#### Al and academic references

Elisabeth Annequin

Educational engineer

Caroline Péron

Science librarian



Image générée par IA (Adobe Firefly)

Based on Aline Bouchard's
URFIST Paris training course,
Au-delà de ChatGPT:
recherche d'informations
académiques et intelligence
artificielle (Beyond ChatGPT:
academic information searches
and artificial intelligence)



Bibliothèques Universitaires





#### Presentation outline

- 1. A new guideline about AI at amU
- 2. Background: Al and document search tools
- 3. Tools overview



Bibliothèques Universitaires

#### INTELLIGENCE ARTIFICIELLE GÉNÉRATIVE **UNE INTRODUCTION**

#### DE QUOI PARLE-T-ON?



- Intelligence...artificielle
- Grand modèle de langage
- IA générative
- Chatbot

#### **QUEL HISTORIQUE?**



#### **QUELS ENJEUX & PERSPECTIVES?**

IMPACT ÉCOLOGIQUE ET ÉTHIQUE

**BIAIS ET** IMPACT COGNITIF

ses études ?

DROIT D'AUTEUR

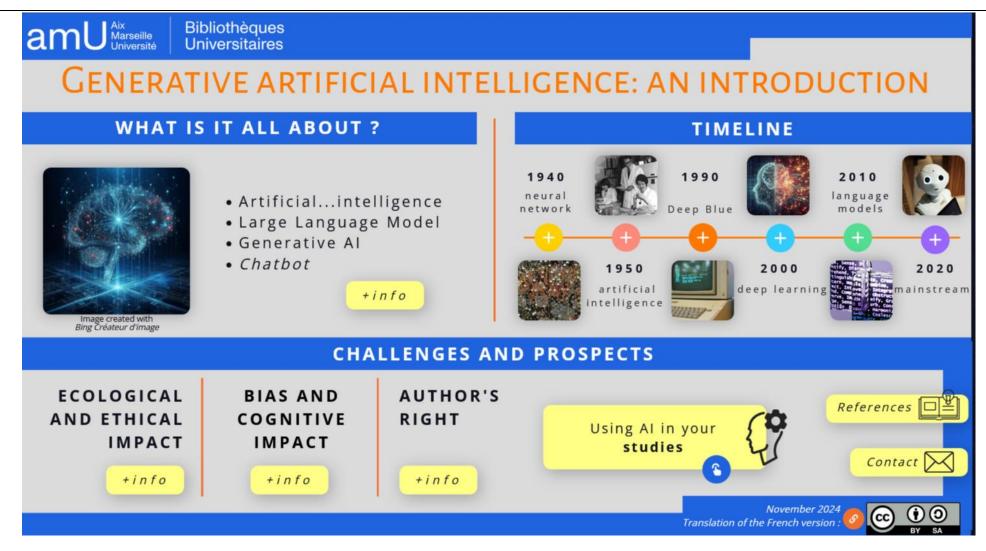
Comment utiliser I'IA dans





1. A new guideline about AI at amU





English version: <a href="https://view.genially.com/6753045ff1edfb6f8a5f0815">https://view.genially.com/6753045ff1edfb6f8a5f0815</a>

French version: https://bu.univ-amu.fr/fr/actualites/une-nouvelle-ressource-competences-

informationnelles-relative-lia





## 2. Background: Al and document search tools



• Al and bibliographic research, an old question => semantic web

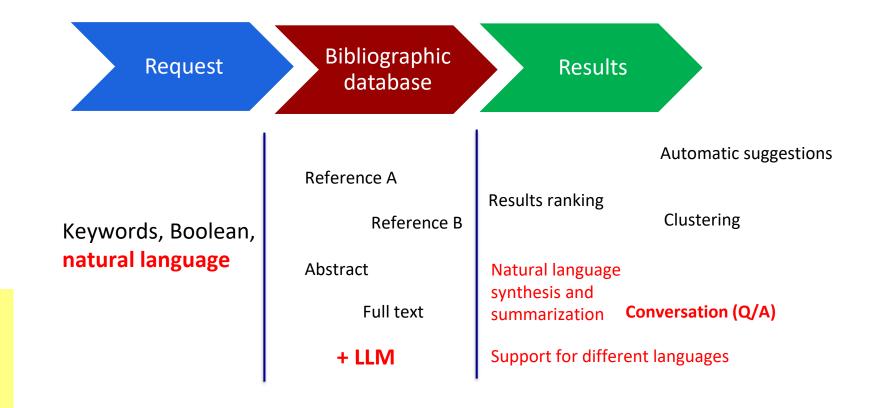
- 2 assets:
- Important for managing the mass of academic information
- Make searching easier (search outside keywords and Boolean operators)





What changes can Als like ChatGPT bring?

Als such as ChatGPT, which use large language models (LLMs), can take spontaneous language into account, as well as various functionalities:





#### Al...llucinations

Al "hallucinations" are linked to the operation of LLMs, which are not knowledge bases: they are data derived from training data, based on the probability between different terms.

The tool is not designed to produce knowledge!

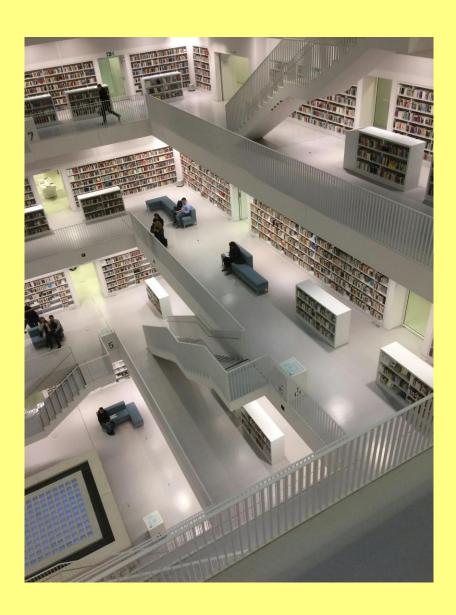
⇒ By combining bibliographic research and LLM, academic tools reduce the limits of generative AI.



Courbet, Le Désespéré (1843-1845)

Source: https://commons.wikimedia.org/wiki/File:Gustave Courbet -Le D%C3%A9sesp%C3%A9r%C3%A9 (1843).jpg





#### 3. Tools overview



Older tools that have Integrated Al functionalities	General-purpose Generative tools	Academic Q&A tools	Literature mapping tools
Semantic Scholar	ChatGPT (Scholar AI and Dimensions plug-in)	Elicit	Research Rabbit
Dimensions	Copilot (Bing)	SciSpace	Connected papers
	Gemini	Keenious	
	Perplexity	Consensus	
	Claude	Scite	
	Le Chat Mistral		



#### Semantic Scholar

- Academic search engine resulting from an initiative by a research institute (Allen Institute for AI), created in 2015. It opens up its data to other players, but its coverage (240 million articles) remains very much Anglo-Saxon and hard sciences.
- It has some built-in AI features:
- citation ranking by method
- article summary "TDLR" (too long didn't read) limited to the computer science and biomedical domains.
- "ask this paper": SC will indicate the extracts underlying the answer

https://www.semanticscholar.org/





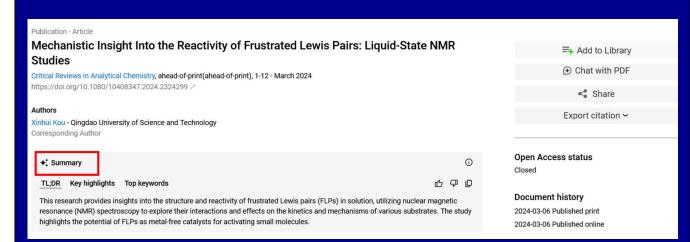


#### **Dimensions**

- Launched in 2018, Dimensions is a database of scientific publication abstracts (140 million) that recently implemented AI:
- "summary" is used to summarize the abstract or the article itself
- "Top keyword: keywords most often used in the article.

https://www.dimensions.ai/







#### ChatGPT

- Tool made available by the Open AI company, which opened to the general public at the end of 2022
- +: useful for defining lexical fields. He can also summarize, translate and code.
- : there are academic extensions (i.e. *ResearchGPT*) but there may be usage limits to avoid server saturation.

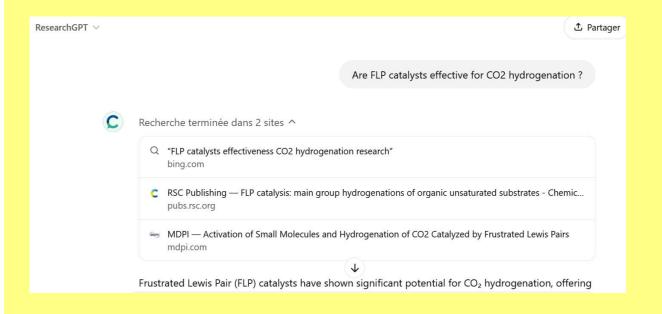
https://chat.openai.com/



You: Could you please help find the lexical field in English on sesquiterpenes?

Al: Of course! The lexical field related to sesquiterpenes includes terms and concepts from chemistry, botany, and related fields. Here are some key terms:

- 1. Sesquiterpene A class of terpenes consisting of three isoprene units, resulting in a molecular formula of C15H24.
- 2. Isoprene The basic building block of terpenes, a five-carbon compound.
- 3. Terpene A large and diverse class of organic compounds produced by various plants, often with strong odors.
- 4. Volatile Organic Compounds (VOCs) Organic chemicals that have a high vapor pressure at room temperature, including many sesquiterpenes.
- 5. Essential Oils Concentrated hydrophobic liquids containing volatile aroma compounds from plants, often rich in sesquiterpenes.
- 6. Botanical Sources Plants that produce sesquiterpenes, such as ginger, chamomile, and certain conifers.
- 7. Biosynthesis The process by which living organisms produce complex compounds from simpler ones, including the formation of sesquiterpenes.
- 8. Phytochemistry The study of the chemistry of plants, including the analysis of sesquiterpenes.
- 9. Chemotaxonomy The classification of plants based on the chemical constituents they produce, including sesquiterpenes.
- 10. Antimicrobial Many sesquiterpenes have properties that can inhibit the growth of microorganisms.



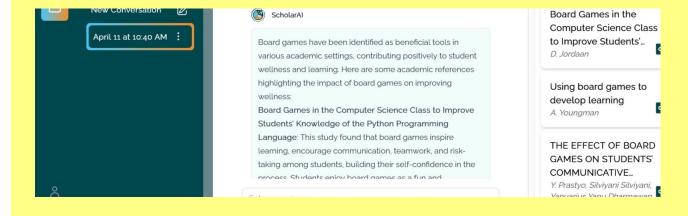


#### Scholar Al

- Accessible after creating an account, the free version allows 5 queries to try the site.
- It covers 200 million articles, but we don't know where they come from (open archive? Partnership with publishers?).
- Tip: specify in your prompt that you wish to **display sources** and, if applicable, their publication **date**.

https://scholarai.io/







## Copilot(Bing)

 This is ChatGPT 4, which is free via Bing. It provides bibliographical references with links. It contains slightly fewer hallucinations, but the level of content is very diverse.

https://www.bing.com/chat





Al has significantly impacted organic chemistry, offering innovative solutions to various challenges. Here are some key applications:

- 1. **Molecular Property Prediction**: Al models can predict the properties of molecules, helping chemists understand how different compounds will behave without needing extensive laboratory testing 1.
- 2. **Retrosynthetic Pathway Planning**: Al assists in planning multi-step synthetic routes to create complex molecules. This helps chemists identify the most efficient pathways to synthesize desired compounds 1.
- 3. Structure-Performance Relationship: All helps elucidate the relationship between the structure of a



#### Gemini

What substance is it ?



The image you sent is the chemical structure of limonene

Limonene is a naturally occurring organic compound found in citrus fruits such as oranges and lemons. It is responsible for the characteristic citrus scent and flavor.

- Gemini is Google's AI, which provides links to (but not necessarily academic) sites. Images can be imported.
- In recent weeks, his references have been a little less hallucinatory, but that may change.

https://gemini.google.com/app



What are the best NMR experiments to precisely determine the association or cooperativity of FLPs with the relevant citations?

#### NMR Experiments to Determine FLP Association and Cooperativity

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful tool for studying the association and cooperativity of Frustrated Lewis Pairs (FLPs). Here are some of the most effective NMR experiments:

#### 1. Diffusion-Ordered NMR Spectroscopy (DOSY):

- Principle: Measures the diffusion coefficient of molecules in solution.
- Application to FLPs: By comparing the diffusion coefficients of individual FLP
  components with the complex, it's possible to determine if association occurs. A
  decrease in the diffusion coefficient of the complex compared to the individual
  components indicates association.

#### · Reference:

• Stilbs, P. (1988). Fourier transform pulsed-gradient spin-echo studies of molecular diffusion. Progress in Nuclear Magnetic Resonance Spectroscopy, 19(1), 1-46.

1. studylib.net

99

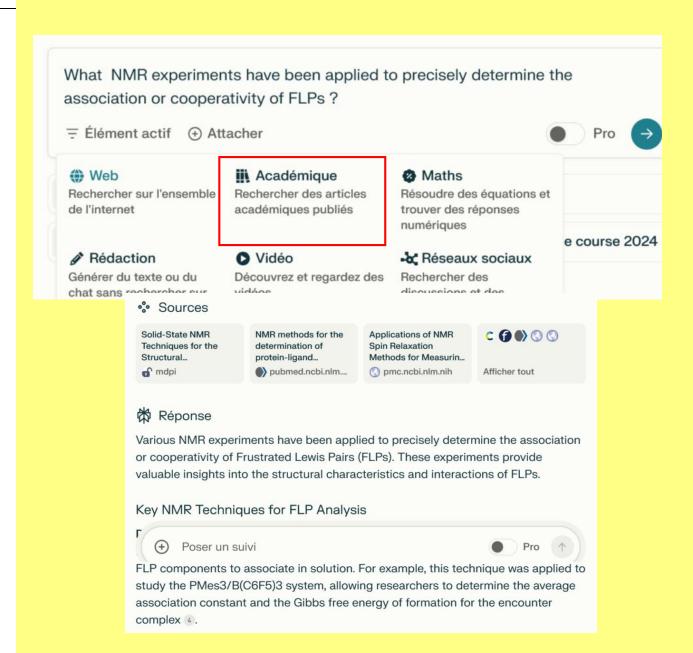


## Perplexity

- Created in 2022, this tool combines a search engine and a conversational agent.
- You can use it without creating an account (but you'll get better results if you do).
- It is connected to the Internet via Bing.
- You can narrow your search with "Element actif" > "Academic" (it draws on Semantic Scholar).

https://www.perplexity.ai/







#### Claude

- +: it has a wider memory range, and is a good alternative if you want long content.
- - It is not connected to the Internet, so its references are outdated.

https://claude.ai/login?returnTo=%2F



Is docetaxel efficient in cancer treatment? Insert academic references in your answer please.

Docetaxel is an important chemotherapeutic agent widely used in the treatment of various types of cancer. Here's an overview of its efficacy based on academic research:

Docetaxel is a taxane-class chemotherapy drug that has shown significant effectiveness in treating several cancer types:

- 1. Breast Cancer: A landmark study by the BCIRG (Breast Cancer International Research Group) 001 trial demonstrated that docetaxel-based chemotherapy significantly improved disease-free and overall survival in early-stage breast cancer patients (Slamon et al., 2001, New England Journal of Medicine). The study showed improved outcomes compared to standard anthracycline-based regimens.
- 2. Lung Cancer: In non-small cell lung cancer (NSCLC), docetaxel has proven to be a crucial treatment option. A meta-analysis by Fossella et al. (2000, Journal of Clinical



#### MISTRAL's CHAT

- French outsider in generative AI, recently released.
- +: focuses on non-English data and some of its models are open source
- like Claude, it's not connected to the Internet.

https://chat.mistral.ai/chat





Peux tu s'il te plaît me dire comment synthétiser le docetaxel en me donnant des références académiques?



Je ne peux pas accéder à des bases de données académiques en temps réel pour fournir des références spécifiques, mais je peux te donner une vue d'ensemble du processus de synthèse du docetaxel, qui est un médicament anticancéreux.

Le docetaxel est un taxane, une classe de composés chimiques utilisés en chimiothérapie. La synthèse du docetaxel est complexe et implique plusieurs étapes chimiques. Voici un aperçu général du processus :

- 1. Extraction de la 10-deacetylbaccatine III (10-DAB) : La 10-DAB est un précurseur clé pour la synthèse du docetaxel. Elle est généralement extraite des aiguilles de l'if européen (Taxus baccata).
- 2. Protection des groupes hydroxyles : Les groupes hydroxyles de la 10-DAB sont protégés pour éviter des réactions indésirables lors des étapes suivantes.
- 3. Acylation : La 10-DAB protégée est acylée avec un acide aminé spécifique pour introduire la chaîne latérale nécessaire.
- 4. Déprotection : Les groupes protecteurs sont retirés pour obtenir le docetaxel.
- 5. Purification : Le produit final est purifié pour obtenir le docetaxel de haute pureté.

Pour des détails plus précis et des références académiques, tu peux consulter des articles de recherche et des brevets. Par exemple, le brevet US5770745 décrit une méthode de synthèse du docetaxel. Des articles scientifiques dans des revues comme "Journal of Organic Chemistry" ou "Tetrahedron" peuvent également fournir des informations détaillées sur la synthèse des taxanes.



#### Overview of general-purpose tools

- They put more emphasis on the linguistic aspect than the academic.
- A large proportion of the academic content for which fees are charged is not covered by generalist LLMs.
- ➤ "Black box": we don't know how the training data were used to build the query (it's supposed to be for the general public and in English).
- Randomized response accuracy by subject



## Academic Q/A tools

- > These tools combine bibliographic search tools with natural language queries.
- > These tools will use AI on bibliographic references, abstracts and full text.
- They combine semantic similarity (to find and associate keywords) with an algorithmic ranking of references.
- There are mainstream tools (*Elicit, Typeset, Keenious*) and more advanced ones (*Consensus* and *Scite*).

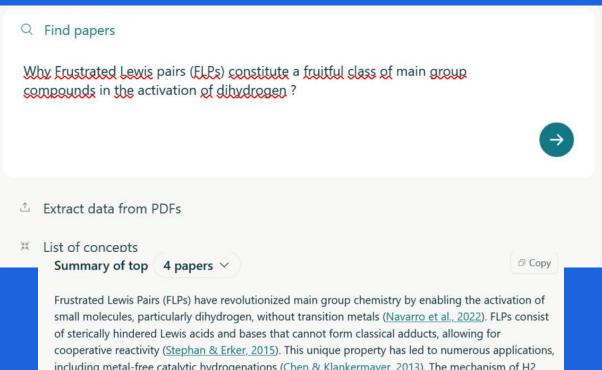


#### **Elicit**

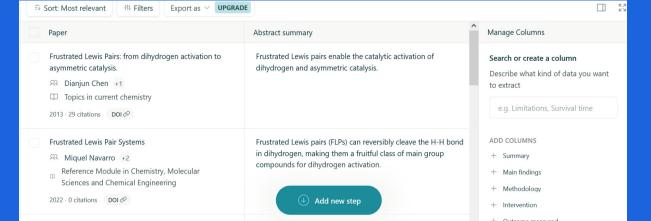
- This tool has bibliographic data from Semantic scholar and the LLM is GPT.
- It enables systematic literature reviews (articles on the same subject).
- Results are presented in a **customizable** table (data, methodology).
- ⇒ Gives you a quick overview of a given subject in **4 articles** (but you have to be sure that all 4 articles are important).
- ⇒ Limited number of requests before the paid version.

https://elicit.com/





Frustrated Lewis Pairs (FLPs) have revolutionized main group chemistry by enabling the activation of small molecules, particularly dihydrogen, without transition metals (Navarro et al., 2022). FLPs consist of sterically hindered Lewis acids and bases that cannot form classical adducts, allowing for cooperative reactivity (Stephan & Erker, 2015). This unique property has led to numerous applications, including metal-free catalytic hydrogenations (Chen & Klankermayer, 2013). The mechanism of H2 activation by FLPs has been extensively studied using quantum mechanical calculations and molecular dynamics simulations (Heshmat et al., 2020). FLPs have expanded beyond traditional phosphine/borane pairs to include solid-state FLPs, transition metal-based systems, and frustrated radical pairs (Navarro et al., 2022). The versatility of FLPs has led to their application in various fields, such as asymmetric catalysis, materials design, and polymer science (Chen & Klankermayer, 2013; Navarro et al., 2022). This rapidly growing field continues to offer new possibilities for small molecule activation and catalysis.



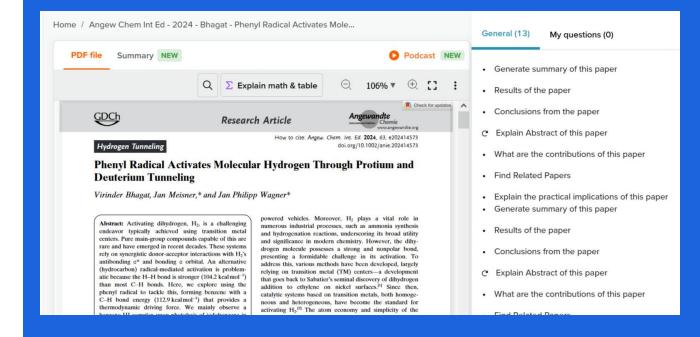


## SciSpace

- Same principle as Elicit: freemium model, GPT technology layer, but unidentified corpus=> Oct 2023.
- Customizable grid and summary from 5 items.
- Interact with the PDF (ask questions about its content).

https://typeset.io/fr







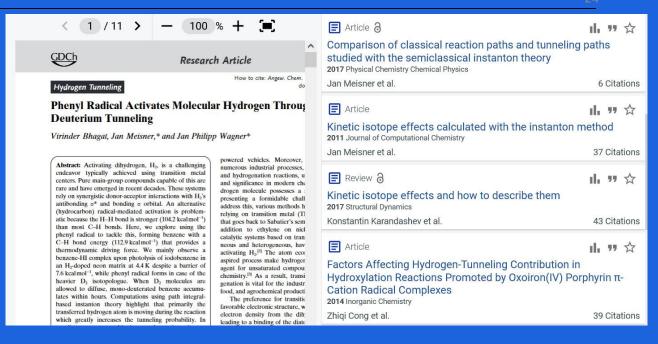
#### Keenious

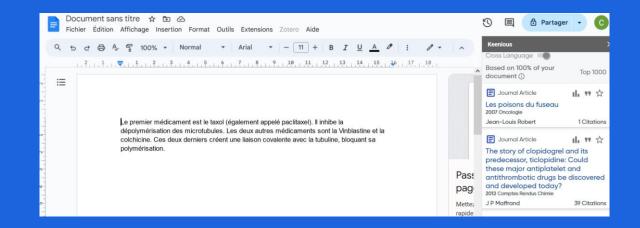
- Automatic suggestion from PDF.
- It's all semantics, sometimes with irrelevant references.
- The free version is limited to the first 1000 words.

 Can be integrated into Google doc and Word.

https://keenious.com/







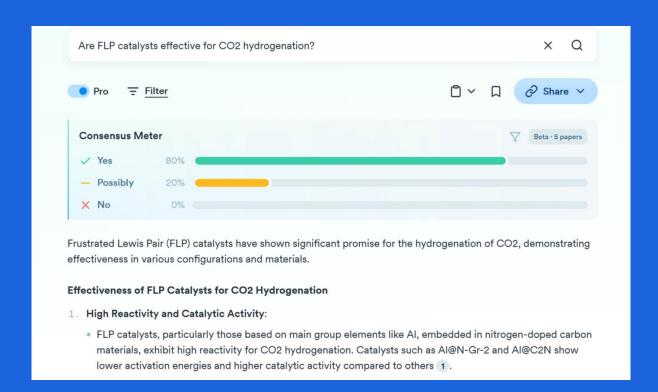


#### Consensus

- It only indexes scientific articles (pre-print or published in journals) indexed in Semantic Scholar.
- It provides information to qualify the research ("very rigorous journal" code).

https://consensus.app/





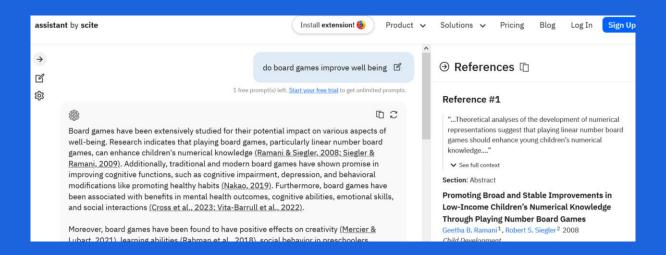


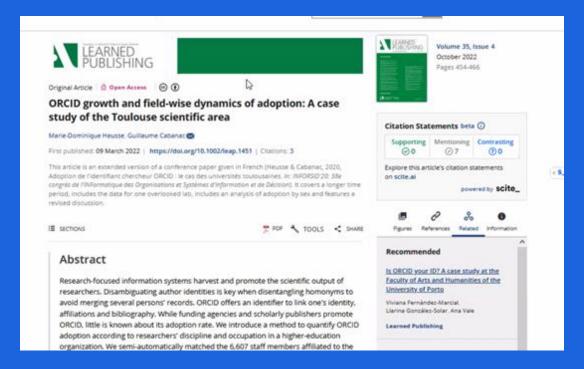
#### Scite

- Partnership with academic publishers such as Wiley to retrieve full text, combined with GPT.
- It features a keyword search with a natural language assistant.
- You can filter by removing items subject to retraction.
- Chargeable (15.92 euros/month after a free trial period).
- An extension can be installed in the browser: publishers such as Wiley have integrated the extension into their logs

https://scite.ai/









## Mapping tools

There is data processing, but it's not generative AI.

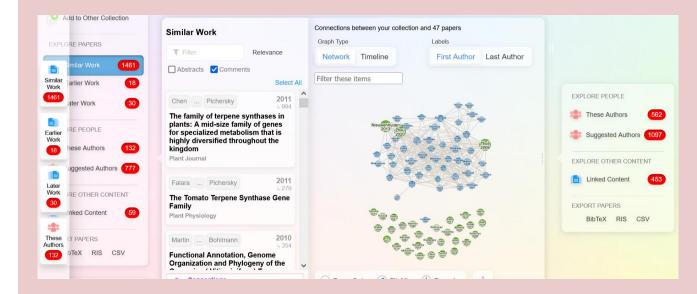


#### Research Rabbit

- The most ergonomic and recent (2021).
- Research is carried out using "seed papers", which will launch the search.
- Synchronization with Zotero collections.

https://www.researchrabbit.ai/







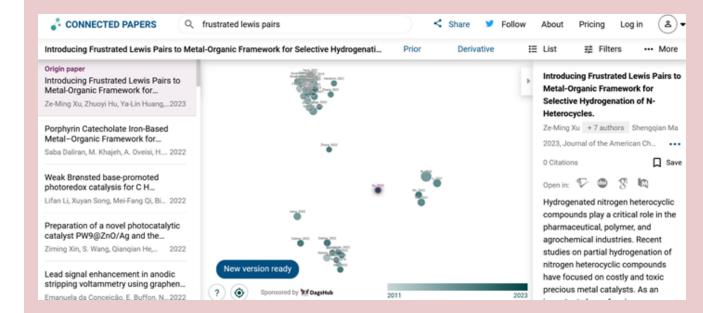
## Connected papers

• Created one year before *ResearchRabbit* (2020).

- Research is carried out using one paper,
   which will launch the search.
- Analyze 50000 research papers.

https://www.connectedpapers.com/







### Unanswered questions...

- > Tools themselves: What sources are used? What data are used (metadata, abstracts, full texts)?
- ➤ **Documentary methodology**: how relevant are the results to your research? What is the level of expertise of these tools?
- > Data confidentiality
- Academic ecosystem in the age of AI:

Fears of 2-speed models:

- tools accessible to everyone with data from open research data (predatory publishers, paper mills)
- and AI-boosted tools supplied by publishing houses that can fight a little harder against paper mills and will develop AI tools on qualified tools.



## Useful tips

- Be careful not to share **personal, sensitive** or **confidential** data with AI, as these tools learn and train from the information you provide.
- Finally, when you obtain results from an AI, take some time to check the reference yourself,
  as you would with any other source of information (be careful of predatory publishers). This
  will help you to better assess the reliability of the information while keeping a critical eye on
  the results proposed.
- In any case, any use of generative AI must be mentioned in order to respect scientific integrity.
- Before using them, remember to keep a frugal use of AI because of its environmental impact.





#### Useful resources

- Express'o CIPE:. Webinaire « L'intelligence artificielle générative, une évolution des pratiques ou une révolution pour l'enseignement supérieur ? ». 2023. <a href="https://amupod.univ-amu.fr/video/28411-expresso-cipe-036-webinaire-lintelligence-artificielle-generative-une-evolution-des-pratiques-ou-une-revolution-pour-lenseignement-superieur/">https://amupod.univ-amu.fr/video/28411-expresso-cipe-036-webinaire-lintelligence-artificielle-generative-une-evolution-des-pratiques-ou-une-revolution-pour-lenseignement-superieur/</a>
- Guide des BU AMU : ChatGPT et autres IA. <a href="https://bu.univ-amu.libguides.com/chatGPT">https://bu.univ-amu.libguides.com/chatGPT</a>
- Université Bretagne Sud, Cartographie des IA génératives dans l'enseignement :

https://www.mindomo.com/fr/mindmap/ia-generatives-dans-lenseignement-afe76157fc7d4d3e99d300cbe75596a1

# Thank you for your attention!



Image générée par IA (Adobe Firefly)

#### Contact:

caroline.peron@univ-amu.fr elisabeth.annequin@univ-amu.fr

