Data Engineering and Semantics
هندسة البيانات و دلالاتها

WIKIMEDIA
TN User Group

SCAN ME

Empowering Wikidata and Wikipedia with Generative AI

Unlocking the Potential of Scholarly Publications

WIKIMANIA
SINGAPORE

Houcemeddine Turki
User:Csisc

Disclosure

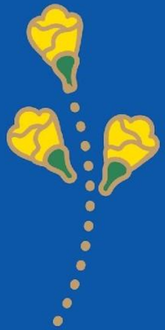
This work is a part of the “Adapting Wikidata to support clinical practice using Data Science, Semantic Web and Machine Learning” Project, funded by the Wikimedia Research Fund of Wikimedia Foundation.

WIKIMEDIA
FOUNDATION

This is a lecture

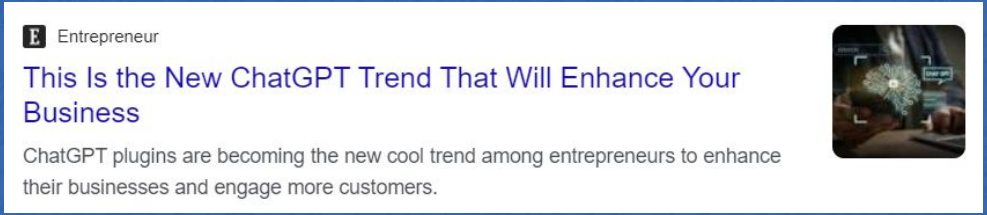
It tries to introduce key concepts and practices about large language models

This presentation tries to solve the misconceptions around large language models by providing a light explanation of how LLMs work and outlining how to use LLMs for editing Wikidata and Wikipedia.



This Year

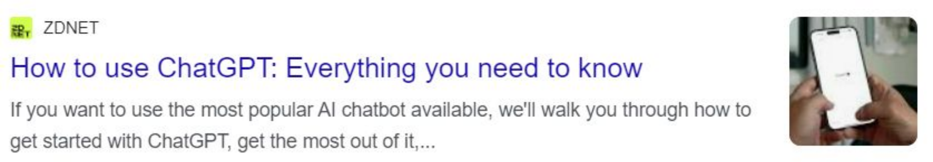
News Headlines



Entrepreneur

This Is the New ChatGPT Trend That Will Enhance Your Business

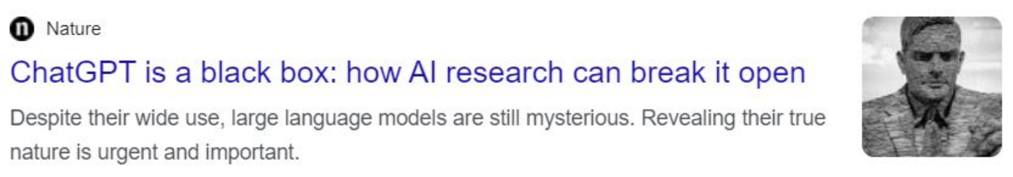
ChatGPT plugins are becoming the new cool trend among entrepreneurs to enhance their businesses and engage more customers.



ZDNET

How to use ChatGPT: Everything you need to know

If you want to use the most popular AI chatbot available, we'll walk you through how to get started with ChatGPT, get the most out of it,...



Nature

ChatGPT is a black box: how AI research can break it open

Despite their wide use, large language models are still mysterious. Revealing their true nature is urgent and important.

This Year

Many LLMs

Bard

ChatGPT

Claude

Assumptions

“We can let LLMs edit Wikipedia articles or Wikidata items”

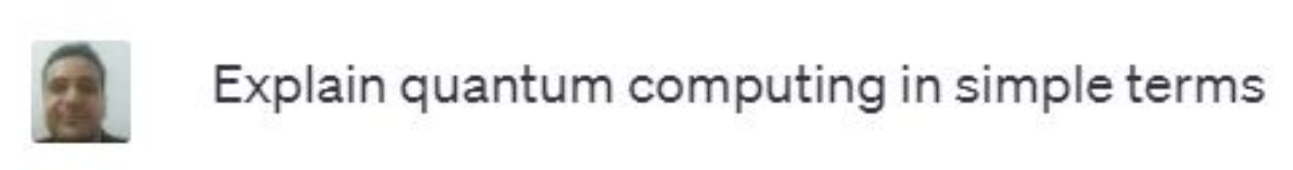
“LLMs can create Wikipedia articles or Wikidata items from scratch”

“LLMs can evaluate Wikipedia or Wikidata”

How an LLM works

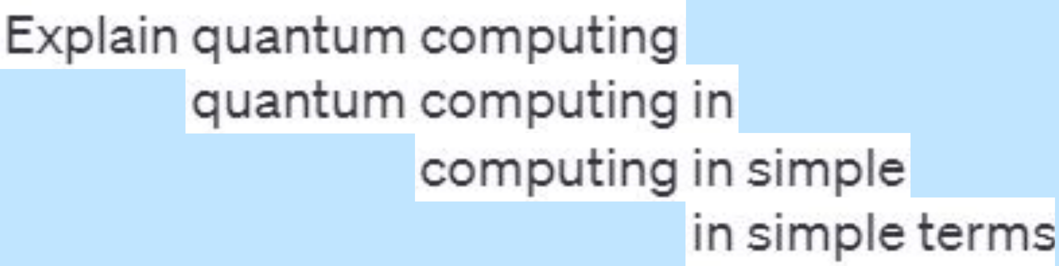
A very simple explanation

This is a **prompt**.



Explain quantum computing in simple terms

This **prompt** will be divided by **tokens**.



Explain quantum computing

quantum computing in

computing in simple

in simple terms

How an LLM works

A very simple explanation

There is a huge amount of **data** in the Internet.

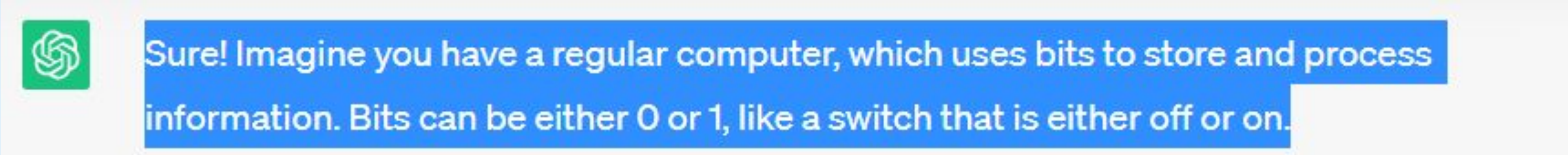
This data is collected and pre-processed using **Embeddings** and **Positional Encoding**.

The pre-processed data is used to train a **transformer** model.

How an LLM works

A very simple explanation

The **transformer** uses the tokens of your prompt to identify the beginning of the answer.

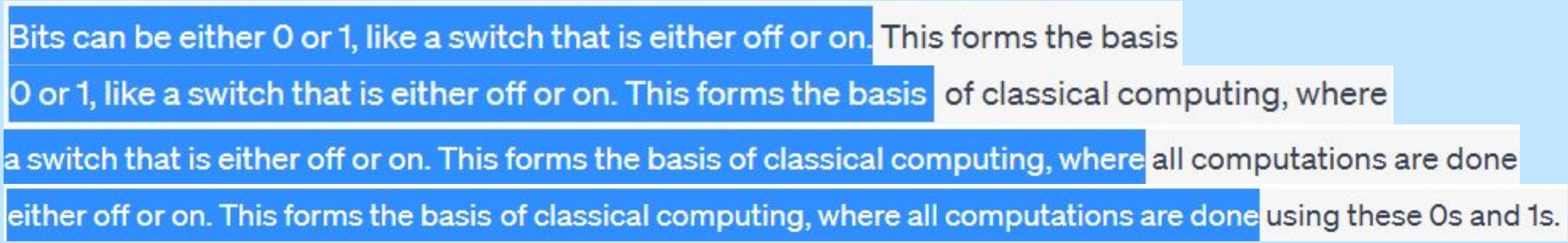


Sure! Imagine you have a regular computer, which uses bits to store and process information. Bits can be either 0 or 1, like a switch that is either off or on.

How an LLM works

A very simple explanation

Then, based on your prompt and a few words from the completed part of the answer, the **transformer** will predict what comes next only based on **probability**. This will be done to incrementally create the



Bits can be either 0 or 1, like a switch that is either off or on. This forms the basis

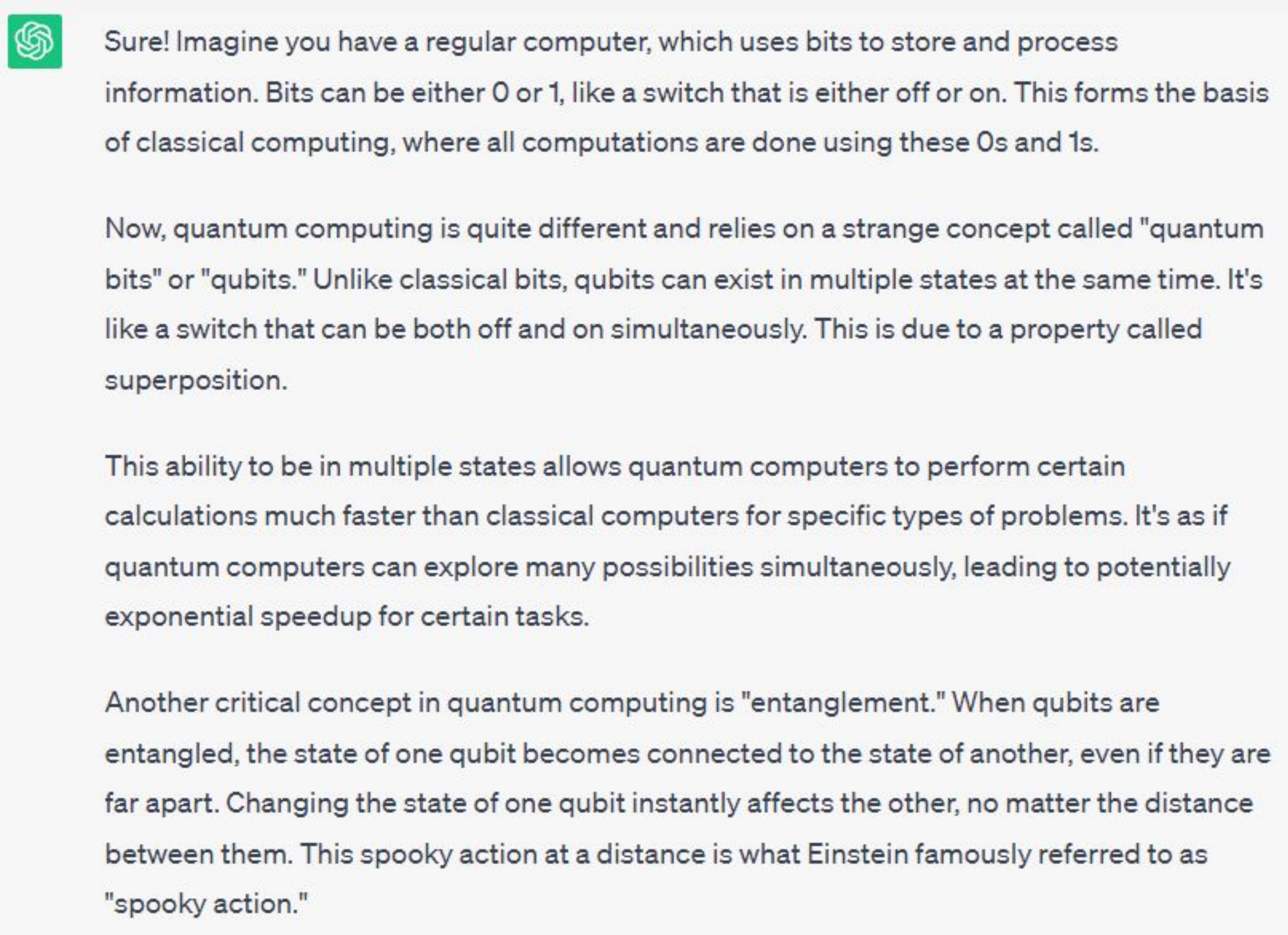
0 or 1, like a switch that is either off or on. This forms the basis of classical computing, where

a switch that is either off or on. This forms the basis of classical computing, where all computations are done

either off or on. This forms the basis of classical computing, where all computations are done using these 0s and 1s.

Final Output

Prompt Answer



Sure! Imagine you have a regular computer, which uses bits to store and process information. Bits can be either 0 or 1, like a switch that is either off or on. This forms the basis of classical computing, where all computations are done using these 0s and 1s.

Now, quantum computing is quite different and relies on a strange concept called "quantum bits" or "qubits." Unlike classical bits, qubits can exist in multiple states at the same time. It's like a switch that can be both off and on simultaneously. This is due to a property called superposition.

This ability to be in multiple states allows quantum computers to perform certain calculations much faster than classical computers for specific types of problems. It's as if quantum computers can explore many possibilities simultaneously, leading to potentially exponential speedup for certain tasks.

Another critical concept in quantum computing is "entanglement." When qubits are entangled, the state of one qubit becomes connected to the state of another, even if they are far apart. Changing the state of one qubit instantly affects the other, no matter the distance between them. This spooky action at a distance is what Einstein famously referred to as "spooky action."

WIKIMANIA
SINGAPORE

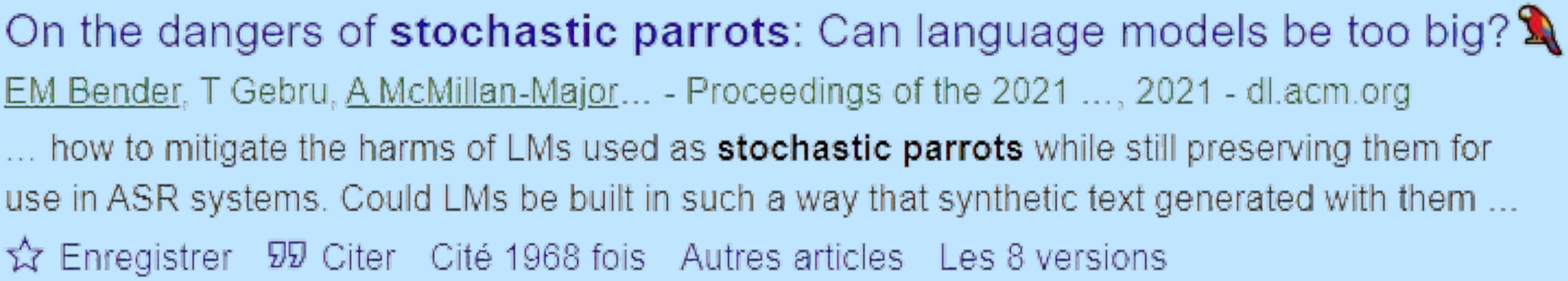
Main features.

Based on probability:

- It will only give you the right answer if it is trained on. No reasoning is there.
- If the information is not used to train the model, it will “hallucinate” by providing the most probable answer.
- But, it can provide the right shape of the answer.

WIKIMANIA
SINGAPORE

Further reading



On the dangers of **stochastic parrots**: Can language models be too big? 🦜

EM Bender, T Gebru, A McMillan-Major… - Proceedings of the 2021 …, 2021 - dl.acm.org

… how to mitigate the harms of LMs used as **stochastic parrots** while still preserving them for use in ASR systems. Could LMs be built in such a way that synthetic text generated with them …

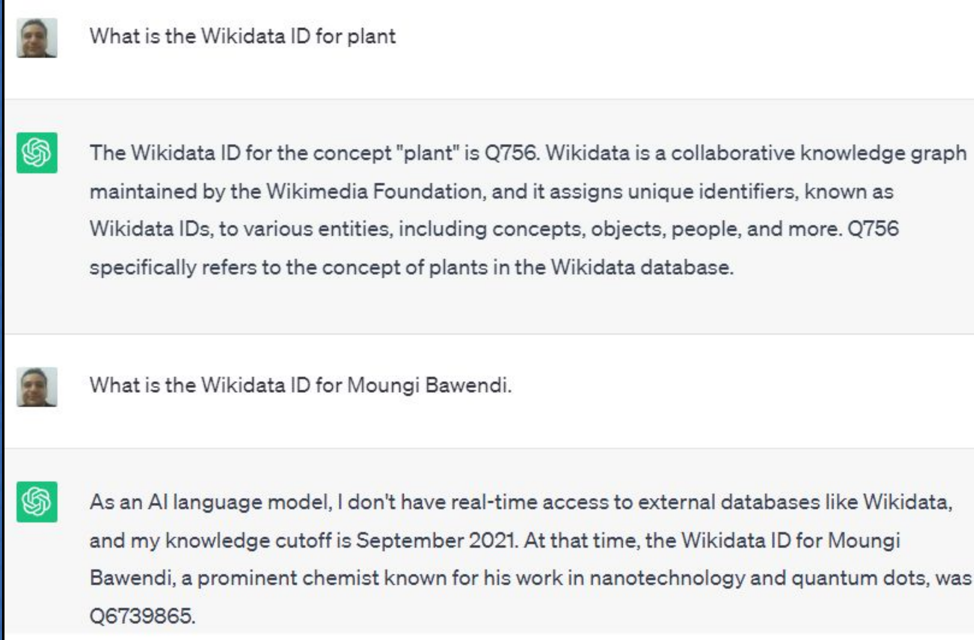
How to edit Wikipedia and Wikidata using LLMs.

Six simple rules for using LLMs

WIKIMANIA
SINGAPORE

Rule 1

- **DO NOT** make LLM-based chatbots make a full Wikipedia page or a QuickStatements batch for creating Wikidata items.
- **DO NOT** rely on Wikidata IDs generated by LLM-based chatbots.

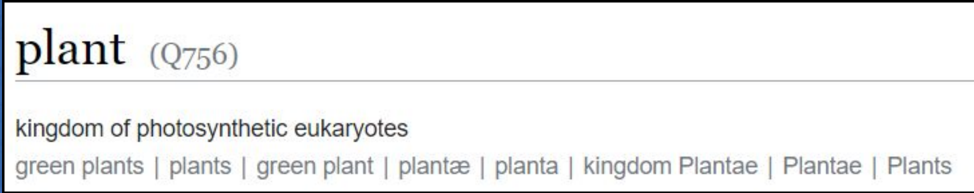


What is the Wikidata ID for plant

The Wikidata ID for the concept "plant" is Q756. Wikidata is a collaborative knowledge graph maintained by the Wikimedia Foundation, and it assigns unique identifiers, known as Wikidata IDs, to various entities, including concepts, objects, people, and more. Q756 specifically refers to the concept of plants in the Wikidata database.

What is the Wikidata ID for Moungi Bawendi.

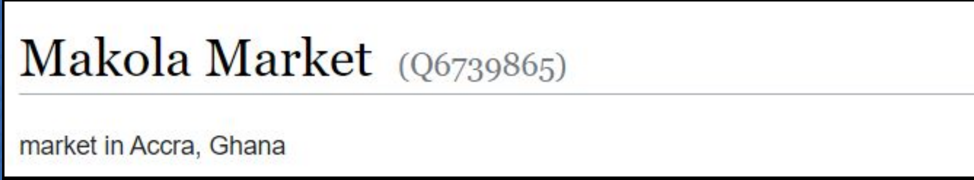
As an AI language model, I don't have real-time access to external databases like Wikidata, and my knowledge cutoff is September 2021. At that time, the Wikidata ID for Moungi Bawendi, a prominent chemist known for his work in nanotechnology and quantum dots, was Q6739865.



plant (Q756)

kingdom of photosynthetic eukaryotes

green plants | plants | green plant | plantæ | planta | kingdom Plantae | Plantae | Plants



Makola Market (Q6739865)

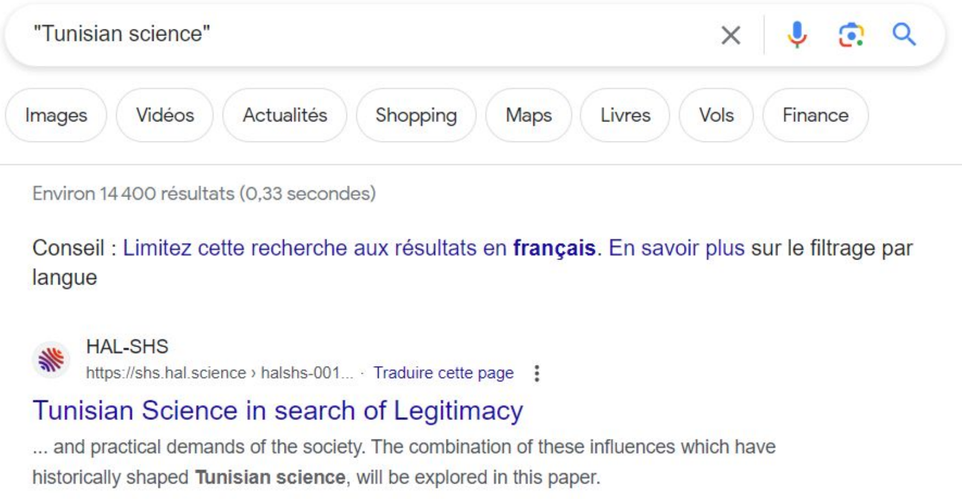
market in Accra, Ghana

WIKIMANIA SINGAPORE

Rule 2

- **DO** your search about the topic of the Wikipedia article or Wikidata item and identify the sources that you will use for your article.
- **DO** your check on whether a source is reliable using the criteria at [[Wikipedia:Wikipedia:Reliable_sources]].

WIKIMANIA
SINGAPORE



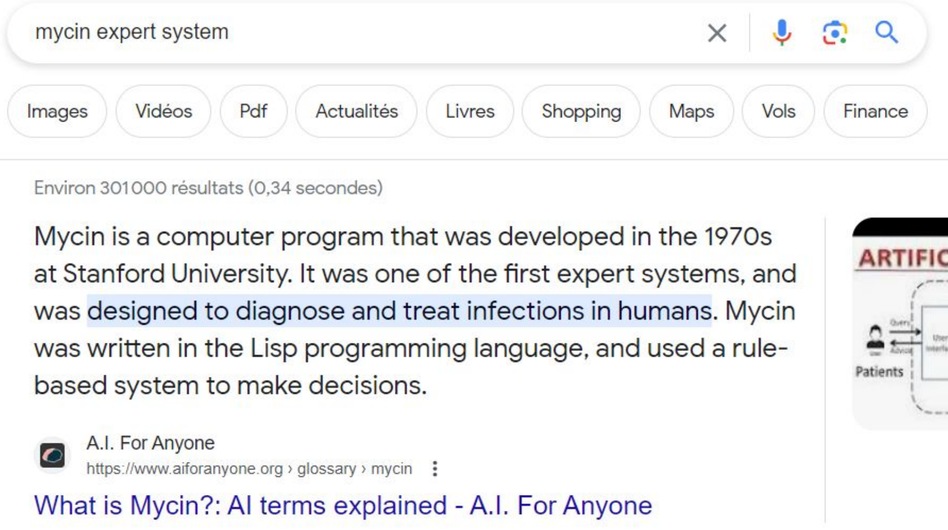
Environ 14 400 résultats (0,33 secondes)

Conseil : Limitez cette recherche aux résultats en **français**. En savoir plus sur le filtrage par langue

HAL-SHS
https://shs.hal.science › halshs-001... · Traduire cette page

Tunisian Science in search of Legitimacy

... and practical demands of the society. The combination of these influences which have historically shaped **Tunisian science**, will be explored in this paper.



Environ 301 000 résultats (0,34 secondes)

Mycin is a computer program that was developed in the 1970s at Stanford University. It was one of the first expert systems, and was designed to diagnose and treat infections in humans. Mycin was written in the Lisp programming language, and used a rule-based system to make decisions.

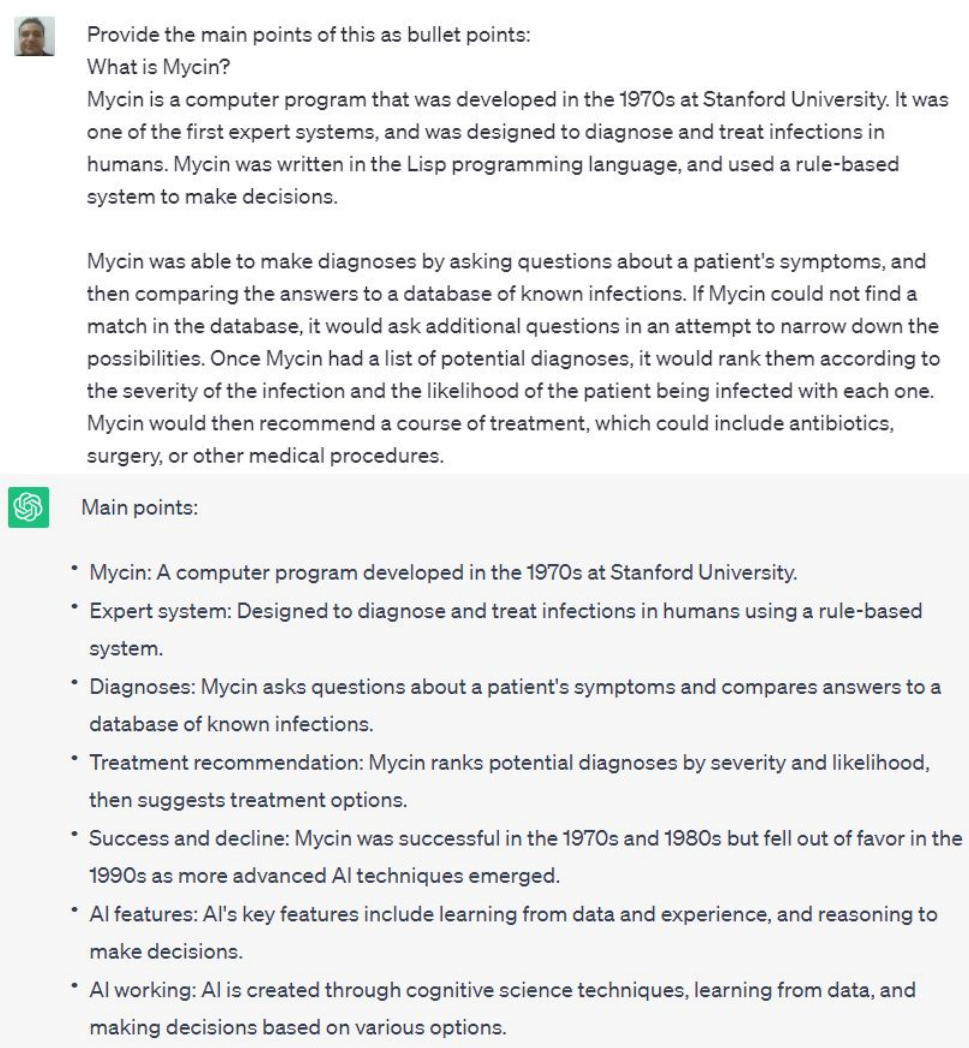
A.I. For Anyone
https://www.aiforanyone.org › glossary › mycin

What is Mycin?: AI terms explained - A.I. For Anyone

Rule 3

- **DO** write the main points from every source to include in Wikidata or Wikipedia as bullet points.
- **DO** use LLM-based chatbots to summarize the sources as bullet points. But, you need to verify the output later to ensure that no bullet point is fabricated or missed.

WIKIMANIA
SINGAPORE



Provide the main points of this as bullet points:
What is Mycin?
Mycin is a computer program that was developed in the 1970s at Stanford University. It was one of the first expert systems, and was designed to diagnose and treat infections in humans. Mycin was written in the Lisp programming language, and used a rule-based system to make decisions.

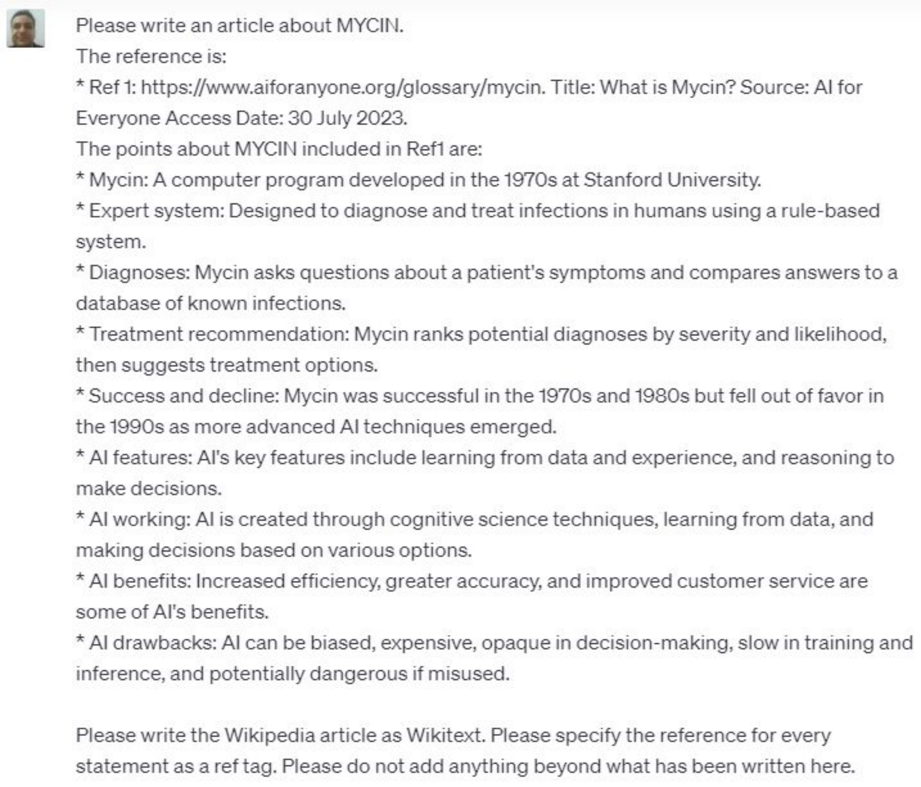
Mycin was able to make diagnoses by asking questions about a patient's symptoms, and then comparing the answers to a database of known infections. If Mycin could not find a match in the database, it would ask additional questions in an attempt to narrow down the possibilities. Once Mycin had a list of potential diagnoses, it would rank them according to the severity of the infection and the likelihood of the patient being infected with each one. Mycin would then recommend a course of treatment, which could include antibiotics, surgery, or other medical procedures.

Main points:

- Mycin: A computer program developed in the 1970s at Stanford University.
- Expert system: Designed to diagnose and treat infections in humans using a rule-based system.
- Diagnoses: Mycin asks questions about a patient's symptoms and compares answers to a database of known infections.
- Treatment recommendation: Mycin ranks potential diagnoses by severity and likelihood, then suggests treatment options.
- Success and decline: Mycin was successful in the 1970s and 1980s but fell out of favor in the 1990s as more advanced AI techniques emerged.
- AI features: AI's key features include learning from data and experience, and reasoning to make decisions.
- AI working: AI is created through cognitive science techniques, learning from data, and making decisions based on various options.

Rule 4

- **DO** write a precise prompt to write an efficient output about the topic. Please include all the references and the bullet points corresponding to each of them.
- **DO** specify that the output should be generated as a wikitext or QuickStatements batch. Please provide the Wikidata IDs for the related entities.
- **DO** require that <ref> tags including the references are added after every new statement.
- **DO** ask the LLM-based chatbot to include categories and infoboxes.



Please write an article about MYCIN.
The reference is:
* Ref 1: https://www.aiforanyone.org/glossary/mycin. Title: What is Mycin? Source: AI for Everyone Access Date: 30 July 2023.
The points about MYCIN included in Ref1 are:
* Mycin: A computer program developed in the 1970s at Stanford University.
* Expert system: Designed to diagnose and treat infections in humans using a rule-based system.
* Diagnoses: Mycin asks questions about a patient's symptoms and compares answers to a database of known infections.
* Treatment recommendation: Mycin ranks potential diagnoses by severity and likelihood, then suggests treatment options.
* Success and decline: Mycin was successful in the 1970s and 1980s but fell out of favor in the 1990s as more advanced AI techniques emerged.
* AI features: AI's key features include learning from data and experience, and reasoning to make decisions.
* AI working: AI is created through cognitive science techniques, learning from data, and making decisions based on various options.
* AI benefits: Increased efficiency, greater accuracy, and improved customer service are some of AI's benefits.
* AI drawbacks: AI can be biased, expensive, opaque in decision-making, slow in training and inference, and potentially dangerous if misused.

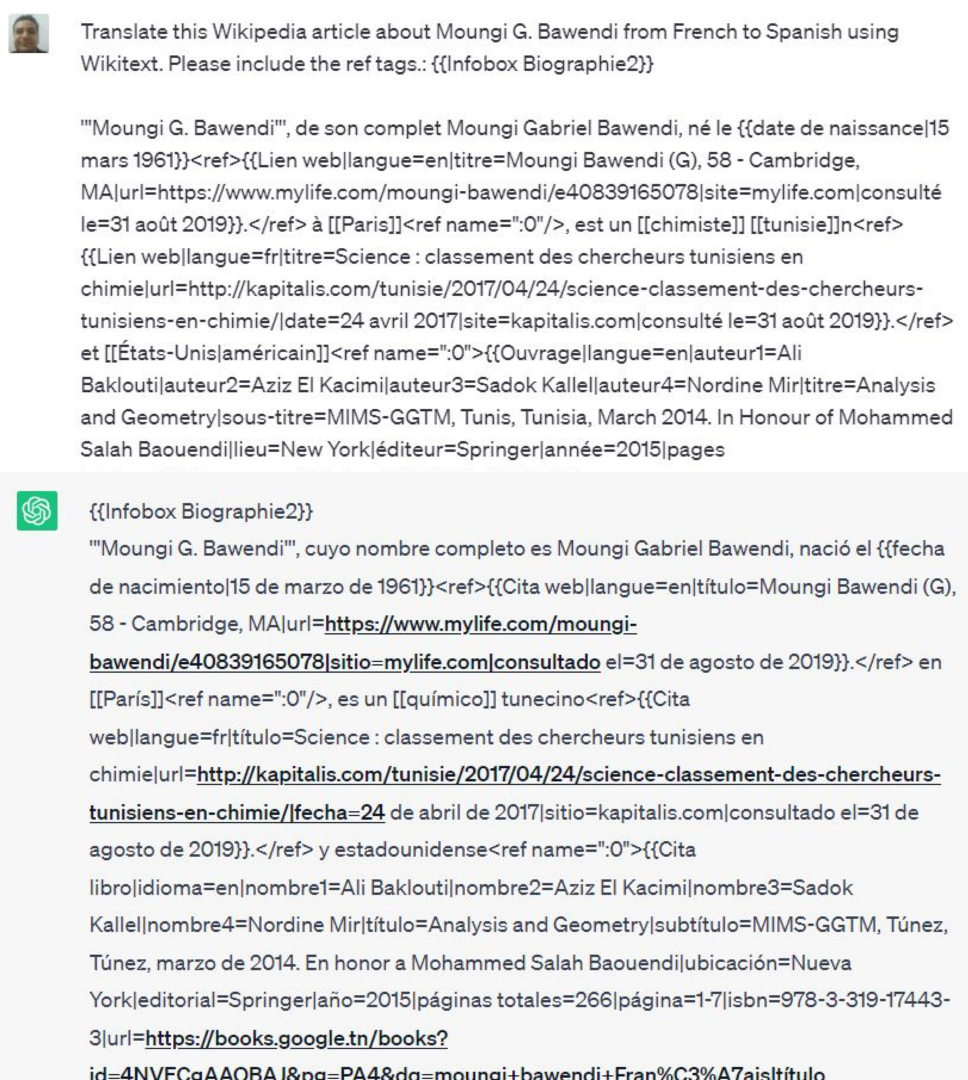
Please write the Wikipedia article as Wikitext. Please specify the reference for every statement as a ref tag. Please do not add anything beyond what has been written here.

WIKIMANIA
SINGAPORE

Rule 5

- **DO** use LLM-based chatbots for language proofreading and translation.
- **DO** check whether the source and target language are efficiently used to train the LLM.

WIKIMANIA
SINGAPORE



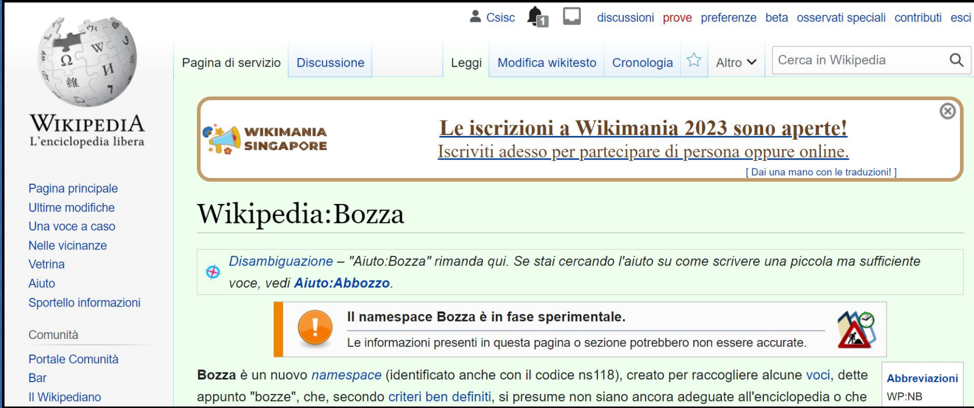
Translate this Wikipedia article about Moungi G. Bawendi from French to Spanish using Wikitext. Please include the ref tags.: {{Infobox Biographie2}}

'''Moungi G. Bawendi''', de son complet Moungi Gabriel Bawendi, né le {{date de naissance|15 mars 1961}}<ref>{{Lien web|langue=en|titre=Moungi Bawendi (G), 58 - Cambridge, MA|url=https://www.mylife.com/moungi-bawendi/e40839165078|site=mylife.com|consulté le=31 août 2019}}.</ref> à [[Paris]]<ref name=":0"/>, est un [[chimiste]] [[tunisie]]n<ref>{{Lien web|langue=fr|titre=Science : classement des chercheurs tunisiens en chimie|url=http://kapitalis.com/tunisie/2017/04/24/science-classement-des-chercheurs-tunisiens-en-chimie/|date=24 avril 2017|site=kapitalis.com|consulté le=31 août 2019}}.</ref> et [[États-Unis|américain]]<ref name=":0">{{Ouvrage|langue=en|auteur1=Ali Baklouti|auteur2=Aziz El Kacimi|auteur3=Sadok Kallel|auteur4=Nordine Mir|titre=Analysis and Geometry|sous-titre=MIMS-GGTM, Tunis, Tunisia, March 2014. In Honour of Mohammed Salah Baouendi|lieu=New York|éditeur=Springer|année=2015|pages

{{Infobox Biographie2}}
'''Moungi G. Bawendi''', cuyo nombre completo es Moungi Gabriel Bawendi, nació el {{fecha de nacimiento|15 de marzo de 1961}}<ref>{{Cita web|langue=en|título=Moungi Bawendi (G), 58 - Cambridge, MA|url=https://www.mylife.com/moungi-bawendi/e40839165078|sitio=mylife.com|consultado el=31 de agosto de 2019}}.</ref> en [[París]]<ref name=":0"/>, es un [[químico]] tunecino<ref>{{Cita web|langue=fr|título=Science : classement des chercheurs tunisiens en chimie|url=http://kapitalis.com/tunisie/2017/04/24/science-classement-des-chercheurs-tunisiens-en-chimie/|fecha=24 de abril de 2017|sitio=kapitalis.com|consultado el=31 de agosto de 2019}}.</ref> y estadounidense<ref name=":0">{{Cita libro|idioma=en|nombre1=Ali Baklouti|nombre2=Aziz El Kacimi|nombre3=Sadok Kallel|nombre4=Nordine Mir|título=Analysis and Geometry|subtítulo=MIMS-GGTM, Túnez, Túnez, marzo de 2014. En honor a Mohammed Salah Baouendi|ubicación=Nueva York|editorial=Springer|año=2015|páginas totales=266|página=1-7|isbn=978-3-319-17443-3|url=https://books.google.tn/books?id=4NVFCgAAQBAJ&pg=PA4&dq=moungi+bawendi+Fran%C3%A7ais|título

Rule 6

- **DO NOT** directly put the output of the LLM in a Wikipedia Page. Please use Sandbox and edit the work before sharing it.
- Same for Wikidata. Please check your QuickStatements batch before uploading it to Wikidata.



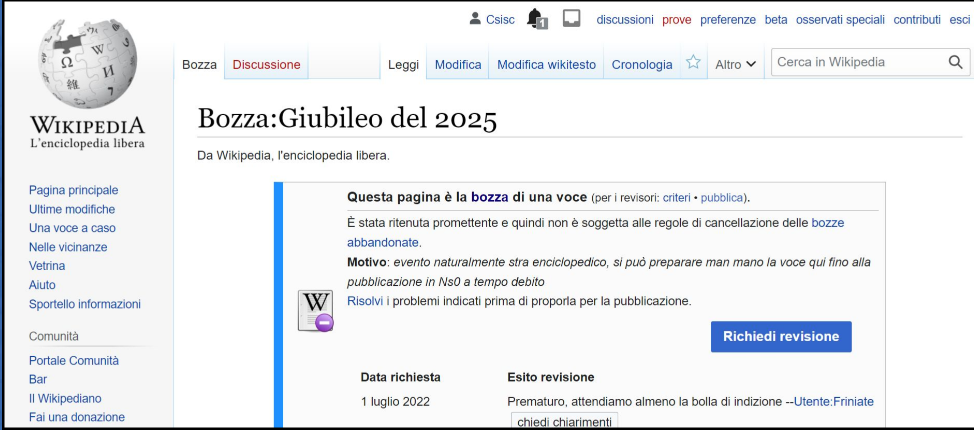
Le iscrizioni a Wikimania 2023 sono aperte!
Iscriviti adesso per partecipare di persona oppure online.
[Dai una mano con le traduzioni!]

Wikipedia:Bozza

*Disambiguazione – "Aiuto:Bozza" rimanda qui. Se stai cercando l'aiuto su come scrivere una piccola ma sufficiente voce, vedi **Aiuto:Abbozzo**.*

Il namespace Bozza è in fase sperimentale.
Le informazioni presenti in questa pagina o sezione potrebbero non essere accurate.

Bozza è un nuovo *namespace* (identificato anche con il codice ns118), creato per raccogliere alcune voci, dette appunto "bozze", che, secondo criteri ben definiti, si presume non siano ancora adeguate all'enciclopedia o che



Bozza:Giubileo del 2025

Da Wikipedia, l'enciclopedia libera.

Questa pagina è la bozza di una voce (per i revisori: criteri • pubblica).

È stata ritenuta promettente e quindi non è soggetta alle regole di cancellazione delle bozze abbandonate.

Motivo: *evento naturalmente stra enciclopedico, si può preparare man mano la voce qui fino alla pubblicazione in Ns0 a tempo debito*

Risolvi i problemi indicati prima di proporla per la pubblicazione.

Data richiesta	Esito revisione
1 luglio 2022	Prematuro, attendiamo almeno la bolla di indizione --Utente:Friniate

WIKIMANIA
SINGAPORE

Any question

Houcemeddine Turki
User:Csisc

E-mail: turkiabdelwaheb@hotmail.fr
Phone: +21629499418
Twitter: @Csisc1994
LinkedIn: Houcemeddine Turki

WIKIMANIA
SINGAPORE

WIKIMANIA
SINGAPORE