12.20.23

THE BRIGHT FUTURE OF AI



An optimist's guide to AI + Publishing

hum

Introduction

> 2023-level AI can already generate suggestive hints and promising leads to a working mathematician and participate actively in the decision-making process. When integrated with tools such as formal proof verifiers, internet search, and symbolic math packages, I expect, say, 2026-level AI, when used properly, will be a trustworthy co-author in mathematical research, and in many other fields as well."
>
> - Terrence Tao

Data and AI is the next big shift in scholarly publishing, yet many publishers remain in the earliest stages of adoption. They're in the "wait and see" phase, dabbling tentatively. Most underestimate the velocity of change.

Even if all progress stopped now, we'd spend 10+ years integrating existing technology.

But if anything, progress in AI is accelerating.

This whitepaper serves as a wake-up call, a guide, and a toolkit:

- you must accelerate your AI efforts or risk being left behind
- you must understand the basics of AI
- you must employ AI tools before you are consumed by them

🚀

- hum

Table of Contents

Understanding AI 4

Big Ideas & New Frontiers 6

Putting AI to Work in Publishing 14

R&D, Future Trends to Watch 28

Preparing for the AI Future Now 34

Understanding AI

What do we mean when we say AI (Artificial Intelligence)?

Intelligence is the ability to learn, understand, and apply knowledge, solve problems, and adapt to new situations.

For a reference point, we can look to intelligence found in the natural world:

Natural Intelligence: Adaptive, found in nature

Involving cognitive processes:

- 👁️ Perception
- 🧠 Learning
- 💭 Memory
- 🧮 Reasoning
- ❓ Problem Solving
- 📢 Communication
- 🐝 Social Behavior









AI is focused on creating machine intelligence.

Practically, artificial intelligence is a mix of data, models, engineered systems that show signs of intelligence, but more importantly, put that intelligence to work.

Artificial Intelligence:

1. **Classic Machine Learning:** Task-specific algorithms with manually crafted features.

2. **Narrow AI:** *Deep learning-based systems** designed for specific tasks. Some feature engineering might still be involved, and while they can handle more intricate tasks than classic ML, they are confined to their training domain. Could also be called old or classic deep learning.

3. **Broad AI:** Systems like ones anchored by GPT-4 (pre-trained foundation models + memory, tools, and internet access). While not truly general, it can operate across a broader set of domains, with expanded capabilities due to external tools and information (via internet access).

4. **Artificial General Intelligence (AGI):** Machines with general problem-solving capabilities across all domains, equivalent to the median human worker in that domain.

5. **Artificial Superintelligence (ASI):** Systems that surpass human intelligence in every conceivable domain.

* **Deep Learning**, inspired by the human brain and its neurons, enables computers to learn from vast amounts of data. Deep learning models are strong at recognizing patterns and making predictions.

Current systems are a mix of 1, 2, and 3.

We've seen incredible progress by combining these new deep learning model architectures with massive amounts of computing power and big datasets.

The key innovation is models built using "attention", enabling the models to pay attention to the most important features of document, photo, song or video, even if they are spread far apart. This allows them to learn from "reading," "seeing," or "listening"

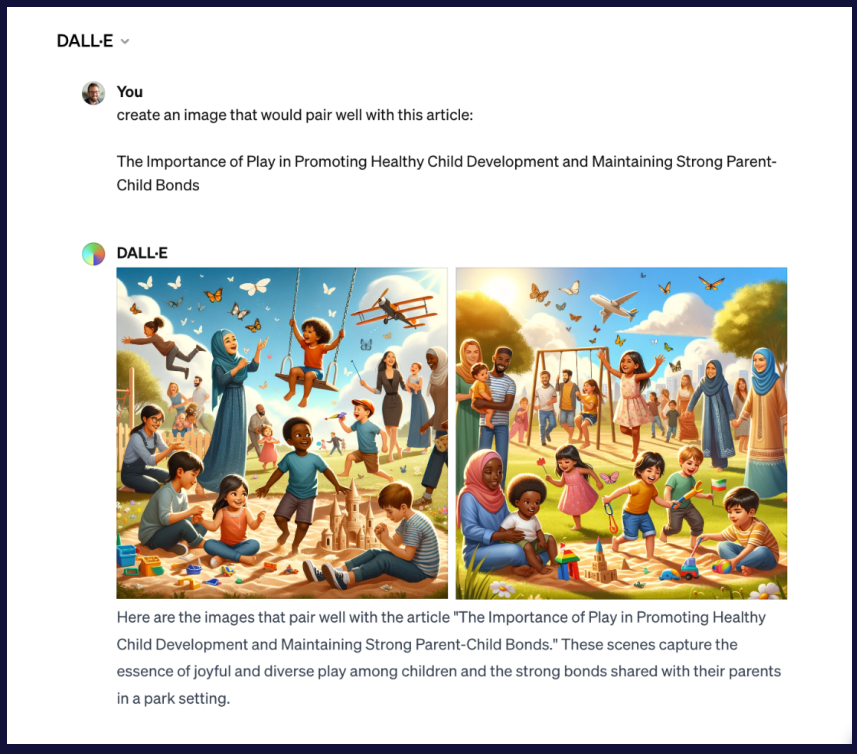
Big Ideas & New Frontiers

The future is LLMS, the future is “multi-modal” AI

ChatGPT made Large Language Models (LLMs) readily available and thrust a version of AI into the public consciousness.

Language models will remain central to AI since they’re able to capture and encode so much intelligence. They’re also crucial since publishers deal with so much text. But we’ll also see models “go multi-modal”, mixing text, images, video, audio and other media (including scientifically relevant ones like proteins, materials, and chemical compounds). For example:

- Climate modeling using satellite data, spotting fires and floods
- New materials prediction using data on the properties of crystals
- Hurricane forecasting using storm data and atmospheric images
- Drug development using protein fitness data to find ideal antibodies

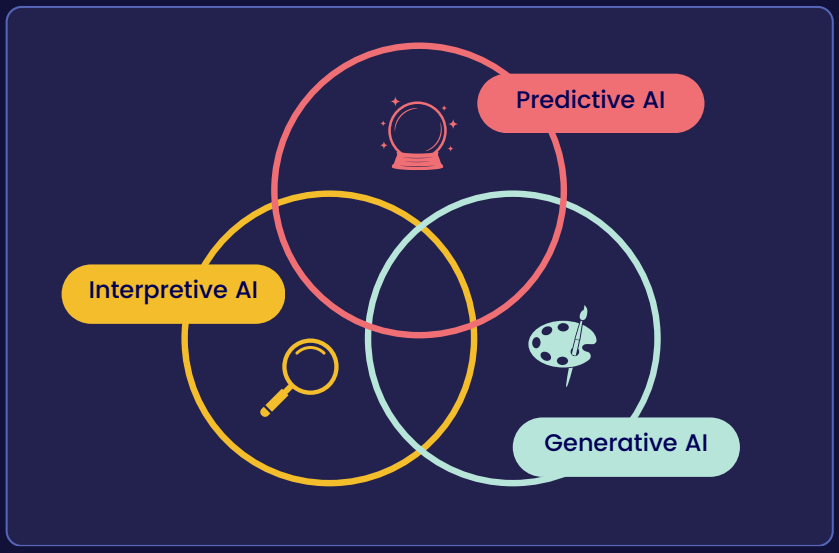


We’ll see this more in generalist models, too. ChatGPT Plus already does this with its integration with DALLE3, OpenAI’s image generator.

This example inputs text only, and outputs a mix of images and text (two modes, or “multi-modal”).

GPT 4.5 is rumored to include text, images, audio, video, and 3D images.

Three major types of AI models



- **Generative AI:** focused on generating new media from prompts, existing media or a combination of both
- **Interpretive AI:** focused on context, understanding and interpretation of media, often used to reason about large corpuses
- **Predictive AI:** using past data to predict future trends and behaviors

When it comes to LLMs, there are two primary flavors: Generative AI known for crafting text and Interpretive AI, best suited for content understanding.

Generative LLMs are trained by “reading” masses of documents (blogs, books, websites) and predicting the next word. They do this millions of time and start to pick up tricks, understanding language, how the world works,

Interpretive LLMs learn by predicting a missing word from masses of documents (blogs, books, websites, similar to generative models). They function by generating "text embeddings" from input content. These embeddings are high-dimensional numeric representations capturing the essence of text—everything from the main point of the author to tone, format & style. These text embeddings, represented with 768 or more dimensions, form a rich landscape for encoding human writing.

They are especially useful when comparing one text with many others–for example, comparing a new paper submission to existing published papers to find corroborating research or missing citations.

Interpretive and Generative AIs are highly complementary, and can be used together to power use cases like conversational search.

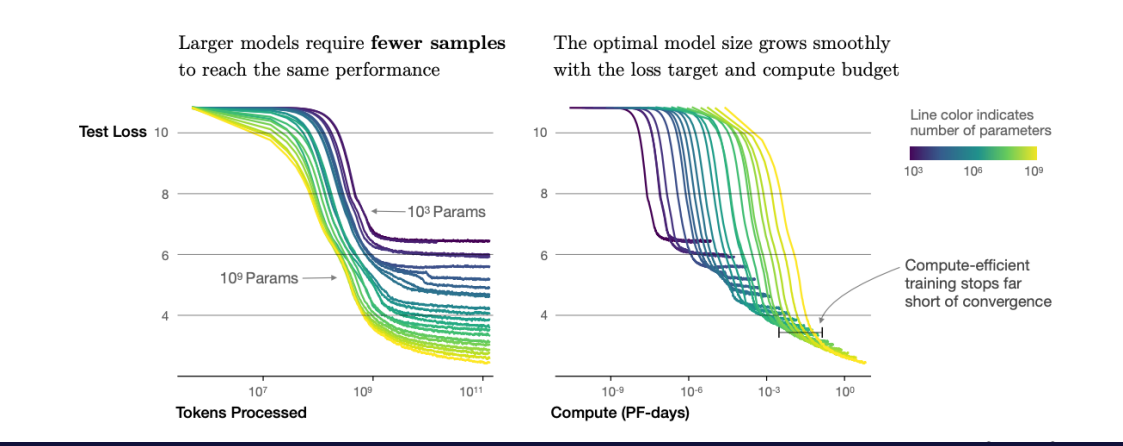
There are blessings to scale

Bigger models
+ More compute
+ More data =
Stronger models *

Increasing the size of models, expanding the datasets they train on, and employing greater computational power leads to stronger AI capabilities.

This pattern suggests a consistent return to scale, where larger investments in these areas yield significant improvements. So far, there's no clear indication of diminishing returns.

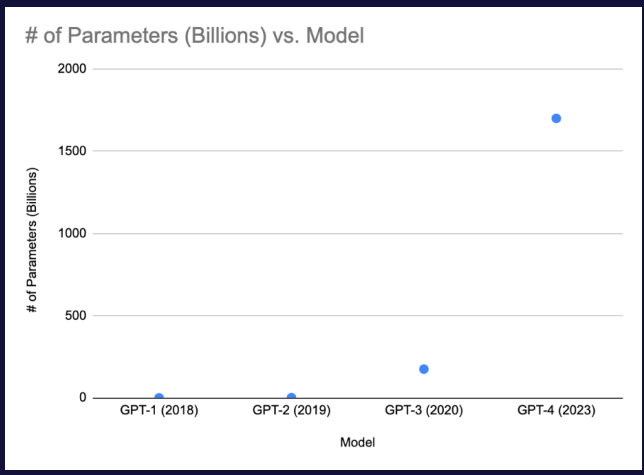
> * **"The scaling hypothesis:** neural nets absorb data & compute, generalizing...manifesting new abilities even at trivial-by-global-standards-scale. The deep learning revolution has begun as foretold."
>
> - gwern branwen



From Scaling Laws for Neural Language Models

AI brains improve as they get larger, consume more data, and use more computing power. We've not seen signs of a slow down in improvement with scale.

Take OpenAI's GPT series of models. You can see the size of the model expand significantly. So much that it's hard to render on the same graph →



LLMs are (massive) multi-task learners

Using a very simple training objective - *predict the next word* - on a diverse dataset, you can teach an LLM many skills.

TASKS	TYPES OF PRE-TRAINING SOURCE DATA	EXAMPLE SENTENCE FROM PRE-TRAINING SOURCE DATA
Grammar correction	Texts with annotations for correct and incorrect grammar usage	"She *go to the store every morning." *(corrected: "She goes to the store every morning.")*
World knowledge acquisition	Encyclopedic entries, factual databases, and verified information sources	"The Nile River is the longest river in the world, flowing through eleven countries in Africa."
Summarization	Long form articles with corresponding summaries	"Quantum computing harnesses the phenomena of quantum mechanics to deliver a huge leap forward in computation to solve optimization problems."
Logical reasoning	Texts containing logical puzzles, arguments, and reasoning scenarios	"All roses are flowers. Some flowers fade quickly. Therefore, some roses fade quickly."
Spatial reasoning	Descriptions of spatial puzzles, geometry problems, and navigation instructions	"If you are facing north and turn 90 degrees to your right, you will be facing east."
Creative writing	Literature, poetry, narrative essays, and various forms of creative text	"The sky was a canvas of oranges and purples as the sun sank into the horizon."
Hypothesis generation	Scientific papers, research proposals, & hypothesis-driven discussions	"If dark matter interacts weakly with light, it would not be visible through standard telescopic observation."

The training task is very simple, but there's massive richness in the data which it extracts.

The models are the data

AI models are structured, compressed, useful versions of the data that fed them.

A model like Llama2 is trained on two trillion tokens of language data from sources like Wikipedia, Reddit, internet blogs, academic papers, etc.

In model training, Llama2 is required to read a piece of text and predict the next token. It gets a reward (yay!) if it gets the word right and a penalty (oh, no!) if it gets it wrong.

> It's becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree... Trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point.
>
> This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It's determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivering compute to approximating that dataset.
>
> Then, when you refer to "Lambda", "ChatGPT", "Bard", or "Claude" then, it's not the model weights that you are referring to. It's the dataset.
>
> *From The "it" in AI models is the dataset.*

Image generation models like those found in Stable Diffusion or DALL-E 3 are trained on text-image pairs. These diffusion models work by gradually adding noise to an image and then learning to reverse the process.



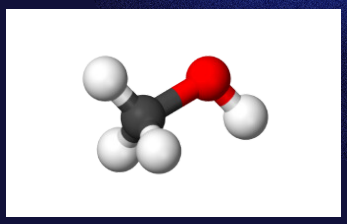
From Dall-E 3

While AI models can synthesize from within their training data (you can combine an avocado + spoon + traditional therapist setup into “a spoon therapist in a session with an avocado patient”), they struggle or fail to represent things outside their training data.

Likewise, if the training data quality is mixed (some good, some bad), the model quality will suffer.



It’s trained to make you happy, so it will give it a go regardless. This is methanol’s actual model:



Data cleanliness and quality are the next frontier

Since models are the data, high quality data improves models.

Early language models trained on mixed-quality internet data, including "low quality web data" including spam, plagiarized content, or poorly-written web forums.

As of writing, one of the most powerful open-source LLMs is DeepSeek-67B. It comes to us from a previously-obscure Chinese team that says this about its data preparation:

- **Data Composition:** Our training data comprises a diverse mix of Internet text, math, code, books, and self-collected data respecting robots.txt. In addition to the diverse content, we place a high priority on personal privacy and copyright protection. All content containing personal information or subject to copyright restrictions has been removed from our dataset.

- **Dataset Pruning:** Our system employs heuristic rules and models to refine our training data. Our filtering process removes low-quality web data while preserving precious low-resource knowledge. It aims to improve overall corpus quality and remove harmful or toxic content.

- **Deduplication:** Our advanced deduplication system, using MinhashLSH, strictly removes duplicates both at document and string levels. This rigorous deduplication process ensures exceptional data uniqueness and integrity, especially crucial in large-scale datasets.

Alibaba just released a data distillation framework, Data Juicer, and accompanying paper.

Data-Juicer is a one-stop data processing system to make data higher-quality, juicier, and more digestible for LLMs.



There has been extensive work by groups like Allen Institute for AI (Dolma) and Together AI (Red Pajama) to prepare and open source large, clean, diverse datasets. Dolma is 3 trillion tokens, and Red Pajama is 30 trillion tokens! Large, clean, diverse.



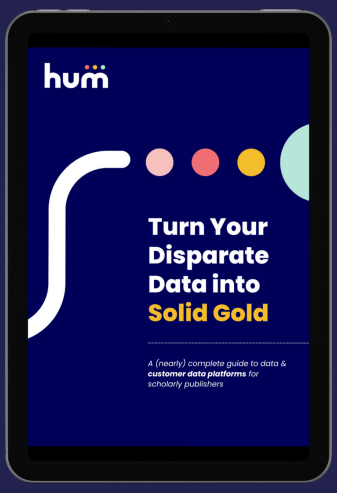
Data filtering pipeline from Allen Institute

Feeding models a healthy diet produces better models.

Brushing up on your data know-how?

The complete guide to data and CDPs for publishers is a great place to start!

>> Get the Whitepaper



Putting AI to Work in Publishing Now

AI in scholarly publishing: Use cases, applications, and examples in the wild

Beyond a prototype scale, you need both AI and data to make AI-driven products, applications and workflows viable. We'll call out which are essential for each of the items below.

INGREDIENTS

AI



Interpretive AI



Predictive AI



Generative AI

Data



Content data



User historical data

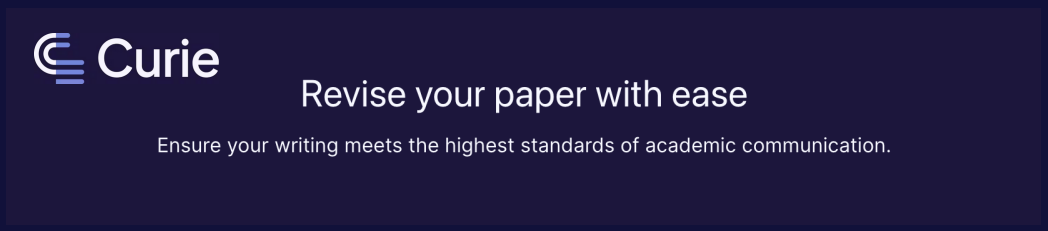


User input

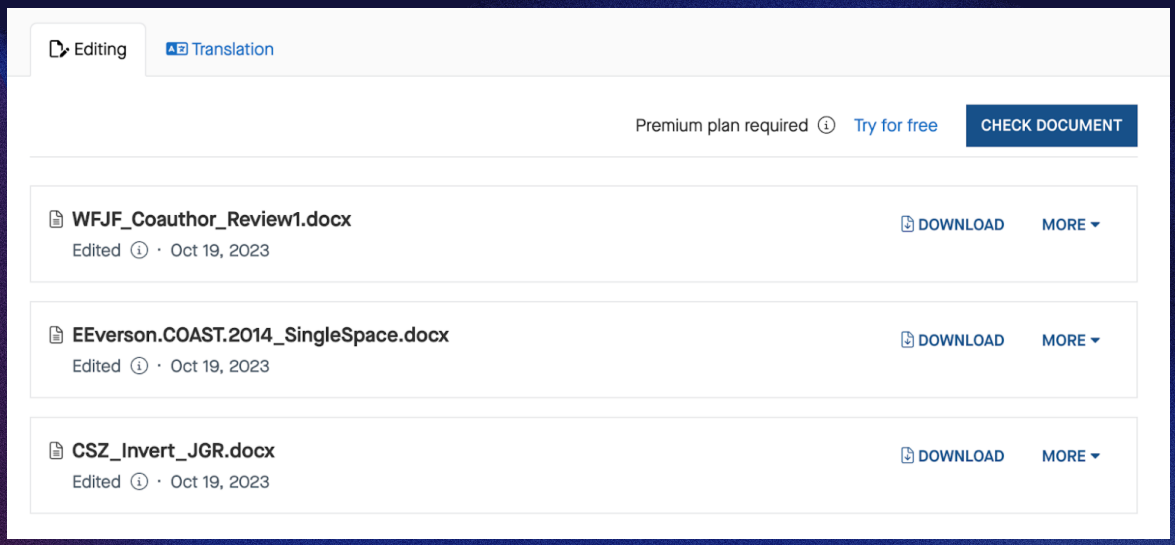
Writing upgrades - e.g. improving writing quality

The idea here is to improve an in-progress or completed manuscript, review, grant proposal or other document.

This requires the user input (written artifact), a careful prompt, and an LLM (likely a frontier model at this point due to variability in the writing).

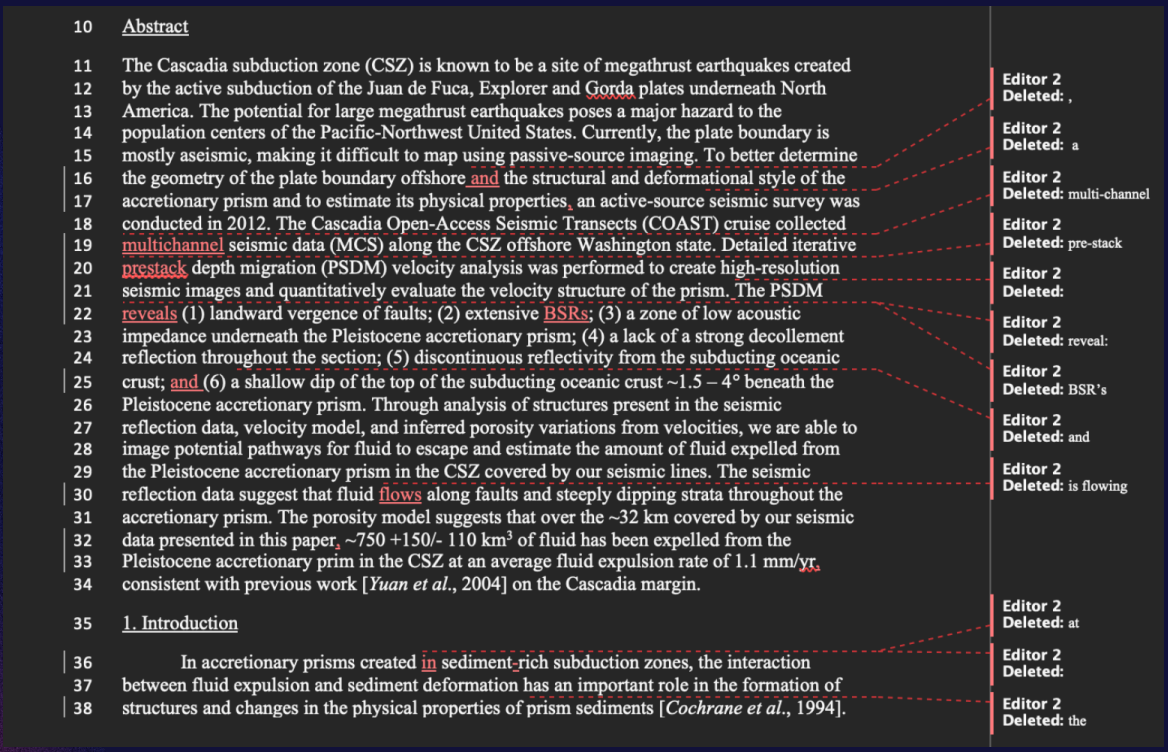


One key example in the wild is Curie, recently released by SpringerNature. It's integrated into their manuscript submission workflow, but you can also use it as a standalone product. For the latter, you upload documents and get a choice of (copy)editing or translation.



We used a few papers and reviews from Will Fortin, Hum's Lead Data Scientist (and former marine seismologist) to test it out.

After upload, you're able to download a Microsoft Word document with the suggested changes in Track Changes mode. This is a nice example of tying into (some!) people's workflow.



10 Abstract

11 The Cascadia subduction zone (CSZ) is known to be a site of megathrust earthquakes created
12 by the active subduction of the Juan de Fuca, Explorer and Gorda plates underneath North
13 America. The potential for large megathrust earthquakes poses a major hazard to the
14 population centers of the Pacific-Northwest United States. Currently, the plate boundary is
15 mostly aseismic, making it difficult to map using passive-source imaging. To better determine
16 the geometry of the plate boundary offshore and the structural and deformational style of the
17 accretionary prism and to estimate its physical properties, an active-source seismic survey was
18 conducted in 2012. The Cascadia Open-Access Seismic Transects (COAST) cruise collected
19 multichannel seismic data (MCS) along the CSZ offshore Washington state. Detailed iterative
20 prestack depth migration (PSDM) velocity analysis was performed to create high-resolution
21 seismic images and quantitatively evaluate the velocity structure of the prism. The PSDM
22 reveals (1) landward vergence of faults; (2) extensive BSRs; (3) a zone of low acoustic
23 impedance underneath the Pleistocene accretionary prism; (4) a lack of a strong decollement
24 reflection throughout the section; (5) discontinuous reflectivity from the subducting oceanic
25 crust; and (6) a shallow dip of the top of the subducting oceanic crust ~1.5 – 4° beneath the
26 Pleistocene accretionary prism. Through analysis of structures present in the seismic
27 reflection data, velocity model, and inferred porosity variations from velocities, we are able to
28 image potential pathways for fluid to escape and estimate the amount of fluid expelled from
29 the Pleistocene accretionary prism in the CSZ covered by our seismic lines. The seismic
30 reflection data suggest that fluid flows along faults and steeply dipping strata throughout the
31 accretionary prism. The porosity model suggests that over the ~32 km covered by our seismic
32 data presented in this paper, ~750 +150/- 110 km³ of fluid has been expelled from the
33 Pleistocene accretionary prim in the CSZ at an average fluid expulsion rate of 1.1 mm/yr,
34 consistent with previous work [*Yuan et al.*, 2004] on the Cascadia margin.

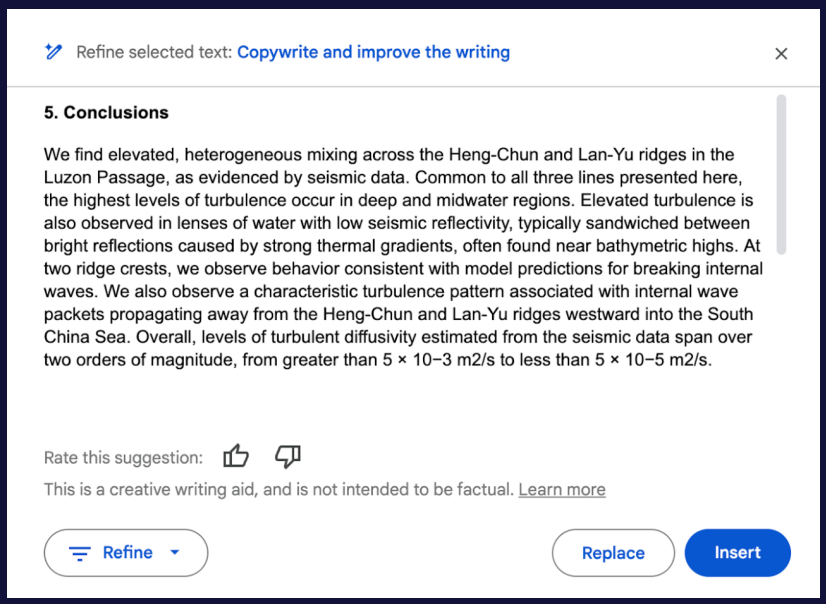
35 1. Introduction

36 In accretionary prisms created in sediment-rich subduction zones, the interaction
37 between fluid expulsion and sediment deformation has an important role in the formation of
38 structures and changes in the physical properties of prism sediments [*Cochrane et al.*, 1994].

Editor 2 Deleted: ,
Editor 2 Deleted: a
Editor 2 Deleted: multi-channel
Editor 2 Deleted: pre-stack
Editor 2 Deleted:
Editor 2 Deleted: reveal:
Editor 2 Deleted: BSR's
Editor 2 Deleted: and
Editor 2 Deleted: is flowing
Editor 2 Deleted: at
Editor 2 Deleted:
Editor 2 Deleted: the

This would be an obvious add on for Overleaf and other services.

You will also see this sort of capability become available via Microsoft's Co-Pilot (powered by OpenAI's GPTs) and Google's Duo (powered by their own LLMs, presently PaLM2).



AI as Peer Reviewer

AI models can and will be used as part of the peer review process. One key question is what is the right part of the workflow?



WATCH:
Artificial intelligence (AI) and peer review

> "[I'm] very interested in the idea of a process where Reviewer 1 is an AI service, and the resulting comments are made available to the subsequent human reviewers. Useful AI comments could be cherry picked for the author's information. The burden of language reviewing (and in the future, statistical evaluation) lifted off the human reviewers, leaving them to focus on the scientific quality."
>
> - Katy Ladbrook, British Geriatric Society

It's easy to think of an AI as "Reviewer #4," working alongside human reviewers, but it's likely more effective as an earlier step in the process.

Below is a very basic prompt we've used with GPT-4:

{Article}

Above is a scientific journal article. You are a world-class editor with deep expertise in oceanography. Your candid and concise feedback to the author is crucial for them to keep their job.

Please:

1. Take a deep breath and think step by step.
2. Provide a tough but fair critique
3. Suggest the top three science improvements. Note: They should be reasonable, factoring in what results or interpretations are plausible given the data and/or experiments.
4. Identify any sections that are overly verbose or poorly written. Cite them specifically

Here is an excerpt of the response:

Top Three Science Improvements
Enhanced Discussion on Limitations: While the seismic method is innovative, it's crucial to acknowledge its limitations in detecting certain types of waves or features. A more thorough discussion on this, perhaps comparing with traditional oceanographic methods, would strengthen the manuscript.

Further Analysis of Turbulent Diffusivity Data: The manuscript could benefit from a deeper analysis of the turbulent diffusivity data. This might include comparisons with existing models or data from other regions to contextualize the findings within the broader field of oceanographic research.

Additional Validation of Results: Considering the novelty of the approach, further validation of the results, possibly through comparison with satellite or in situ data, would add robustness to the findings.

Sections Needing Revision for Verbosity or Clarity
***Section 3**, Methods: The explanation of the slope spectral methods and the calculation of turbulent diffusivity is quite technical and dense. Simplifying this section or providing a more intuitive explanation of the methods could make the paper more accessible to a wider audience.*

***Section 4**, Results: Some parts of the results, especially where seismic line data are described (e.g., Line 4, Line 5, and Line 9 descriptions), are overly detailed for the average reader. Focusing on the key findings and reducing the level of granular detail could improve readability.*

***Section 1.2**, Seismic Oceanography: While informative, this section feels slightly disconnected from the main narrative. Integrating it more smoothly with the rest of the introduction or reducing its length could improve the flow of the manuscript.*

You can see the full example output HERE.



ACS Journal Finder from ChronosHub

Using a similar approach to the ACS Journal Finder, imagine a researcher dropping an article title, abstract, or full manuscript into a web interface.

That blob of text is then semantically matched (using an interpretive LLM and content embeddings) to produce a rank ordered list for ideal journals to publish in (or federation of publishers, like in the case of GeoScienceWorld or BioOne).

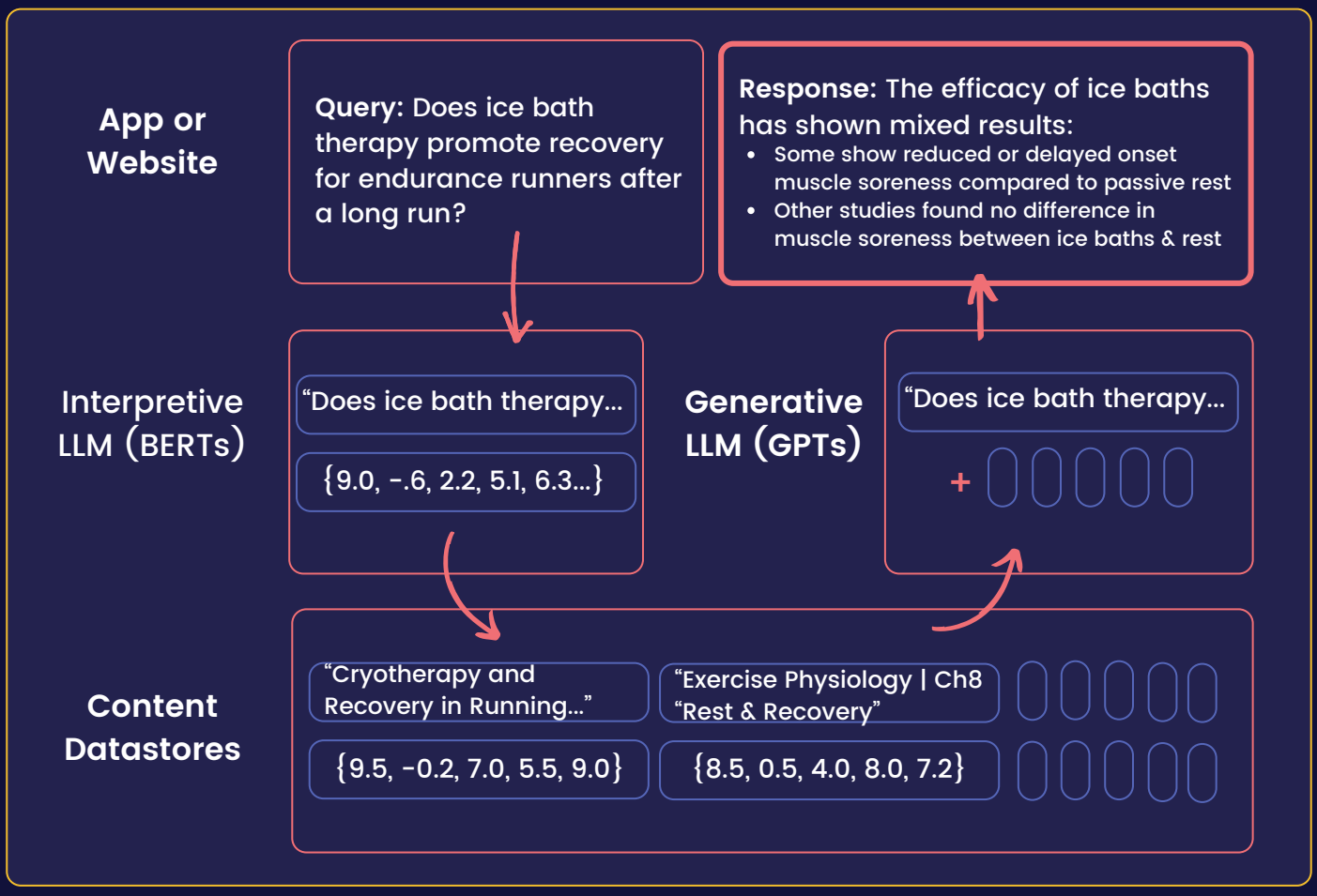
The user might have further filters and controls on things like time to publish, impact factor, or APC cost.

New User Experiences: Conversational Search

Imagine you wanted to build "Conversational Search" for your publishing site. Something like the offerings found from Scite, Consensus, Digital Science, and Scopus.

You need to rely both on the breadth of interpretive AI models to understand your full content corpus (and get around Generative AI models' context window limitations), and on the creativity of GPTs to generate an informed response.

Here's the approach:



Interpretive Layer:

First, you'll need to translate your content corpus into embeddings format. We're biased, but we'd recommend Hum's Lodestone-v1, an open-source embeddings model optimized for long sequences, or Hum's Embeddings as a Service, based on a specialized version of Lodestone-v2.
You can also use OpenAI's ada embeddings, Allen Institute's SciBERT, or others.

Next, you'll translate the user query into embeddings format - the query is a phrase or sentence that can be matched to other phrases, sentences, paragraphs, and documents within your corpus.

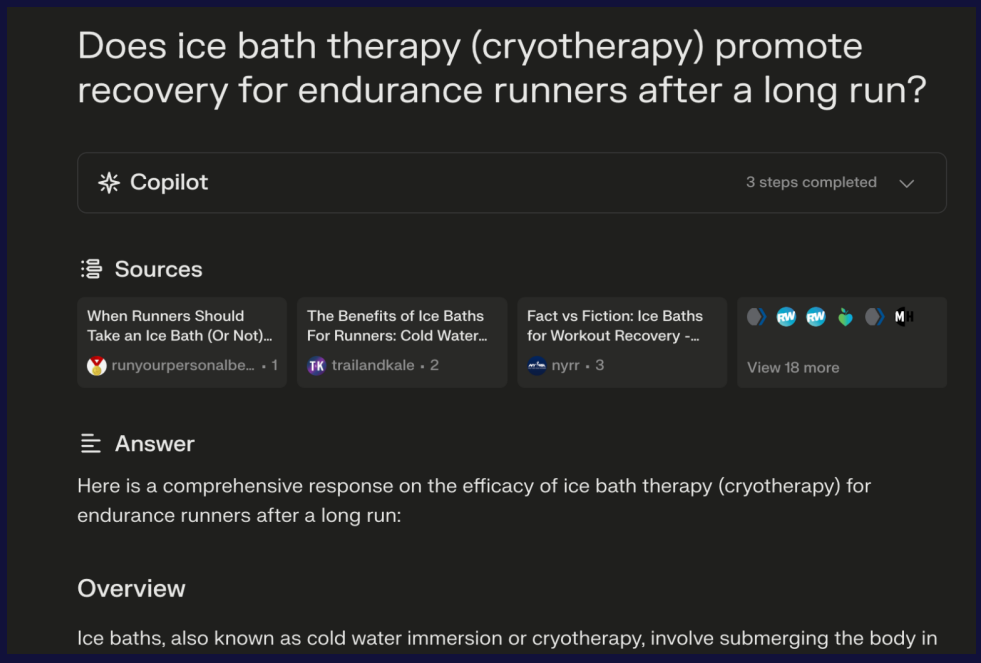
The best matching chunks will then be fed into the generative LLM, along with the original prompt

Generative Layer:

For this to work, you'll pass three things to the LLM:

1. *The initial user query - e.g. "Does ice bath therapy promote recovery for endurance runners after a long run?"*
2. *The retrieved chunks of context (reference points to use in addition to the LLM's own training data)*
3. *An explanation of what the chunks are and how to use them.*

What you're then able to output is a response that's grounded by your content corpus, potentially enhanced with citations from individual pieces of content.



Here's an example output using the example query above using Perplexity.AI, which uses either Claude2 or GPT-4 (which was used here) as its generative LLM under the hood and has internet access (it will pull chunks from various internet sources).

New User Experiences: AI-Generated Media



LLMs are great at creating and transforming text. In this case, the key use is working from existing content to create other useful media.

While a journal article or book captures key information well, it's time consuming for experts to consume and often not accessible to the general public.

We'll demonstrate five examples below:

- 9th Grade Summary: unpack the article for a smart 9th grader
- Expert Summary: dense summary to help an expert determine whether to read the paper
- Research Highlights: generate slide input for a paper author for presentations & seminar talks
- Questions for the Author: help audience members prepare to engage an author on their research
- Generate Hypothesis: help early career researchers use this research as a starting point for future work

Let's start with an astronomy paper by UVA astronomer Ilse Cleeves and team.

The Coupled Physical Structure of Gas and Dust in the IM Lup Protoplanetary Disk

Show affiliations

Cleeves, L. Ilsedore ; Öberg, Karin I. ; Wilner, David J. ; Huang, Jane ; Loomis, Ryan A. ; Andrews, Sean M. ; Czekala, Ian

The spatial distribution of gas and solids in protoplanetary disks determines the composition and formation efficiency of planetary systems. A number of disks show starkly different distributions for the gas and small grains compared to millimeter-centimeter-sized dust. We present new Atacama Large Millimeter/Submillimeter Array observations of the dust continuum, CO, ^{13}CO, and C^{18}O in the IM Lup protoplanetary disk, one of the first systems where this dust-gas dichotomy was clearly seen. The ^{12}CO is detected out to a radius of 970 au, while the millimeter continuum emission is truncated at just 313 au. Based upon these data, we have built a comprehensive physical and chemical model for the disk structure, which takes into account the complex, coupled nature of the gas and dust and the interplay between the local and external environment. We constrain the distributions of gas and dust, the gas temperatures, the CO abundances, the CO optical depths, and the incident external radiation field. We find that the reduction/removal of dust from the outer disk exposes this region to higher stellar and external radiation and decreases the rate of freeze-out, allowing CO to remain in the gas out to large radial distances. We estimate a gas-phase CO abundance of 5% of the interstellar medium value and a low external radiation field ($G_0 \approx 4$). The latter is consistent with that expected from the local stellar population. We additionally find tentative evidence for ring-like continuum substructure, suggestions of isotope-selective photodissociation, and a diffuse gas halo.

9th Grade or “Lay Summary”

Overview: This research focused on understanding the amounts and types of gases like carbon (C), nitrogen (N), and oxygen (O) in a young planetary system called the IM Lup protoplanetary disk. They used observations from a powerful telescope called ALMA.

Why This Matters: Protoplanetary disks are clouds of gas and dust around young stars, where planets start to form. By studying these disks, scientists can learn about the materials that eventually make up planets. This particular study looked at the gases in one of these disks, which is important because the gas can end up on planets as their atmosphere.

Key Discoveries:

- Gas Ratios: They found that the ratio of carbon to oxygen (C/O) in the gas is about 0.8, which is higher than the sun's ratio of about 0.54. This means there's more carbon compared to oxygen in this young planetary system than in our sun.
- Water Ice and Carbon: A lot of the oxygen is locked up in water ice and larger dust particles that settle in the middle of the disk. Some of the carbon is also trapped in less volatile (less easily vaporized) molecules.
- Nitrogen: Nitrogen, on the other hand, is mostly found as a gas (likely N2) and doesn't get trapped in ice or dust as much as oxygen and carbon.

What This Means for Planets: Understanding these gas ratios is crucial because they can influence the type of atmospheres that planets develop as they form in these disks. For instance, having more carbon than oxygen could make the atmospheres of new planets very different from Earth's.

The Bigger Picture: These findings might help predict the composition of gases around other young stars and the atmospheres of planets forming there. It's like getting a sneak peek into how different planets might be born and what their air could be like, which is super important for understanding the variety of planets in our universe.

Expert Summary

Employing ALMA observations and chemical modeling, this study on the IM Lup protoplanetary disk focuses on gas-phase C, N, and O abundances. It finds a super-solar C/O ratio and a significant N/O ratio variation in the molecular layer, particularly evident between radii of 20 and 300 au in a pseudo-steady state. The transformation of H2O ice to gas-phase atomic oxygen and nitrogen’s stability in N2 form are key factors affecting these ratios, crucial for understanding planet formation processes.

Questions for the Author

1. Methodological Clarifications: Can you elaborate on how the ALMA observations were specifically used to determine the C/O and N/O ratios?
2. Findings Interpretation: How does the super-solar C/O ratio in the IM Lup disk compare with other known protoplanetary disks?
3. Chemical Processes: Could you explain why nitrogen appears less chemically processed in the disk compared to carbon and oxygen?
4. Broader Impact: How do these findings impact our understanding of the chemical composition of exoplanet atmospheres?
5. Future Directions: What are the next steps in this line of research following your findings on the IM Lup disk?

Generate Hypothesis for Future Research

Hypothesis:
"Differential Processing of Volatiles in Protoplanetary Disks and Its Effect on Planet Formation"

Rationale: Building on the findings from the IM Lup disk study, which uncovered a super-solar C/O ratio and distinctive processing of nitrogen, this hypothesis proposes that the differential processing of volatiles like carbon, oxygen, and nitrogen in protoplanetary disks plays a crucial role in determining the chemical composition and potential habitability of forming exoplanets.

Potential Future Research Directions:

- Comparative Studies: Investigate the volatile compositions of other protoplanetary disks using similar methodologies to understand if the IM Lup disk's characteristics are common or anomalous.

- Modeling Variations in Volatile Processing: Develop and test models that simulate how different levels of volatile processing impact planet formation and the resulting atmospheric compositions.

- Link to Exoplanet Atmospheres: Explore the potential correlations between the volatile compositions of protoplanetary disks and the observed atmospheric compositions of exoplanets to validate the hypothesis.

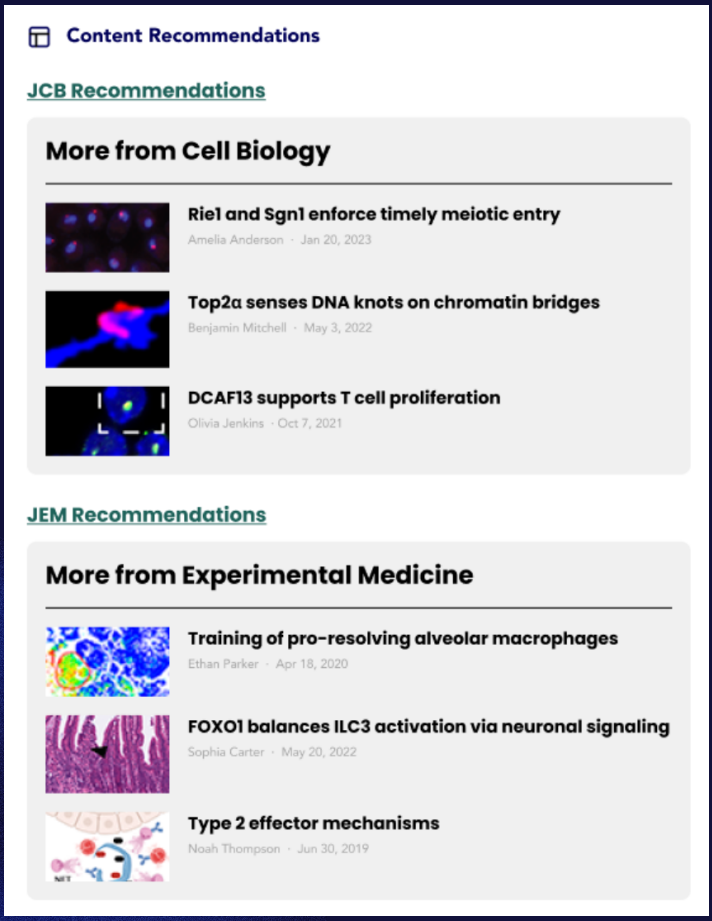
Better User Experience: Content Recommendations

Leading publishers have built custom systems to power recommendations, but AI has the potential to make this process more accurate and scalable. For instance, this is a native feature in Hum's CDP.

Today's audiences, entrenched in constant customization available from platforms like Amazon and Netflix, expect increasing levels of personalization and curation from digital experiences. Using reader data and content embeddings, Alchemist can match users to new and existing content articles similar to their past engagement. These recommendations improve as the system learns from incoming behavioral signals.

Beyond on-site suggestions, unlocking the right data + AI means publishers can deliver:

- Personalized emails with curated recommendations.
- Context-aware recommendations, like suggesting related articles on a topic page.
- Recommendations of appropriate journals for submitted manuscripts, using abstract embeddings and journal fingerprint profiles.

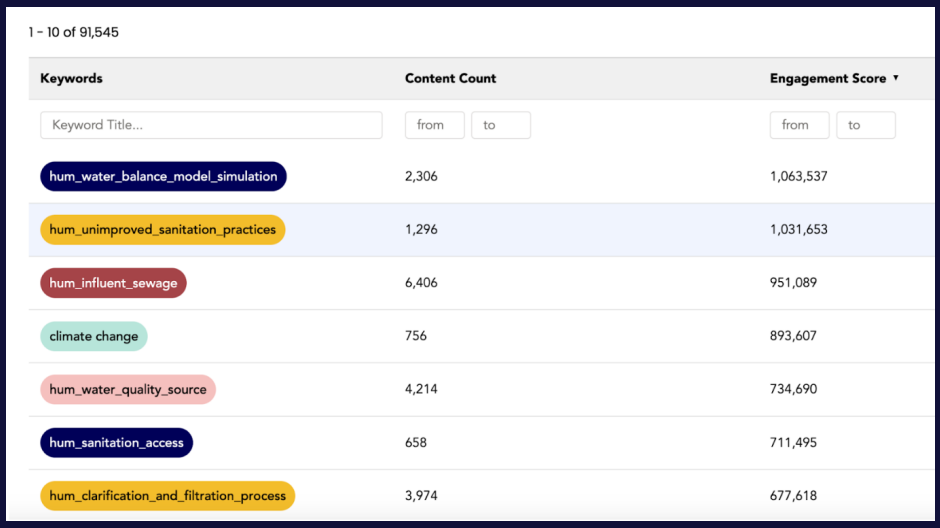


Content Tagging



Good tagging can improve discoverability, recommendations, search, audience intelligence and more. But content tagging has always been a problem. How do you decide on a taxonomy (or keyword set)? How do you apply that to content?

AI can solve both of those problems, something we know well from our work with the Alchemist Tagging Engine.



Automatic Taxonomy Application

You can "embed" taxonomies into interpretive AI models. This works for an Alchemist derived taxonomy, third party taxonomies (IAB, MESH, OpenAlex Concepts), or an in-house taxonomy. It also works for word lists and keyword sets. Once the terms are embedded into the model format, you can match them to documents, looking for the "semantic similarity" of the term/topic/keyword to that of the document.

Taxonomy Building

If you turn your content into embeddings format, you can "see" it in a high dimensional space, including interesting clusters. We pull candidate terms from clusters to find good candidate keywords. We then use several other clustering and combination techniques in order to combine like terms and get the topics to the right level.

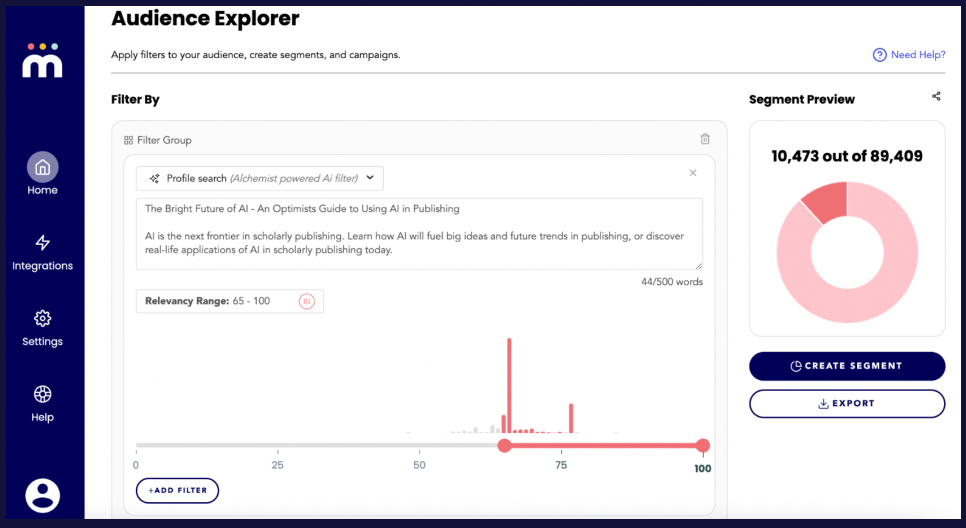
Behavioral Segmentation

Interpretive AI: Funneling People in the Embeddings Space

Hum has been working on a frontier application of embeddings: representing people in the embeddings space! It works on the principle of "you are what you eat." With the proper systems in place to capture behavioral data (e.g. a Customer Data Platform like Hum), you can observe how people travel within the same embedding space as the content they're consuming. Just as text embeddings allow you to connect the content from a proposal to existing content within your journal, you can take the description of a new special topic section and find people who's content consumption suggests they'd be interested in reading or even submitting to it.

Applications include:

- People Search: Identifying people matching descriptions from products, papers, or events.
- Predictive Interests: Anticipating an individual's future actions/.interests based on audience trends.
- Live Adaptation: developing models and web features that adapt as new content is added to your corpus - without expensive retraining.



Take, for example, a segment we started building in Hum to narrow down people in our audience that might be interested in the whitepaper you're reading now. Hum's Profile Search feature matches our content description to audience members who are likely to have interest in the ways AI will influence the future of scholarly publishing.

R&D, Future Trends to Watch

Ways AI Continues to Improve

Future Trend 1: Massive models, small models

Massive models

Expect models to continue to scale, and for there to be benefits to scale.

Anthropic says that it plans to build a “frontier model” — tentatively called “Claude-Next” — 10 times more capable than today’s most powerful AI, but that this will require a billion dollars in spending over the next 18 months. (TechCrunch)

DeepMind co-founder and Inflection CEO Mustafa Suleyman predicts that AI will continue its exponential progress, with orders-of-magnitude growth in model training sizes over the next few years. AI training models will be "1,000x larger in three years" (Future Timeline)

Bigger models (+ more data and compute) have shown to be more powerful models. Expect massive models to demonstrate frontier capabilities. They’ll have a role in performing unique capabilities, especially where use cases involve diverse data or capabilities (you want the biggest “brain” brain in the room for those).

Small models

Small models can cost between 100-1,000x less than frontier Large Language Models.

> My prediction: in 2024, most companies will realize that smaller, cheaper, more specialized models make more sense for 99% of AI use-cases. The current market & usage is fooled by companies sponsoring the cost of training and running big models (especially with cloud incentives).
>
> - Clement Delangue (CEO, HuggingFace)

The biggest opportunity in small language models is using cleaner data. Since the models are the data, if you remove junk from the training set, you can pack the same intelligence into a smaller brain.

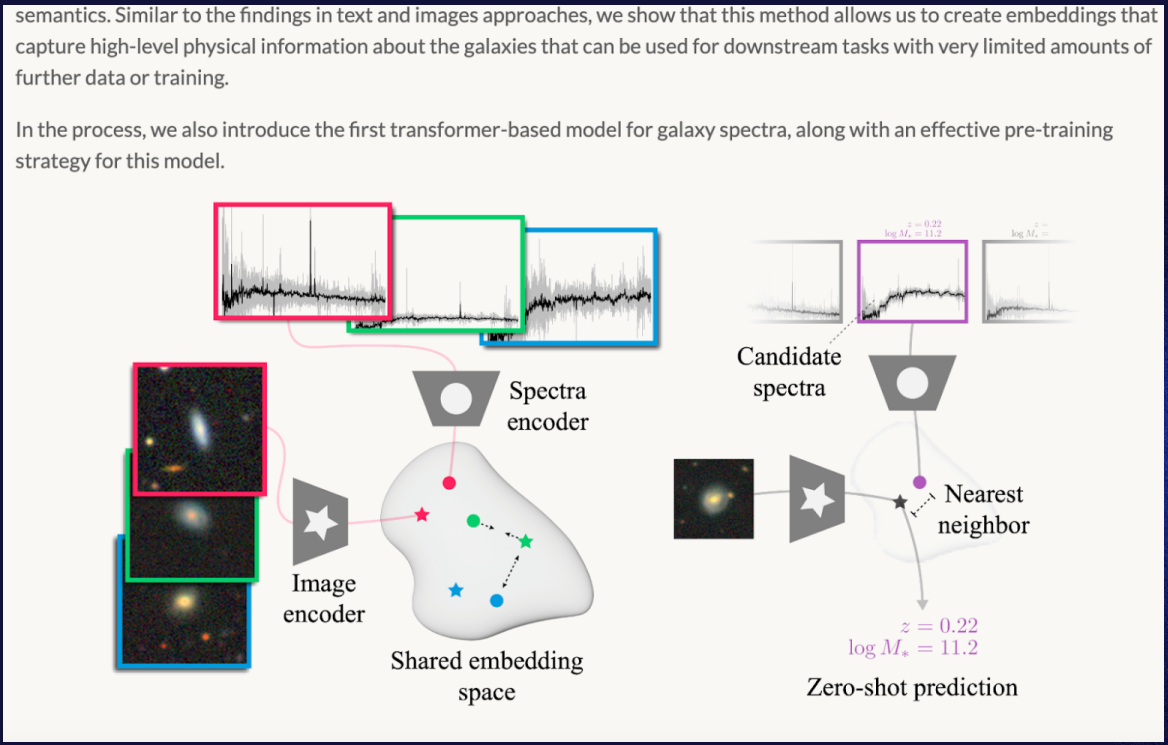
A recent example is Phi-2 by Microsoft:

We are now releasing Phi-2(opens in new tab), a 2.7 billion-parameter language model that demonstrates outstanding reasoning and language understanding capabilities...on complex benchmarks Phi-2 matches or outperforms models up to 25x larger, thanks to new innovations in model scaling and training data curation.

For more straightforward tasks like summarization, small models are often the best choice.

Future Trend 2: New model architectures

Polymathic - Multimodal models for science



From Cross-Modal Pre-Training for Astronomical Foundation Models

New startup Polymathic is doing very interesting things in trying to capture text, figures, and images all in the same knowledge (embeddings) space. Their aim is "to develop, train, and release such foundation models for use by researchers worldwide."

Where generic LLMs or even other multi-modal models may struggle with scientific inputs, this approach may succeed. In big part because the models are the data.

System 1 vs System 2 "Thinking"

LLMs produce answers with a fixed amount of computation. This is akin to the human fast and subconscious System 1 decision process as described by Daniel Kahneman in "Thinking Fast and Slow".

True reasoning and planning would allow the system to search for a solution, using potentially unlimited time. This is akin to human deliberate and conscious System 2 decision process.

Some AI systems have planning abilities, namely those that play games or control robots. These systems are still fairly limited in their planning abilities, compared to animals and humans.

To have more general planning abilities, an AI system would need to possess a world model. With a world model, a system can plan a sequence of action so as to fulfill an objective.

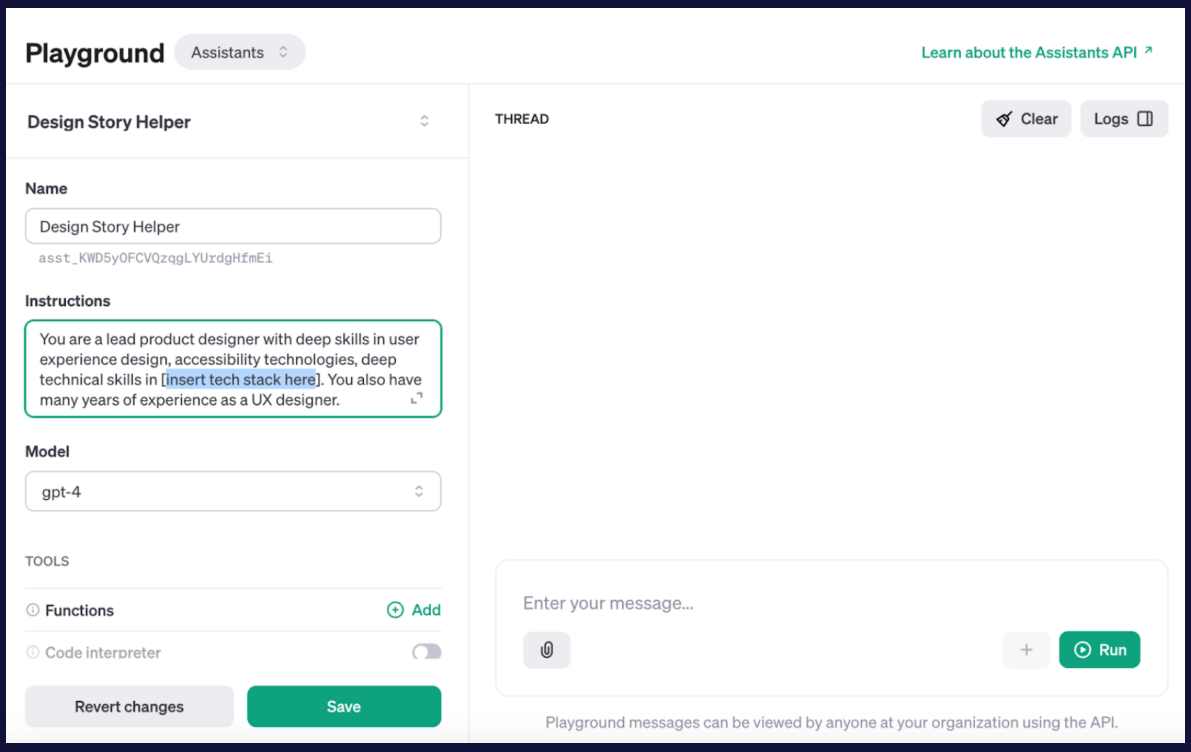
Building and training world models is still a largely unsolved problem. Decomposing a complex objective into a sequence of sub-objectives is even more complex. This would enable hierarchical planning, something that humans and many animals can do effortlessly but is still completely out of reach of AI systems.

For instance, you might want to give an AI "brain" a set of gnarly problems and the ability to "think on them for a while".

Future Trend 3: Agents and tools

AI agents, also known as intelligent agents, are autonomous programs that interact with their environment, make decisions, and take actions to achieve specific goals. They can perceive their environment through sensors, process the information they gather, and act upon that environment through actuators or effectors. This cycle of perceiving, thinking, and acting is continuous and allows the AI agent to function independently, without constant human intervention.

While a language model can provide text output given a prompt, agents can break down prompts into logical steps for accomplishing goals. They maintain internal state and memory to track progress, instead of just responding reactively. They can also act beyond conversing, through APIs, automation, and even robotics.



Hum uses an early version of an AI agent to help write and refine “story cards”, which are individual pieces of software work.

In a more advanced future, an e-commerce agent could autonomously navigate sites, compare products, and make purchases.

Hum uses an early version of an AI agent to help write and refine “story cards”, which are individual pieces of software work.

In a more advanced future, an e-commerce agent could autonomously navigate sites, compare products, and make purchases.

For publishers, agents unlock new use cases like automated content creation, document analysis, and process automation. Internally focused "research assistant" agents could help employees work more efficiently. Externally, reader-facing agents could provide a sophisticated personal concierge service, understanding user contexts and needs over time rather than single interactions.

The “multi-agent” future is one you will have to prepare for: human and robot audiences; human and robot employees.

Things get crazier when you introduce tools:

- Ability to access the internet
- Database access (e.g. Wolfram Alpha, structured knowledge)
- Other LLMs
- Calculators and software programs
- Bespoke code...that the agents themselves write!

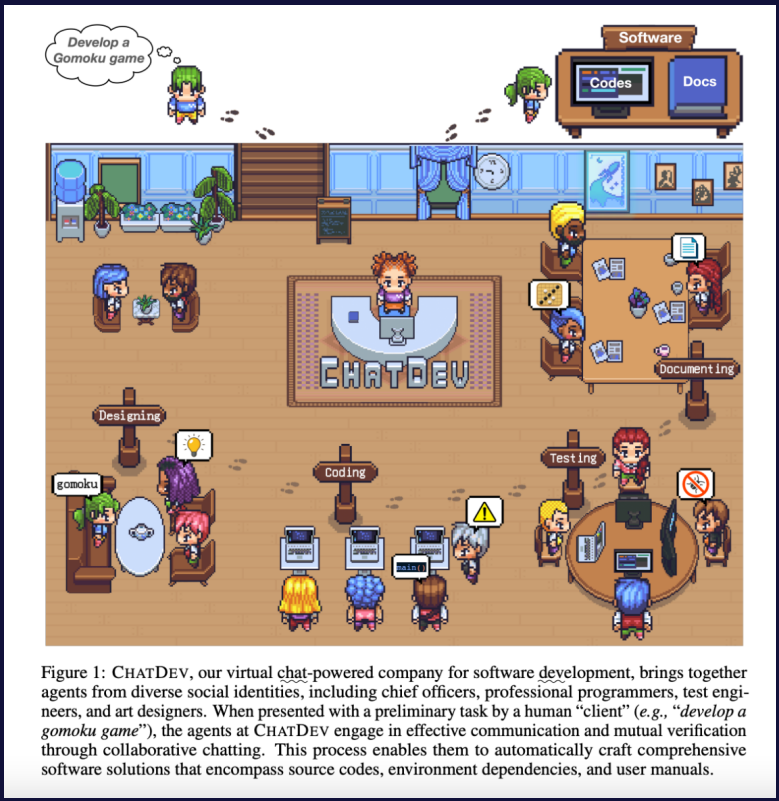


Figure 1: CHATDEV, our virtual chat-powered company for software development, brings together agents from diverse social identities, including chief officers, professional programmers, test engineers, and art designers. When presented with a preliminary task by a human “client” (*e.g.*, “*develop a gomoku game*”), the agents at CHATDEV engage in effective communication and mutual verification through collaborative chatting. This process enables them to automatically craft comprehensive software solutions that encompass source codes, environment dependencies, and user manuals.

In the same way this experiment looked at building a swarm of agents serving different software development roles, you could imagine the same for publishers.

Preparing for the AI Future Now

Editor's note:
We did not address two key issues regarding AI in scholarly publishing:

1. *Detecting AI-generated content in submitted manuscripts and peer review.*
2. *What to do about licensing content to AI model builders like OpenAI.*

In this white paper we focused on the (arguably underexplored) bright future aspects of AI, but we plan to cover these two issues in an upcoming blog post.

Determine key problems to solve, or needs to address

Think in terms of opportunities. Where can this emergent technology meet an unmet need?

We see many publishers doing workshops and hackathons to formulate a list of early candidates to address. And then they launch into an experimental phase.

Some needs to get you started!

- Increase researcher prestige
- Save researcher time
- Speed student learning
- Save an editor time
- Speed peer review
- Detect fraudulent manuscripts
- Navigate copyright and licensing
- Speed literature reviews
- Demonstrate impact to the public
- Translate research for other audiences

Experiment, test, & learn

Below is a framework for assessing where an organization, or individual project/product is on a maturity scale:

1. Getting Oriented: This initial stage involves understanding the AI landscape, identifying relevant technologies and trends, and recognizing potential applications within the context of scholarly publishing.

2. Ideation and Idea Selection: In this phase, organizations brainstorm potential AI applications, considering their unique challenges and opportunities, and then select the most promising ideas for further exploration.

3. Experimentation: At this stage, organizations begin practical experimentation with AI technologies, typically through small-scale, low-risk projects to test hypotheses and learn from practical experiences.

4. Prototypes and Testing: This involves developing functional AI prototypes based on the learnings from the experimentation stage and rigorously testing these prototypes to evaluate their effectiveness and feasibility. Often prototypes get early user or focus group feedback.

5. Alpha/Beta Product: Here, a more refined version of the AI solution is developed, often as an Alpha or Beta product, which is tested with a select group of users for feedback and further refinement. Potential for public release, depending on the product.

6. Scaled Production Product: In the final stage, the AI solution is fully developed and scaled up for broad distribution and integration into the organization's primary digital publishing processes, reflecting a mature application of AI technology.

You can think of using it to “climb the ladder” toward an AI product working at scale.



Experiment, test, & learn

You'll have several key considerations as you reach scaled products:

Cost. As you start out, you should tend toward off the shelf frontier models like GPT-4. As you scale, you'll want to consider smaller fine-tuned models that fit a particular use case. The cost difference can be as much as 100x. That makes a big difference for how much you can take advantage of AI.

Data privacy. OpenAI and other services say that they don't train on your data if you use their APIs. Many publishers are keeping their content out of OpenAI, Anthropic and the other generic model providers (they're worried about "accidental escape"). Alternatives include deploying your own open source models or using private AI clouds like the one provided by Hum.

> "OpenAI are training on every piece of data they see, even when they say they aren't" is the new "Facebook are showing you ads based on overhearing everything you say through your phone's microphone".
>
> - Simon Willison

Trust. Can you trust the model outputs? Hallucinations are a feature, rather than a bug of the initial wave of AI models. You'll want to carefully evaluate your models and model outputs. You may want to include a human-in-the-loop to ensure AI outputs, at least for a time. Recognizing the importance of trust, one of Hum's key areas of research in Generative AI is trustworthy LLMs.

Build internal capabilities

Where to start?

1. Use best-in-class technology. Optimize for speed/cost later.
2. To move quickly, you may have to buy the expertise or fill gaps with partners.
3. Longer-term, you'll need to build this muscle internally by doing and practicing. Now is the time to begin upskilling your team, and plan to bring in new talent: Data engineers, UX designers, and AI product managers.

Ship and iterate quickly

> "The single most important thing to understand about AI is how fast it is moving."
>
> - Dario Amodei, CEO of Anthropic

In an AI-driven world, scholarly publishers need to move quickly to experimenting with and integrating new AI tools. Because of how fast AI is moving, you need to be "building with it" to have more than a surface level understanding and keep up. It's not just about competition within the industry.

Publishers have the opportunity over the next 1-2 years to adopt this technology before emergent AI players (like OpenAI). Publishers also have the opportunity to build new types of platforms, products, and experiences that will directly serve their customers long into the future.

In the realm of AI, all publishers are startups. The number one predictor of success for a very young startup: rate of iteration.

Thank you!

Hum is an AI company that is focused on customer data – offering platforms, tools, and services that help publishers understand and activate their audience and content.

www.hum.works

Hum CDP:
The first customer data platform to put LLMs at the core. Capture first party data about people, content, topics and organizations. Use that to drive targeted marketing and audience engagement. Deep predictive capabilities driven by LLMs.

Alchemist Tagging Engine:
Creates a customer-specific taxonomy by reading a full content corpus, then applies that to content automatically. Can support applying third party taxonomies (e.g. IAB, OpenAlex, MeSH), in-house taxonomies, or custom vocabularies as well. Available as part of the Hum CDP or directly as an API.

Embeddings as a Service:
Powered by Lodestone-v2, and fine-tuned on a customer's own content, this service turns content into an AI-native storage format (embeddings) to be used anywhere in a publisher's stack. This helps with AI apps, as well as improving search, retrieval, and recommendations. Essential for any conversational search applications.

Generative AI Workbench:
Cost-effective, trustworthy "tools" like summarization (expert, 9th grade), hypothesis generation, short pitches derived from running generative LLMs on individual pieces of content. Each tool is validated for effectiveness by Hum's data science team. Key focus is on synthetic media to improve customer data, enable better audience engagement, and create new customer experiences.

Drop us a note if you want to hear or see more: hello@hum.works