you need (2017)
- GPT2 (2019)

Années 2010

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

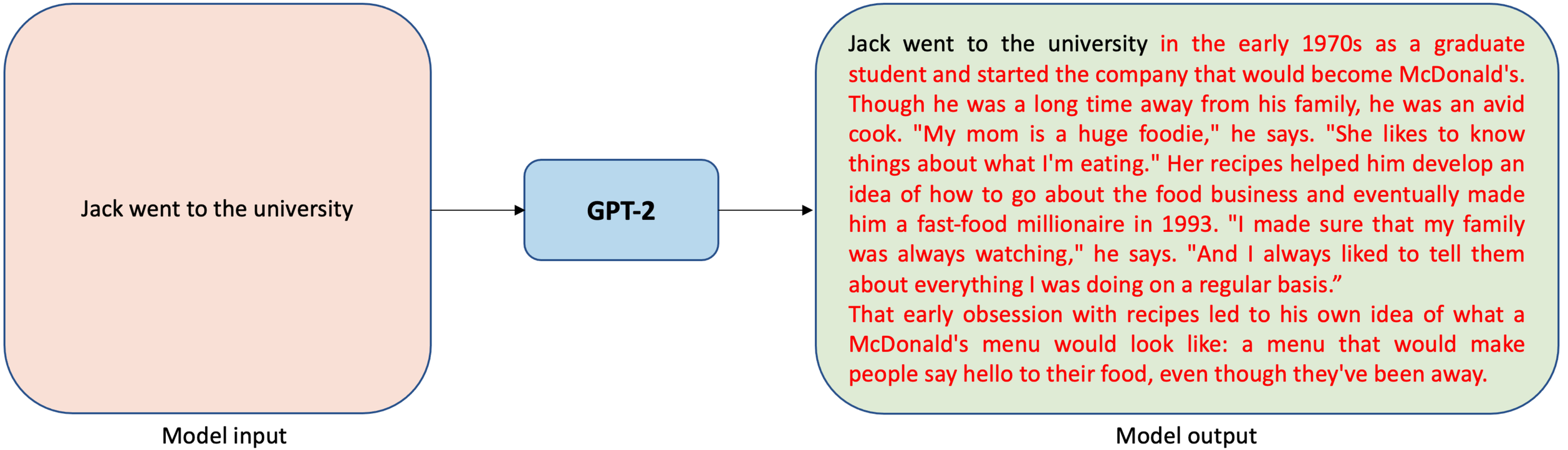
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

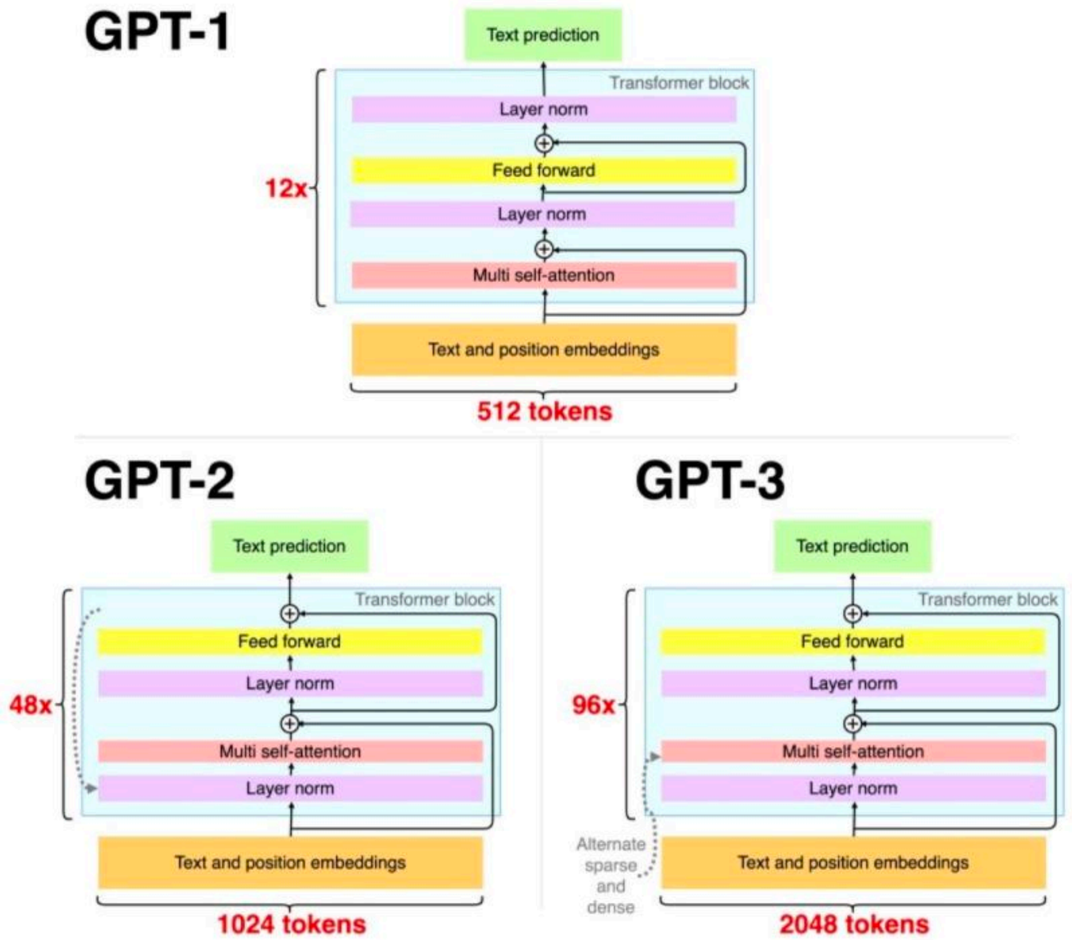
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including

Années 2010


Jack went to the university
GPT-2
Jack went to the university in the early 1970s as a graduate student and started the company that would become McDonald's. Though he was a long time away from his family, he was an avid cook. "My mom is a huge foodie," he says. "She likes to know things about what I'm eating." Her recipes helped him develop an idea of how to go about the food business and eventually made him a fast-food millionaire in 1993. "I made sure that my family was always watching," he says. "And I always liked to tell them about everything I was doing on a regular basis."
That early obsession with recipes led to his own idea of what a McDonald's menu would look like: a menu that would make people say hello to their food, even though they've been away.
Model input
Model output

Années 2010

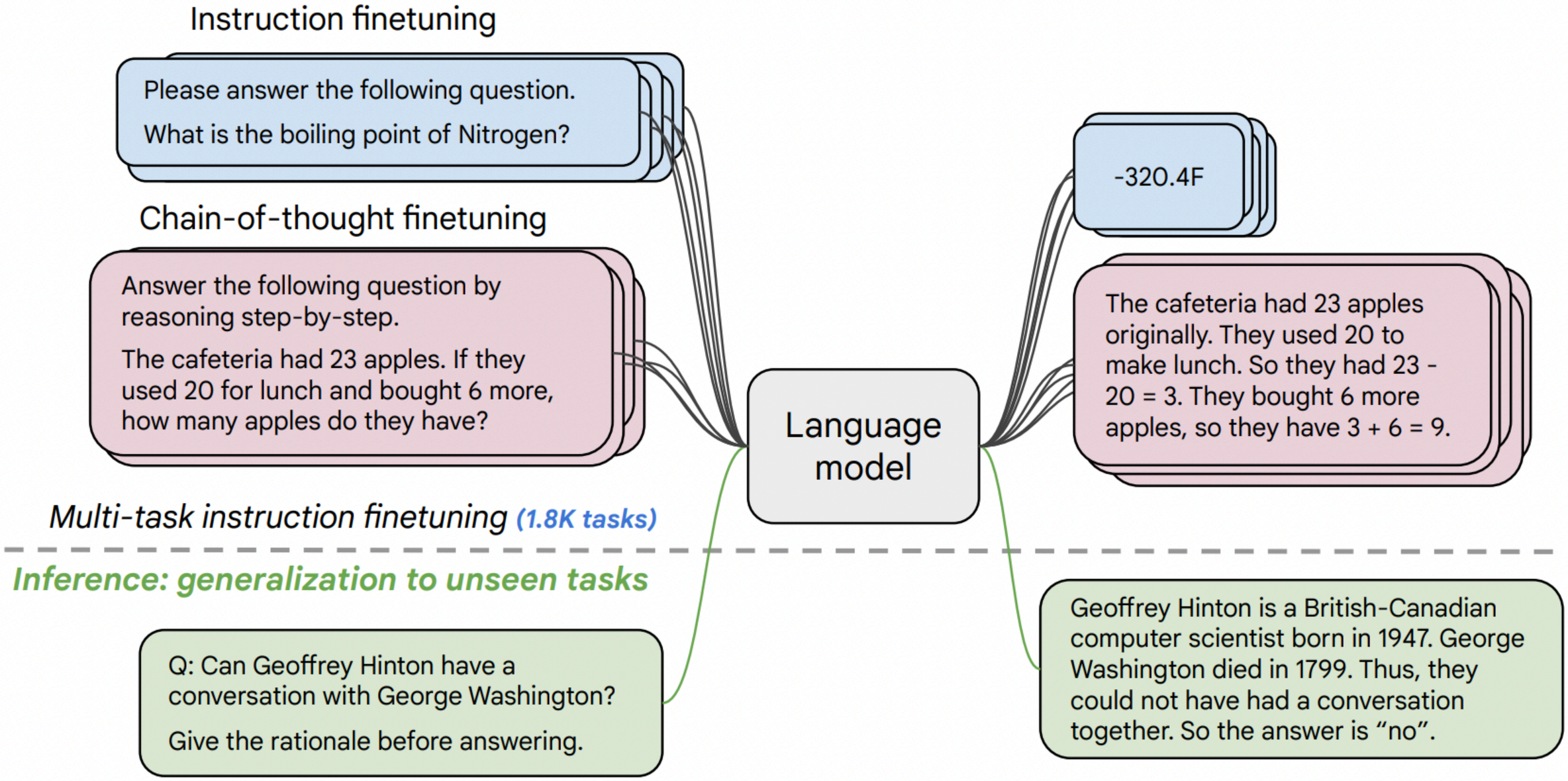
GPT-1 vs GPT-2 vs GPT-3



Années 2020

- GPT3 (2020)
- GPT Instruct (GPT Davinci 003, 3.5, 5)
- Flan T5
- Modeles open source (LLama, GPT NEO, MPT, Falcon, Mistral, ...)

La révolution des modèles à instruction


Instruction finetuning
Please answer the following question.
What is the boiling point of Nitrogen?
Chain-of-thought finetuning
Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?
Multi-task instruction finetuning (1.8K tasks)
Language model
-320.4F
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.
Inference: generalization to unseen tasks
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is “no”.

Single-shot learning

Classifie un texte en utilisant les categories suivantes:

Cuisine, Sport, Aviation, …

Texte: “Le PSG gagne 3-0 contre Monaco”

Categorie:

Sport

Zero-shot learning

Classifie ce titre

Texte: “L’allocution télévisée de emmanuel macron”

Politique

Categorie:

Few-shot learning

Classifie un texte en utilisant les categories suivantes:

Cuisine, Sport, Aviation, …

Exemples:

Texte: “Le PSG gagne 3-0 contre Monaco”

Categorie: Sport

Texte: “La recette de cuisine de la purée aux pommes de terre:

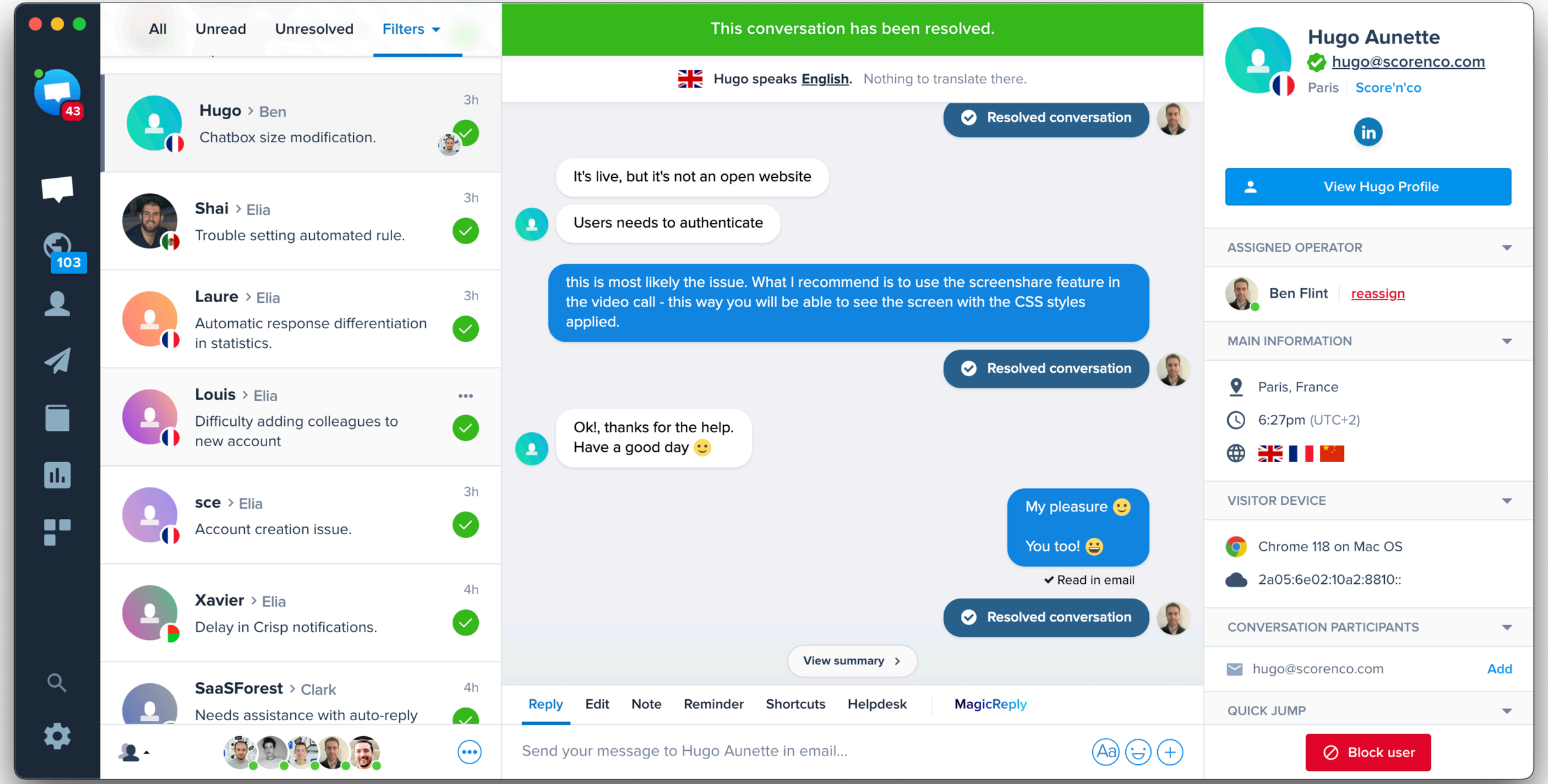
Categorie: Cuisine

Sport

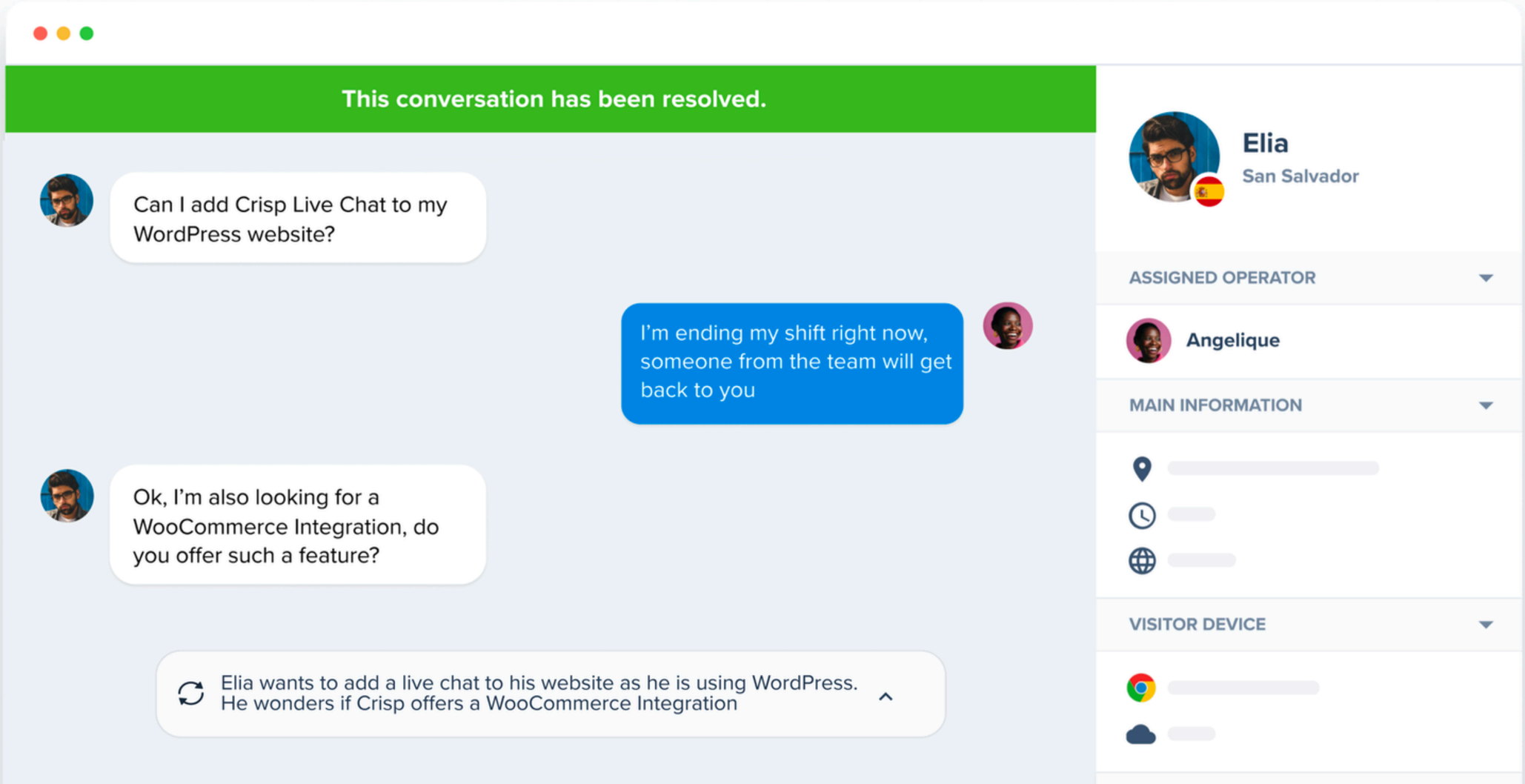
Instruct GPT

Use Case	Example
chat	The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly. Human: Hello, who are you? AI: I am an AI created by OpenAI. How can I help you today? Human: I'd like to cancel my subscription. AI:
chat	Marv is a chatbot that reluctantly answers questions with sarcastic responses: You: How many pounds are in a kilogram? Marv: This again? There are 2.2 pounds in a kilogram. Please make a note of this. You: What does HTML stand for? Marv: Was Google too busy? Hypertext Markup Language. The T is for try to ask better questions in the future. You: When did the first airplane fly? Marv:
chat	This is a conversation with an enlightened Buddha. Every response is full of wisdom and love. Me: How can I achieve greater peace and equanimity? Buddha:
closed qa	Help me answer questions about the following short story: {story} What is the moral of the story?

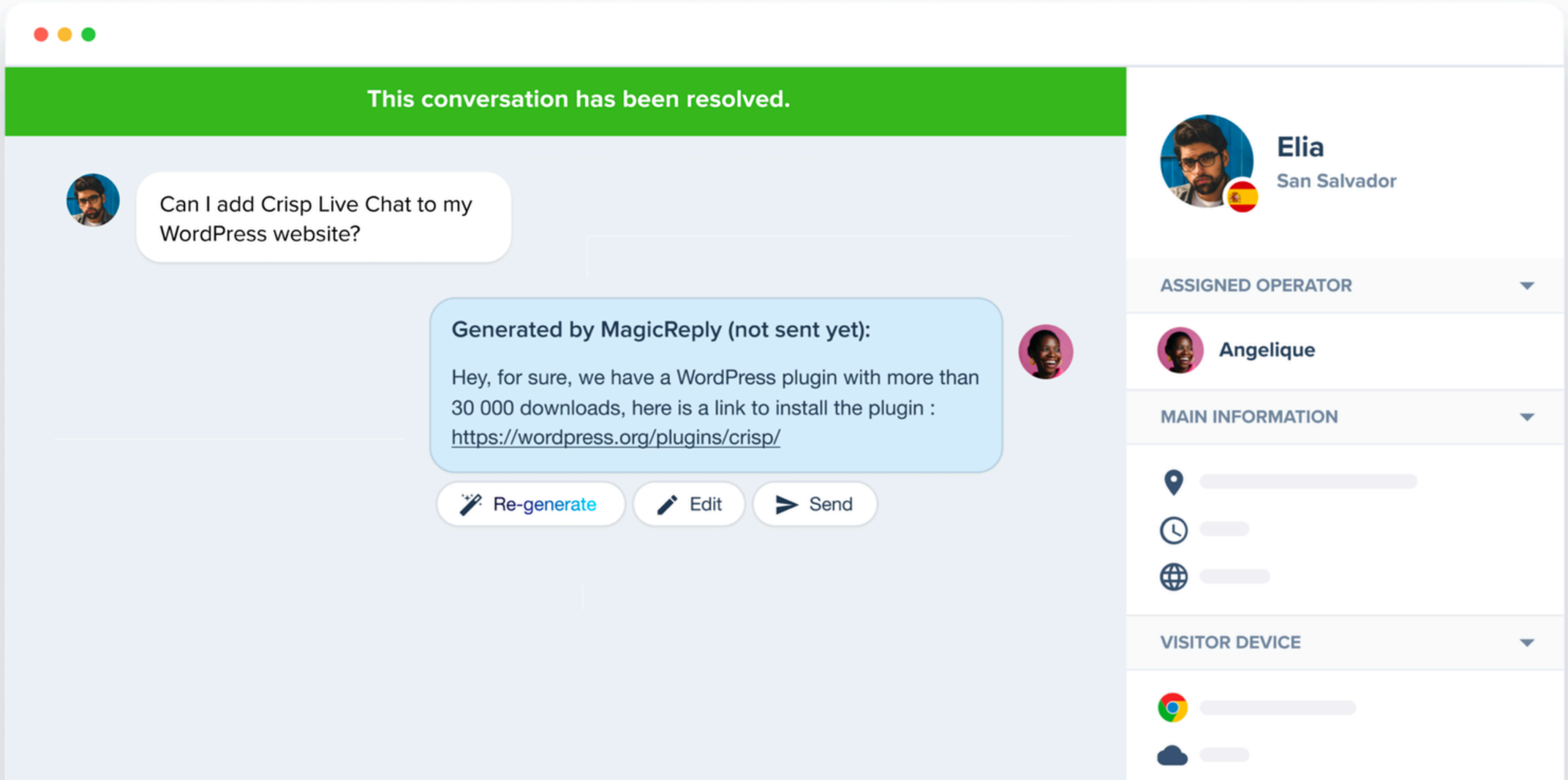
Quelques use-case Pour Crisp

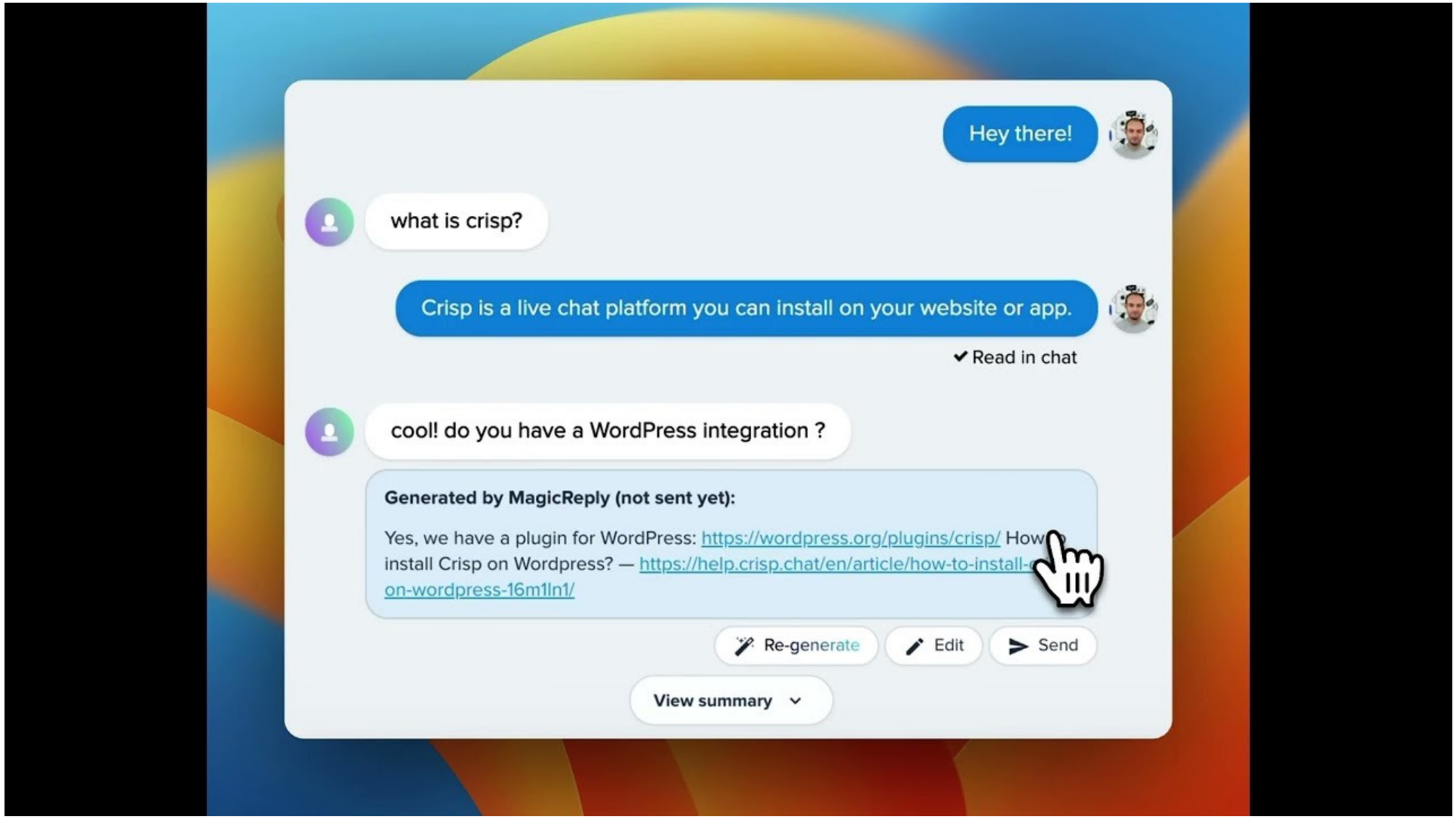

All
Unread
Unresolved
Filters
This conversation has been resolved.
Hugo › Ben
Chatbox size modification.
3h
Shai › Elia
Trouble setting automated rule.
3h
Laure › Elia
Automatic response differentiation in statistics.
3h
Louis › Elia
Difficulty adding colleagues to new account
sce › Elia
Account creation issue.
3h
Xavier › Elia
Delay in Crisp notifications.
4h
SaaSForest › Clark
Needs assistance with auto-reply
4h
Hugo speaks English. Nothing to translate there.
Resolved conversation
It's live, but it's not an open website
Users needs to authenticate
this is most likely the issue. What I recommend is to use the screenshare feature in the video call - this way you will be able to see the screen with the CSS styles applied.
Resolved conversation
Ok!, thanks for the help. Have a good day
My pleasure
You too!
Read in email
Resolved conversation
View summary
Reply
Edit
Note
Reminder
Shortcuts
Helpdesk
MagicReply
Send your message to Hugo Aunette in email...
Hugo Aunette
hugo@scorenco.com
Paris | Score'n'co
View Hugo Profile
ASSIGNED OPERATOR
Ben Flint | reassign
MAIN INFORMATION
Paris, France
6:27pm (UTC+2)
VISITOR DEVICE
Chrome 118 on Mac OS
2a05:6e02:10a2:8810::
CONVERSATION PARTICIPANTS
hugo@scorenco.com
Add
QUICK JUMP
Block user

Summarize conversations
and make shifts easiers between your teams



Empower your teams with AI-generated answers




Hey there!
what is crisp?
Crisp is a live chat platform you can install on your website or app.
Read in chat
cool! do you have a WordPress integration ?
Generated by MagicReply (not sent yet):
Yes, we have a plugin for WordPress: https://wordpress.org/plugins/crisp/ How to install Crisp on Wordpress? — https://help.crisp.chat/en/article/how-to-install-crisp-on-wordpress-16m1ln1/
Re-generate
Edit
Send
View summary

Mirage – Augment your apps

mirage-ai.com

Mirage™

Support Docs Login to Dashboard

Augment your apps with AI

Mirage is an API for everything AI. Integrate in minutes.

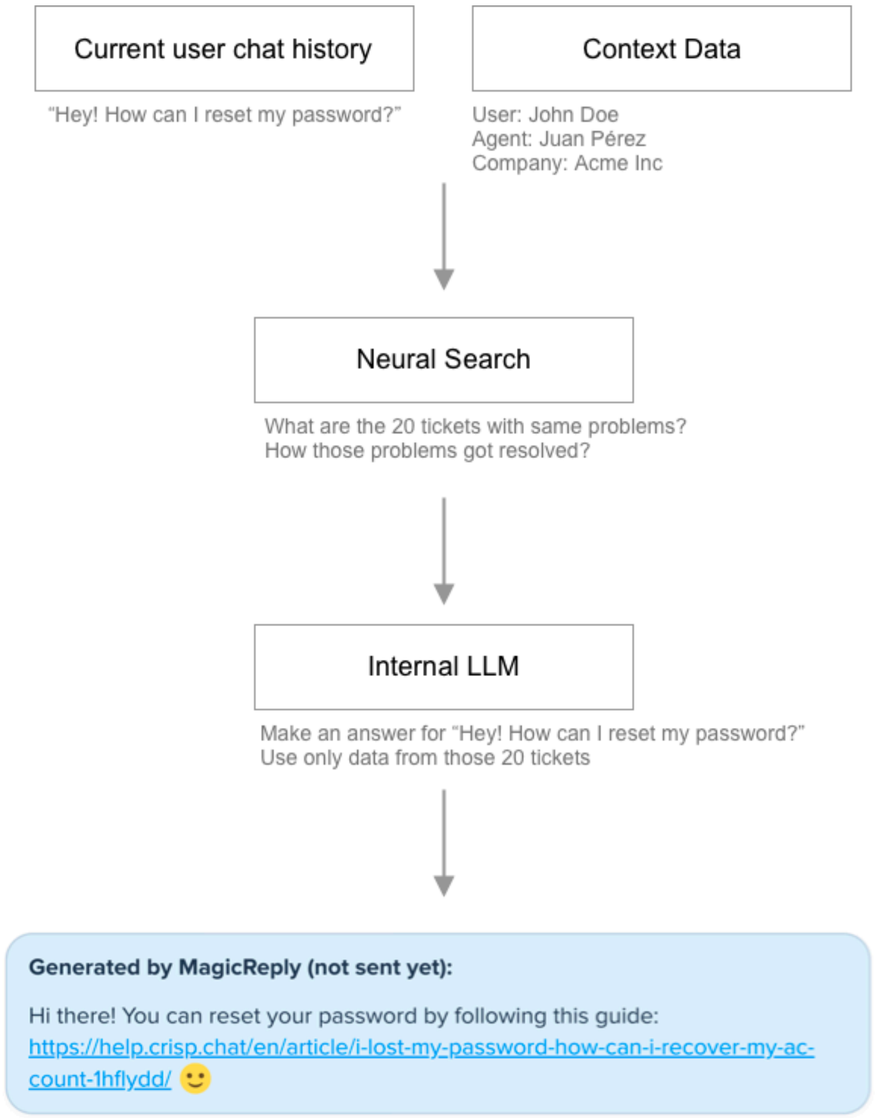
Text translation, answering, speech-to-text & more.

Get your Mirage API Key →

Try for free, no credit card required.

- Transcribe Speech to Text — Try me! →
- Answer Question in a Chat — Try me! →
- Summarize a Conversation — Try me! →
- Categorize a Conversation — Try me! →

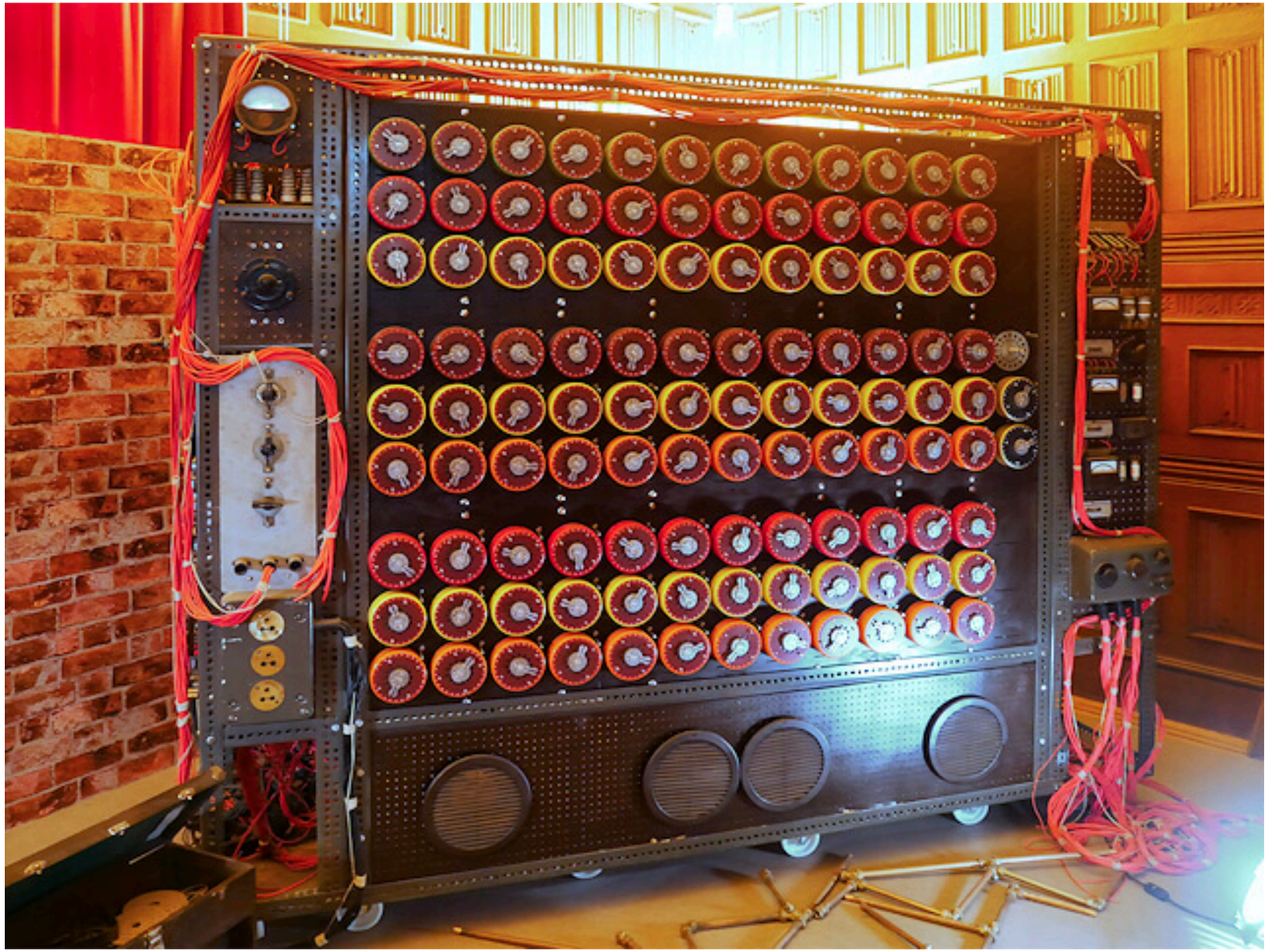
Créer un LLM pour résoudre des taches



Remplacer GPT4 par des modèles OSS

Fine tuner c’est quoi?

- Apprendre à un modele pré-entraine a faire une tache
- On donne au modele n-exemples
- On le “Brute force”
- Le modèle se spécialise à résoudre cette tache



Step 1: Labelling

- Le robot c’est toi
- On résout la tache sois même
- 1000 exemples environ

#1
1 of 1

[1]: 7/10

You are a helpful Bot that takes part of Crisp's team who help the support agent by providing helpdesk articles that can provide answers to the user problems.
[HELPDESK]
[1] How to add a live chat to my website?
website, chat with us. We will be happy to help! You can also go to our website chat widget feature page that describes everything you need to know, in terms of price or possibities. How to create an account on Crisp Simply reach out to https://app.crisp.chat/initiate/signup/ and follow the process

---

[2] How to use the Crisp Inbox for the first time
Take advantage of a dedicated video that gives you an overview of Crisp and how to use it for the first time. You just joined a team using Crisp? Welcome onboard! We made a video tutorial to help you using Crisp. Looking at this video will help you to understand

---

[3] How to edit my profile and add a profile picture?
Completing your Crisp profile helps your teammates and customers to learn more about you. Here's how to add details about yourself, upload a photo, and set your language Go to https://app.crisp.chat Reach out to your settings Click on "Account" Edit your information Here is the menu where you can edit

---

[4] How to install Crisp Live chat on WordPress?
https://help.crisp.chat/en/article/how-to-add-the-crisp-chatbox-code-to-my-website-10wcj3l/
Installing Crisp Live chat widget on WordPress is really easy! || If you use Woocommerce and would like to see your customers carts and orders, see our Woocommerce installation guide here To check details on our WooCommerce helpdesk, click here. Video tutorial: ${youtube}Video tutorial on how to install Crisp on WordPress Want to know more about Crisp features and pricing? Click here to see the detail: live chat widget Instructions: First, you have to connect on your WordPress dashboard You have to click on Plugins / Add New and search "Crisp" Activate the Plugin On the left menu, click "Crisp

---

Skip
Submit

dataset_final.jsonl

```
{
    "text": "Agent: Malheureusement ce n'est pas possible de cr\\u00e9er tel bouton directement avec nos outils, vous aurez besoin d'un d\\u00e9veloppeur pour cr\\u00e9er ce bouton personnalis\\u00e9.\nN'h\\u00e9sitez pas si vous avez d'autres questions!\n\nAgent: Bonjour, merci de votre retour!\n\nCustomer: Bonjour j'ai trouv\\u00e9 comment d\\u00e9plac\\u00e9 l'icone message mais est-ce que je peux la mettre sur le cot\\u00e9 droit comme sur la photo d'avant ?\n\nAgent: Bonjour! Notre \\u00e9quipe de Fran\\u00e7ais est actuellement indisponible, est-ce que \\u00e7a va si je continue avec notre traducteur ?\n\nCustomer: Je ne trouve pas l'onglet position ?\n\nAgent: Vous pouvez personnaliser la position de la bulle de chat avec le plugin Customization:\n\nhttps://help.crisp.chat/en/article/how-to-customize-my-chatbox-1hj537j/\n\nSous \"Positions\"\n\nCustomer: et aussi, comment faire pour que l'icone de discussion ne soit pas en bas \\u00e0 droite mais sur le cot\\u00e9 \\u00e0 droite ( photo ci joint)\n\nCustomer: ok merci\n\nAgent: Le trigger ne va pas avertir parce que la conversation n'es pas vraiment commenc\\u00e9e, donc pas d'affichage ni de 1 jusqu'as que l'utilisateur ne r\\u00e9pond pas\n\nCustomer: Non un petit \"1\" pour dire 1 nouveau message, comme une notification\n",
    "target": "Customizing chat bubble location."
}
```

Step 2: Fine-tuning

- Choix du modele: Plus c’est gros + c’est facile!
- Il faut un ou plusieurs GPU
- Plus facile avec des modèles encoder/ decoder
- Flan T5 XXL (11 milliards de params)
- Auto-train Hugginface ou Lora / Full fine tuning


Mistral
Falcon
Llama 2
AI Devs
Flan T5

Step 3: L’evalution

- On compare les outputs des modèles
- On les note

Step 4: Ameliorer

- Distillation
- RLHF

Distillation

Step 1:
On faire 1000 exemples a la main

Step 2:
On apprend à un gros modèle (Flan XXL) à soudre la tache

Step 3:
On génère 40 000 prompts et on les infère avec Flan XXL

Step 4:
On fine tune un petite modèle avec les 40k exemples de XXL

Distillation

GPT3.5: Qualité 67%
GPT4: Qualité 73%
Modèle XXL: Qualité 78%
Modèle Base (10 fois + petit): Qualité 81%

Step 5: Plus de data!

- Il est important de logger tous les prompts et leurs output
- Loggez et filtrez vous outputs GPT

Optimiser les modèles
Pour la production

GPU

- Un GPU est x50 + performant
- Quantisation
- Utilisation d’un inference engine natif
- Batching

Inference engines

- Transformers
- TS_Server
- GGML (Llama.cpp)
- Ctranslate2 (parfait pour des plus petit modèles)


Merci!
baptiste@crisp.chat