AMBOSS

Evaluating AI-powered Search Features

Valentin v. Seggern, Product Manager

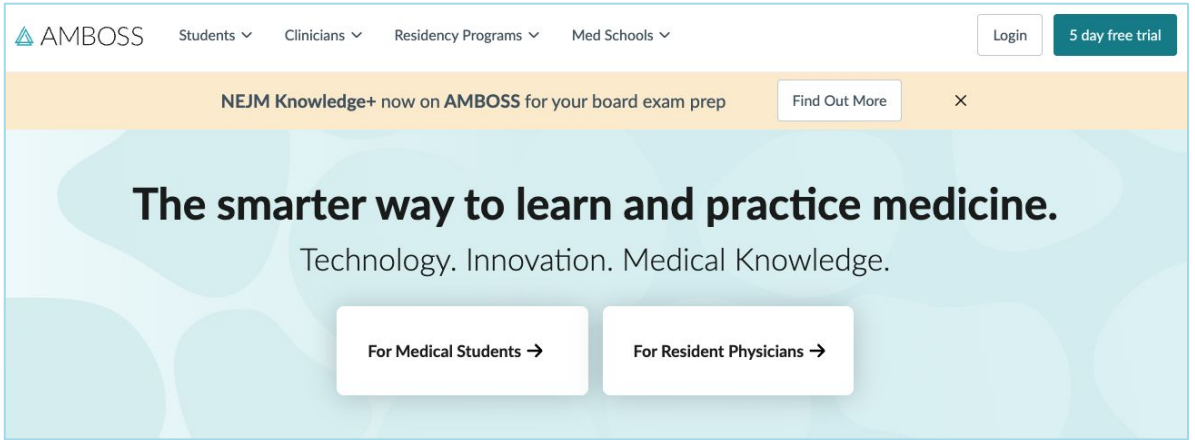
AMBOSS

The Company

SaaS for medical students & resident physicians to study and reference during patient care.

~500 FTE Offices in the US (NY), Germany (Berlin, Cologne) and Italy (Cagliari).

AMBOSS is used by >1.1 million medical students and practising physicians to prepare for exams and during patient care around the world.

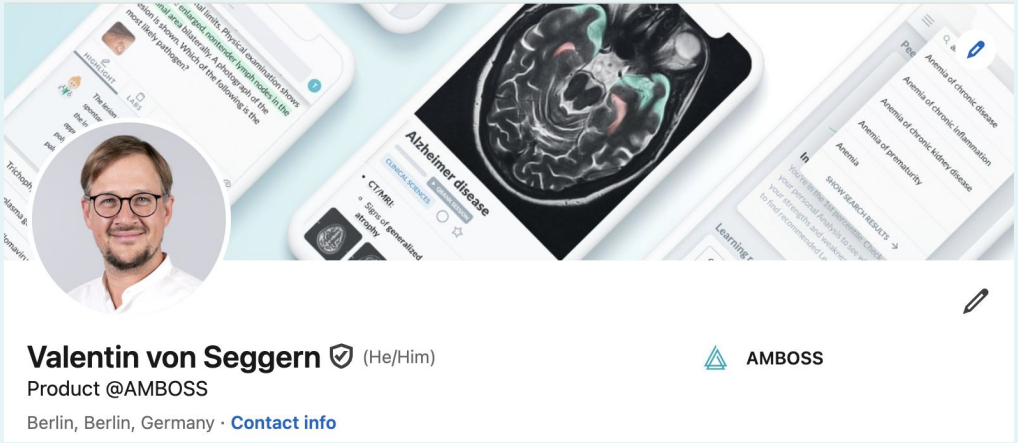


AMBOSS

The Speaker

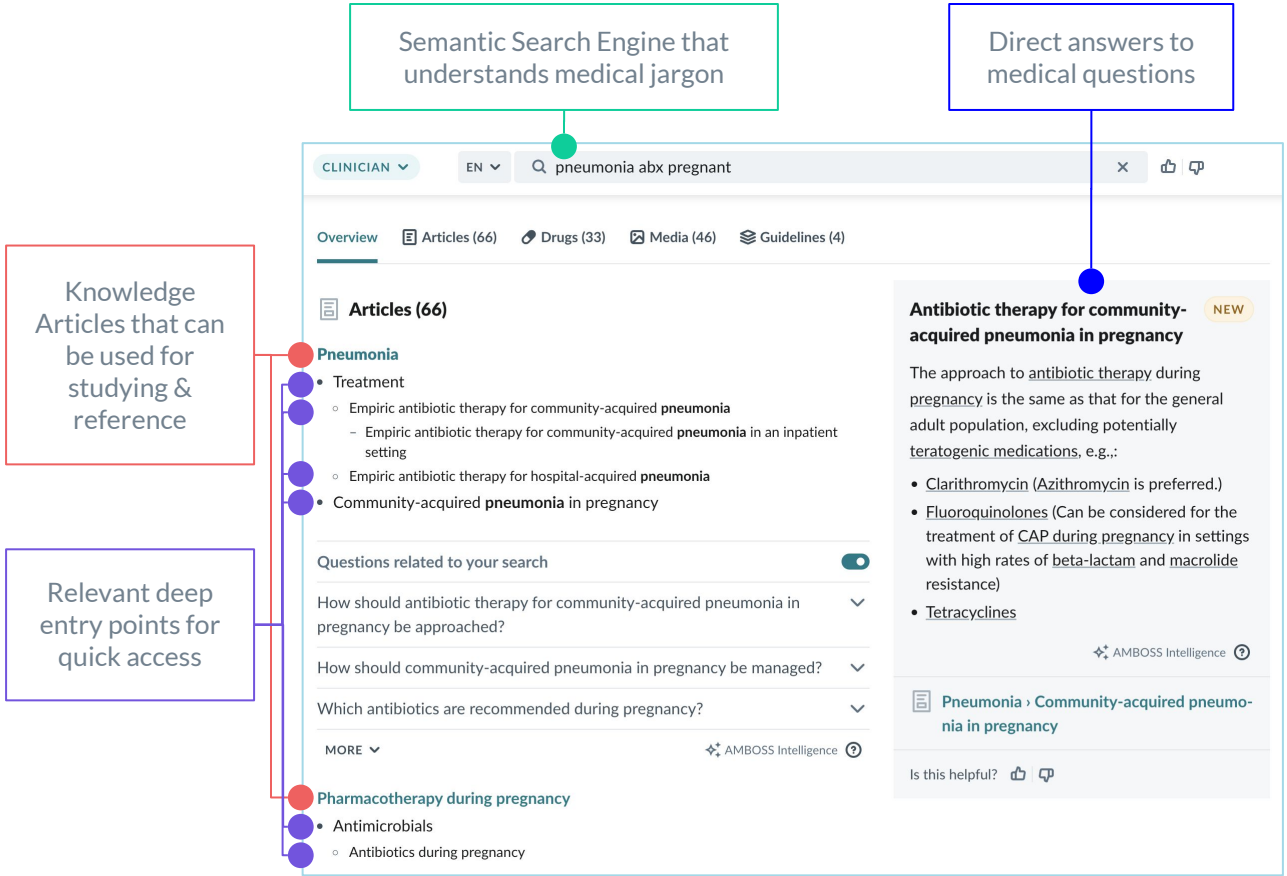
I started with AMBOSS 5 years ago as a Product Manager to work on Search. Before that I was a Software Engineer for about 20 years (Ada Health, T-Mobile) & Founder. Since 2016, I work in the medical field.

I came here to connect with other PMs & Engineers that work on Knowledge Search, **please reach out!**



AMBOSS

Search Engine



AMBOSS

Search Engine

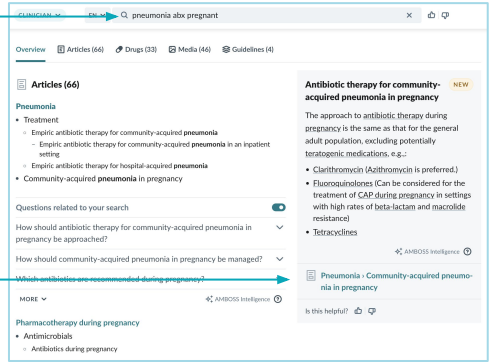
Key points about AMBOSS search engine:

- Expert users that exactly know what SHOULD be there
- Precision > recall
- Usually 1 to 5 relevant results per query
- Professional tool: Low latency, must be fast
- Frequent Query / Document language mismatch: Abbreviations, Synonyms, Clinical Shorthands, Spelling Mistakes

Today I will talk about:

- How did we start building out a dedicated search team
- Our journey through **hybrid embeddings & BM25 search**
- Our latest **Generative AI-native release**: Knowledge Panels on the SERPs

All from an ***evaluation*** point of view.



AMBOSS

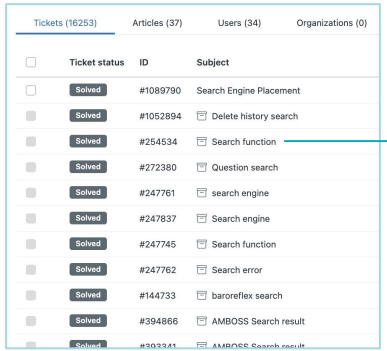
Conclusion

- Evaluating Knowledge Search Features is hard.
- No “high signal”-transactions.
- Many things to measure: Qualitative Methods, Online, Offline.
- All methods have different strengths and weaknesses.
- KPIs such as CTR/abandonment rate need to be contextualized.

AMBOSS

Listen to your users

My journey at AMBOSS started with many, many, many customer support tickets. I read thousands unaggregated user complaints & their queries and result lists.



Search term used:
Spinal stenosis or Lumbar stenosis

Information they were looking for:
I typed "spinal stenosis" which is an amboss page that I can find through google search but I cannot find it when I type it into the search bar logged in at https://next.amboss.com/us

Note from editorial:
Technical issue where user could not see any content when performing searches on AMBOSS

"...Improve the search function to operate with American and British spellings as well as related disorders..."

"...Better search function and dark mode on phone..."

Today with GPT, it's easier to get an aggregated view, but raw view is still important to build empathy

Vector Search

AMBOSS

Vector Search

Pre 2022

Thousands of hand configured synonyms, stemming, stopword removal, a HUGE query … highly tuned BM25 based engine using ElasticSearch.

*"...I need to know **exactly** how it is called on AMBOSS to find it..."*

In 2022

We started to explore embeddings for retrieval

In 2023

We released the first semantic search upgrade using SAPBert

In 2024

We invested into an extensive evaluation pipeline to evaluate embedding models for our domain and switched to OpenAI embeddings

AMBOSS

Relevance Judgments

We started off our relevance journey with:

- 3 * 150 queries (head, torso, tail)
- ~5000 relevance judgements by domain experts in Quepid 🐸

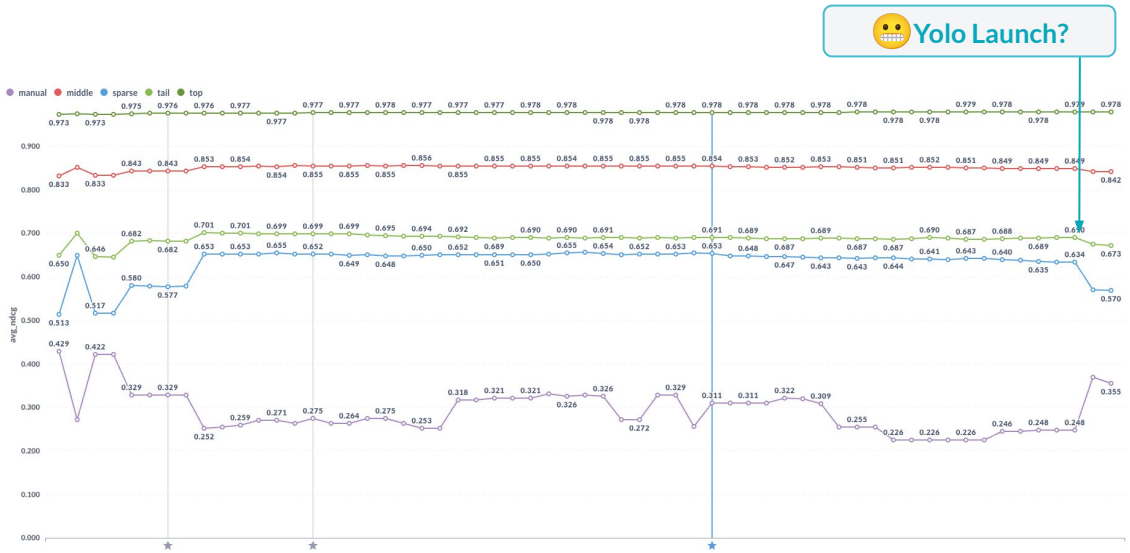
Initially the data helped to establish a baseline & systematic approach to evaluation, but the human ratings were noisy, biased and expensive to maintain.

We switched to implicit judgements based on user clicks (see AI-powered Search book, Chapter 11). Basic idea: Results users examine and don't click are not relevant, results that users click are relevant.

AMBOSS

Vector Search Offline Evaluation

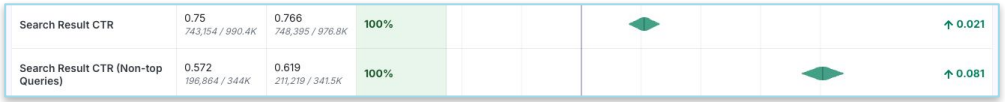
Problem: Presentation Bias - results that users don't see are not labeled as relevant. Lag in seeing an effect.



AMBOSS

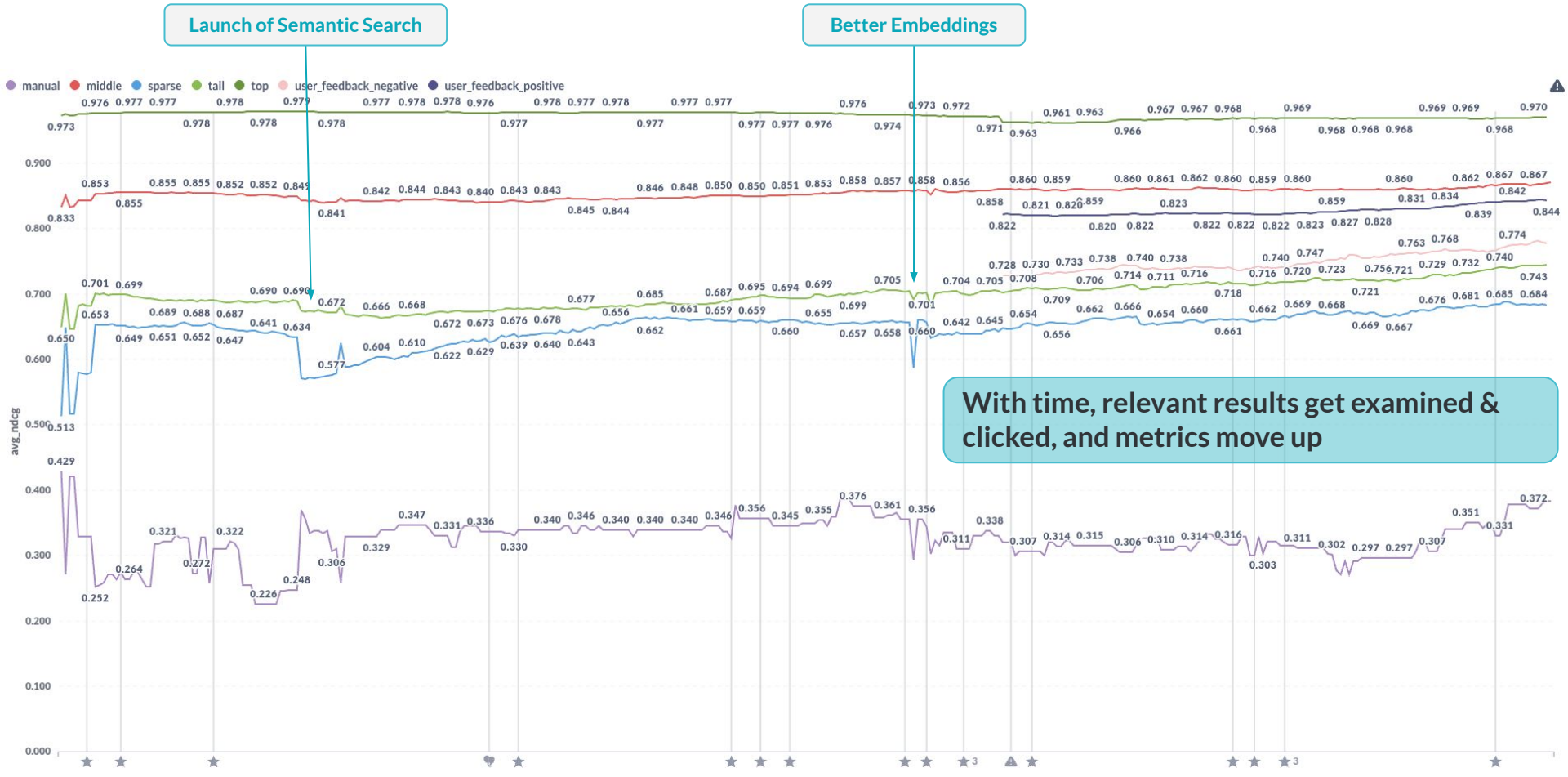
Vector Search Online

For the online experiment of adding semantic search, we launched an A/B test looking at Search Result Click-Through-Rate (CTR).



Search Result CTR	0.75 *743,154 / 990.4K*	0.766 *748,395 / 976.8K*	100%	↑ 0.021
Search Result CTR (Non-top Queries)	0.572 *196,864 / 344K*	0.619 *211,219 / 341.5K*	100%	↑ 0.081

We could show that the user group with **Semantic Search clicked more!**
We released the feature & hoped for the best 😅


Launch of Semantic Search
Better Embeddings
manual
middle
sparse
tail
top
user_feedback_negative
user_feedback_positive
avg_ndcg
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
With time, relevant results get examined & clicked, and metrics move up

AMBOSS

Embedding Model Offline Evaluation

To get started with Embeddings, we used SAPbert, a model optimized for BioMedical Entity recognition.

After our initial launch & online evaluation, we wanted to improve on our semantic search performance and tried to find the best possible embedding model.

How do you do that?

Rank (Borda)	Model	Zero-shot	Memory Usage (MB)	Number of Parameters	Embedding Dimensions	Max Tokens	Mean (Task)	Mean (TaskTy
1	gemini-embedding-exp-03-07	99%	Unknown	Unknown	3072	8192	**68.32**	**59.64**
2	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.21
3	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	56.00
4	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.23	55.17
5	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.93	54.00
6	GritLM-7B	99%	13813	7B	4096	4096	60.93	53.83
7	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.13	54.32
8	GritLM-8x7B	99%	89079	57B	4096	4096	60.50	53.39
9	e5-mistral-7b-instruct	99%	13563	7B	4096	32768	60.28	53.18
10	Cohere-embed-multilingual-v3.0	⚠ NA	Unknown	Unknown	1024	Unknown	61.10	53.31
11	gte-Qwen2-1.5B-instruct	⚠ NA	6776	1B	8960	32768	59.47	52.75

AMBOSS

What's a good embedding model?

We created a data set that represents our problem well. Queries in different forms, documents in different forms, spelling mistakes, languages mixed.

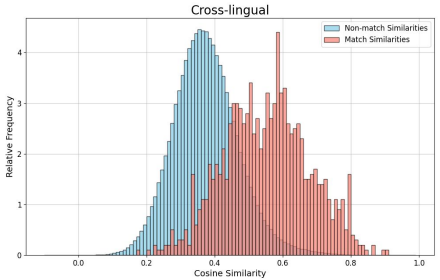
Similarity score should be high

Similarity score should be low

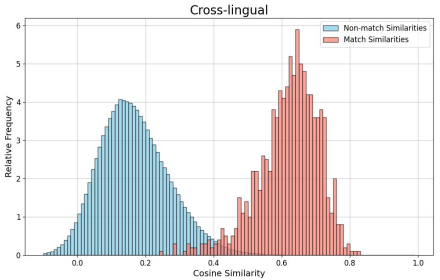

```
{
  "topic": "Pair Bond: Neurobiological Mechanisms",
  "query_form": "question",
  "query_lang": "English",
  "query": "how does the brain support pair bonding in animals",
  "document_form": "title",
  "document_lang": "German",
  "document": "Neurobiologische Grundlagen von Paarbindungen bei Tieren"
},
{
  "topic": "Single Umbilical Artery: Clinical implications",
  "query_form": "keywords",
  "query_lang": "English",
  "query": "single umbilical artery health impact",
  "document_form": "title",
  "document_lang": "German",
  "document": "Klinische Auswirkungen der einzelner Nabelarterie"
},
{
  "topic": "Injection Site Reaction: Long-term effects",
  "query_form": "question",
  "query_lang": "German",
  "query": "welche langffristigen auswiirkungne kann eine inje,tionssgtellenreaktion habne",
  "document_form": "passage",
  "document_lang": "English",
  "document": "Persistent injection site reactions can occasionally lead to chronic inflammation or granuloma
},
```

AMBOSS

Embedding model evaluation



SAPBert performance on Cross-Lingual Dataset



text-embedding-large-3 performance

Embedding Evaluation Leaderboard

Rank	Model	D Embed	Score	P@1	H_Dist	P@1 EN	P@1 DE	P@1 XLang
1	openai-text-embedding-3-large	3072	**0.857**	**0.914**	**0.938**	**0.953**	**0.917**	0.873
2	openai-text-embedding-3-large-768	768	0.826	0.894	0.924	0.946	0.899	0.837
3	openai-text-embedding-3-large-384	384	0.812	0.884	0.918	0.943	0.889	0.820
4	voyage-multilingual-2	1024	0.806	0.898	0.898	0.934	0.855	0.904
5	algolia-large-multilang-generic-v2410	1024	0.763	0.875	0.872	0.878	0.835	**0.912**
6	cohere-embed-multilingual-v3	1024	0.758	0.848	0.894	0.905	0.820	0.819
7	openai-text-embedding-3-small	1536	0.751	0.841	0.893	0.923	0.831	0.769
8	google-text-embedding-004	768	0.744	0.836	0.890	0.944	0.779	0.785
9	nvidia-nv-embed-v1	4096	0.739	0.838	0.882	0.911	0.830	0.774
10	openai-text-embedding-3-small-768	768	0.732	0.827	0.886	0.918	0.814	0.748
11	google-text-multilingual-embedding-002	768	0.728	0.826	0.882	0.852	0.824	0.801
12	openai-text-embedding-3-small-384	384	0.714	0.810	0.881	0.901	0.798	0.732
13	baai-bge-m3	1024	0.703	0.808	0.870	0.846	0.751	0.828
14	mistral-embed	1024	0.614	0.853	0.720	0.916	0.856	0.787
15	cambridgeltl-sapbert-all-lang	768	0.604	0.698	0.864	0.717	0.632	0.746
16	intfloat-multilingual-e5-base	768	0.581	0.723	0.804	0.834	0.691	0.644
17	snowflake-arctic-embed-m	768	0.433	0.639	0.679	0.850	0.576	0.490
18	nomic-embed-text-v1	768	0.372	0.551	0.676	0.779	0.523	0.350
19	cambridgeltl-sapbert	768	0.271	0.470	0.578	0.815	0.268	0.326
20	ncbi-medcpt-encoder	768	0.236	0.467	0.505	0.818	0.284	0.299

AMBOSS

Embedding model evaluation

Is it perfect? - No

Picking the best model has many different trade-offs besides good vectors for your problem:

Latency

Price

Reliability

AMBOSS

Summary Vector Search

- Qualitative Insights lead to starting development.
- Online Evaluation to see if we solve a problem.
- Offline Evaluation to tune parameters and improve how we solve the problem.

Knowledge Panels

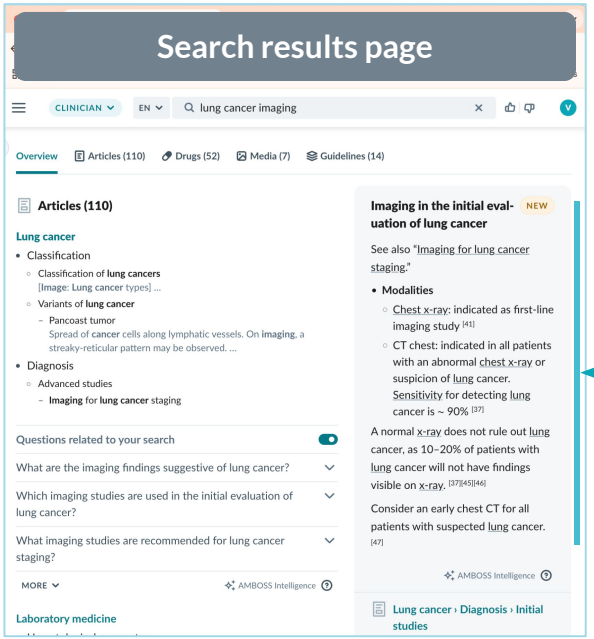
AMBOSS

Knowledge Panels

AMBOSS is used hundreds of thousand times per day by physicians under time pressure as a quick reference on the go.

"...I don't need a whole article, but just a quick fact & reminder when I'm on my way to the patient..."

How might we reduce Time to Knowledge further?



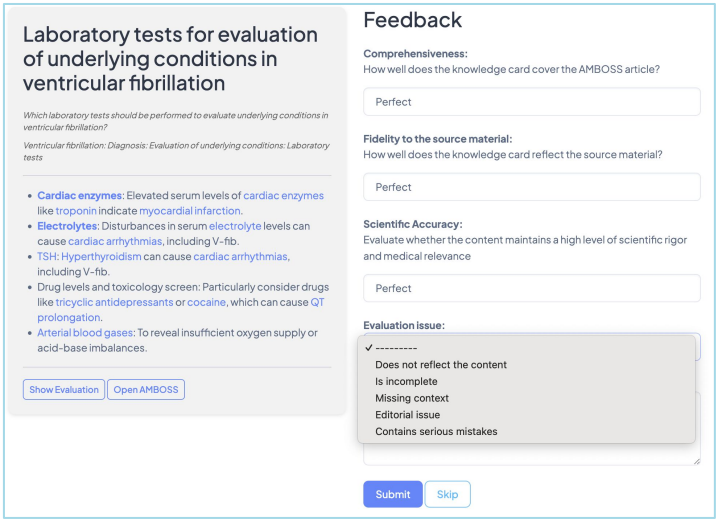
AMBOSS

Knowledge Panel Offline Evaluation

We vibe-coded a small labeling tool for our physician colleagues to rate Knowledge Panel along the axis:

- Comprehensiveness
- Fidelity
- Accuracy

We rolled it out and collected a few hundred expert judgements.

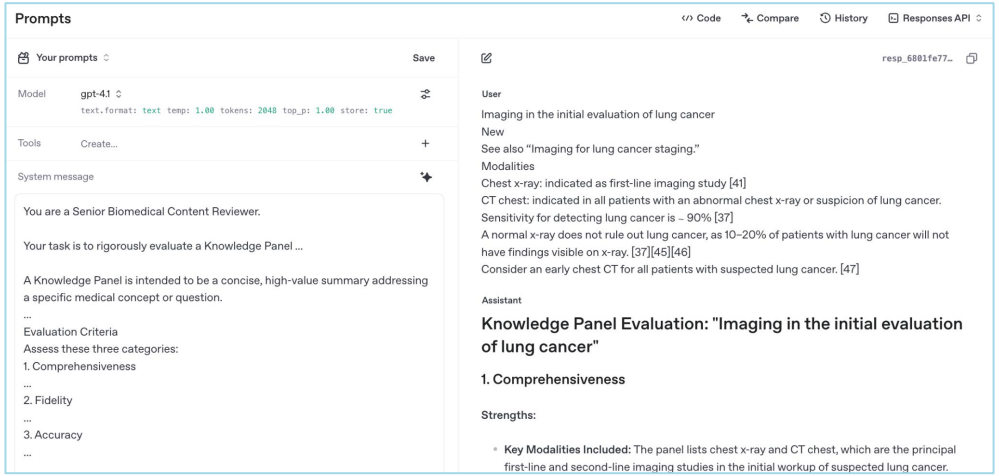


AMBOSS

Knowledge Panel Offline Evaluation

LLM as a judge. With the learnings from the manual labeling, we configure a GPT prompt and scale up judgments to thousands of judgements that we can run as batch jobs.

Target for evaluation prompt development is to optimize Inter-Rater-Reliability (IRR) measured with Cohens Kappa (κ) between physician and GPT.



Exemplary grading in OpenAI Playground, real prompt is longer, resulting discrete scores, prompt includes examples, criteria & definitions and is using the APIs

AMBOSS

Knowledge Panel Offline Evaluation

Usually the Rater prompt can be improved with error analysis, few-shot examples and other prompt engineering tricks, but avoid overfitting!

With a Generator prompt and a Rater prompt, it is tempting to generate synthetic datasets for fine tuning, … we tried a few times, but never had success.

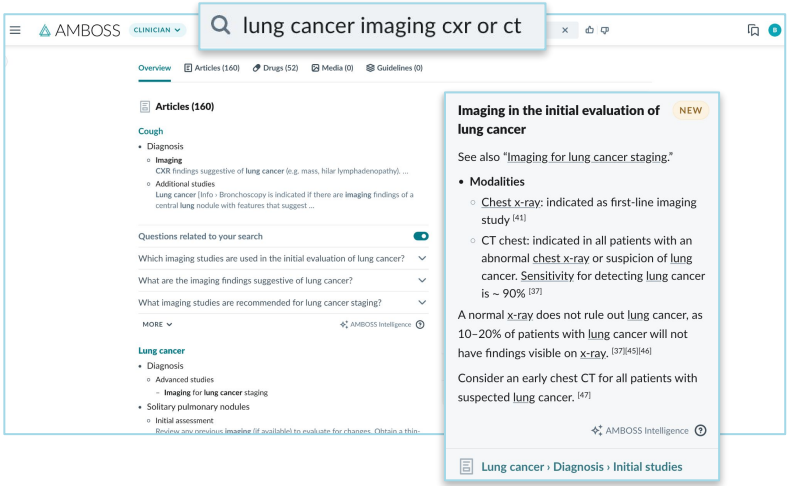
For continuous offline evaluation however, it works very well.

AMBOSS

Knowledge Panel Online Evaluation

Normally it's a good sign if people click results on the Search Result Pages and we try to ***REDUCE*** **abandonment rate.**

For Knowledge Panels, we assumed that easy questions can be answered on the SERPs, and tried to ***INCREASE*** **abandonment rate.**



AMBOSS

Summary Knowledge Panels

- Qualitative insights: identification of the opportunity & start of development
- Online Evaluation to see if we solve a problem. This time target KPIs where inverse of the KPIs for vector search.
- Offline Evaluation to tune parameters and improve how we solve the problem.

AMBOSS

Conclusions

Do the unscalable thing a lot before scaling it: We got a lot of value from spending time with raw user feedback, watching users use AMBOSS, looking at raw search queries and individual result lists with domain experts.

A great search is relevant & with great UX, so you need offline AND online metrics.

Combine offline & online evaluation to identify wins & iteratively improve the product.

GPT provided judgements & the ability to rate/judge extremely cheaply & with low latency is a new superpower for AI-powered search experiences, but it is still important to measure what matters to users.

AMBOSS

It’s a team thing

The AMBOSS Search Team is an interdisciplinary bunch of developers, data scientists, designers and medical doctors from 6 different countries.

Diversity of backgrounds & viewpoints matters!