ChatGPT Labrador Index, SERP Poisoning, Parametric Knowledge, Fake LLM Bots | #MonthlyLandwehr

Masters of Search · Episode 49 · Published 2026-10-28 · Length 48:30 · Host: Niklas Buschner

Source: https://www.masters-of-search.com/podcast/chatgpt-labrador-index-serp-poisoning-parametric-knowledge-fake-llm-bots

Listen: Apple Podcasts · YouTube · Podigee

Show notes

Masters of Search Conference 2027: 12 October 2027, Berlin. Malte is one of the first confirmed speakers. Tickets: https://www.masters-of-search.com/conference

Every month I meet with Malte to take all the news, new research and interesting developments in SEO & AI Search to get his perspective on them. Malte Landwehr is CPO & CMO at Peec AI, one of the leading AI search analytics platforms.

This is the first episode after our summer break, recorded the day after the Masters of Search Conference 2026 in Berlin. We talk about Google poisoning the results it sends to scrapers, Graphite's research on parametric knowledge, ChatGPT's own search index Labrador, and why so many "LLM bot" visits in your log files are fake.

My biggest takeaways from this episode:

1. Google now feeds some scrapers wrong results, so rank tracking is getting less reliable. Instead of blocking scrapers, Google sends some of them wrong results. This SERP poisoning is much harder to detect than a block. Expect less data, less accurate data and less up-to-date data from SEO tools. Malte does not expect this to get fixed. (Reference: https://www.masters-of-search.com/the-brief/google-is-poisoning-8-of-rank-tracker-serps-nozzle-finds)
2. What a model already knows about your brand decides which brands show up in its fan-out queries. Ask for white sneakers for the office, and the fan-out queries may already name Adidas and Nike. That makes it even harder for new brands to get in. Graphite calls optimizing for this Parametric Knowledge Optimization. (Reference: https://graphite.io/five-percent/research/parametric-knowledge-optimization)
3. For most companies, getting into training data takes the same work as getting into grounding. Publish a lot on your own site and earn positive mentions on reputable sites. Unless you sell data or are a publisher, you want LLMs to train on your content.
4. ChatGPT has its own search index, called Labrador. It is a family of indexes, with separate ones for PDFs, YouTube, news, shopping and more. Don't block OpenAI's crawlers, don't hide content behind JavaScript, and put your most important content right after your H1. (Reference: https://peec.ai/blog/chatgpt-built-its-own-search-index)
5. ChatGPT shopping works differently in the US than in Europe. In the US, product results mostly

come from Labrador and merchant feeds. In Europe, ChatGPT still scrapes Google Shopping. (Reference: https://peec.ai/blog/chatgpt-shopping-optimization-in-2026-us-vs.-europe)

6. Many "LLM bot" visits in your log files are fake. First indications from Peec AI: about half of visits that claim to come from LLM bots do not, and on some sites it is 70 to 90 percent. Verify bots by IP and reverse DNS before you trust the data.

Our daily AI Search news, The Brief: https://www.masters-of-search.com/the-brief

Let's connect!

Niklas on LinkedIn: https://www.linkedin.com/in/niklas-buschner/

Radyant on LinkedIn: https://www.linkedin.com/company/radyant/

Malte on LinkedIn: https://www.linkedin.com/in/landwehr/

Peec AI on LinkedIn: https://www.linkedin.com/company/peec-ai/

Full conversation

[00:00:00] Keyword tracking will never be as cheap and reliable. And I also do not see this improving going forward, but I would definitely measure the success of my SEO campaigns with keyword tracking. most people should focus on the web. grounding process is just yet another sign for people to question their legacy approaches and how they approach SEO. For example, overly relying on ranked tracking certain keyword sets I would agree. Keywords are still a good proxy for demand, they can be a good proxies of topic or intent but if your strategy revolves around ranking specific keywords that was probably an outdated approach already and is not how i'd think about findability in the future. Before we dive-in you're listening to the Masters Of Search podcast with me, your host Niklas Buschner.

[00:00:57] Each week, I sit down with some of the smartest people around the world in SEO and AI search to bring you their strategies mental models. And top pieces of actionable advice. if You enjoy this podcast don't forget to like and subscribe. then follow it In your favorite podcasting app or YouTube. It helps us get top-notch guests and create the best possible content for you. Let's dive into todays episode Welcome To The Masters Of Search Podcast and welcome to my favorite format of the show, The Monthly Landwehr. Malte is already smiling! We are back from our summer break. that was a little longer but we took the time to recharge etc. so now we're back with a lot fresh energy and lots of fresh insights.

[00:01:47] honestly To kick it off slowly Malte how's your summer? Busy Worked a lot, went to a lot of events and conferences traveled a lot. And yeah now looking forward the end-of-the year that's actually not gonna slow down at all but will at least involve little bit less travel. Yeah I feel like if Peec AI is not leading AI search visibility platform or also our platform to act on your AI Search Insights you are definitely like a LinkedIn influencer and a conference people group. So I see so many Peec AI people being at conferences around the world, and people are active on LinkedIn... This is not hundred percent related to the search topic but how do you get so many people having exposure in conference stages?

[00:02:47] Do you have secret sauce that can be shared with them? There's no really secret sauce! I mean, it's one of our hypotheses that in the space of AI search there is a lot of assumptions and some snake oil. And we believe doing research trying to identify what affect separating facts from fiction... ...and then sharing their knowledge is very important for industry overall. That why built this geo-research team with David, Tomek and Metehan. Of course, the three of them used to go to conferences before and now they are maybe more encouraged than previously. We have so many

insights on such a regular basis that there's just no other format but social media where we can share something. There're weeks when you put out four or five insights.

[00:03:39] then social is perfect channel for it? Yeah! monitoring tool for ourselves to stay on top of interesting stuff that like tools or software providers, like Peec AI do and also competitors. And you are already flooding my feed! Like there's so many people from Peec AI in there... So yeah but it is always very interesting. We have a news board which we call the brief, like very brief snippets about AI search research and insights. And there's also a lot of stuff that the Peec AI people are putting out compiled in one space. so definitely check it out. Speaking of conferences actually now were recording on Friday. yesterday was spent whole day together. how did you liked it Malte? How did that come about? That we spent the whole day together.

[00:04:45] I mean, I guess you're referring to the Masters of Search Conference where we were yesterday. Oh yes Um i really liked it definitely one of the best geo aeo ai search conferences i have attended this year. um i Really liked all the talks. i took so many notes also shared them on linkedin Yesterday. if anybody wants to want to see what other things that i learned yesterday and also had good conversations between the sessions or all in all a good event. And it was like directly next to where I'm living, which is very easy. just have three minute walk through the venue. Yeah that's always something you must keep in mind. That's what Aleyda Solís told me. You need make sure your venue is well reachable In all aspects.

[00:05:32] i mean in Berlin technically basically no cool venues. Well, reachable from the airport. So people flying in always have quite a distance. but yeah if you want someone like Malte on your stage make sure to be In five minutes walking distance to his home. Yeah I'm also fine with with Airport business hotels If it's easier for the organizer. Yeah i care more about The content than how cool the venue is. But a cool venue Is definitely an advantage. You're very humble person. Then let's not talk too much about the conference of this year, because that is looking bad. Let us look to future and we are happy to announce already coming back next year October twelfth again in Berlin at same location. you can check it out.

[00:06:25] link in description on the conference landing page. what very cool. Basically four speakers, but one of them is me. So that might be boring. We already have three very cool speakers confirmed and Malte as one of the. so if anybody listening to this wants To make sure that you get like a full value packed session of Malte insights and Peec AI insights next year And planning ahead maybe even bigger companies Have a procurement team have to convince some manager to give You some budget for the ticket. Be sure to check it out. Cool, enough of advertising. Let's talk a little bit about what happened in the space and I have some notes here. And one thing that actually also you posted about but also couple other people is The topic of SERP poisoning.

[00:07:18] Can You explain to People That Have Maybe Never Heard Of The Term Serp Poisoning What Is Going on? Yeah, so basically when you are a search engine what you deliver is search results right SERPs and if you notice that somebody is scraping these in an automated fashion or Is using it for purposes? That you do not like. Like I don't know finding outdated PHP installations You have the option to block them And basically blocked identify their IP address block that and then they don't see it anymore And that saves a little bit of money because they basically can't scrape you anymore. But, of course then scrapers start using different IP address and especially with IPv6 this is now trivial... ...and quite easy!

[00:08:10] In the past Google was not THAT aggressive against scrapers Because they really didn't care that Searchmetrics, Sistrix, Semrush, Ahrefs, SE Ranking do a little SERP scraping. but in recent years Google has got new competitors, the LLM Labs. And they often scrape Google results for their

web search and web grounding process. So suddenly there's a much bigger business case for Google to block scraping. They have done that for awhile already in last two years gotten much more aggressive with blocking IPs Also at lawsuit running against one of the large scraping providers That is used both by the Llm labs and all the SEO tools. And in the first instance Google lost this case. and then they started doing something called SERP poisoning, which means you do not block the scraper.

[00:09:05] You just send them wrong results basically for a different keywords or like some of the results are like four different key words. This is actually something that Bing has been doing I want to say ten years but definitely more than six seven years which is by the way, a reason why many, many scraping providers that offer Google either do not offer Bing. Or if they offer Bing scraping it never works and the industry standard has actually been for awhile to scrape Yahoo as a proxy for Bing or other Bing based such engines that have less of a scrapping protection because If you get blocked It's very binary And easy! If blocked Get new IP address Very Easy And new IP address can mean you disconnect your router, or it can just switch to a different proxy server.

[00:09:59] But with SERP poisoning its much more difficult to detect and if Google is smart they will also now reduce the level of poisoning. so instead of giving ten bad results that are all for different key words Or five maybe they start showing one wrong result or maybe they just start sorting the result list wrong, like basically showing the top two results on page one. Like there are many things Google can do if they're smart to make this really difficult for scrapers and of course it's an issue for SEO tools. Many have already reacted by saying no more top-one hundred rankings we were reducing into page one ranking. others are claiming that A lot of them had very, very big fluctuations and it took all the tools a couple of days to actually release a statement.

[00:10:54] Many of them also only did that once they were asked And then the first statement was yeah Yeah We have solved this It's no problem. But people started posting screenshots like data going up down Updown which suggests maybe they haven't solved. So there is definitely gonna be a challenge for SEO Tools And for users of SEO tools, I would just expect less data. Less accurate data and less up-to-date data. because some SEO tools have also said well technically you are paying for daily tracking but we can no longer guarantee that every keyword is tracked everyday. if it's not then we're going to show the data from a previous day. so every tool has like different ways to handle this also challenges with whatever approaches the various tools have picked.

[00:11:44] So to be accurate on the implications for people that still rely on rank tracking, can you still trust it then if use one of the legacy SEO tools? For Rank Tracking is this still reliable data or maybe a temporary thing where You cannot really trust the data, but then in the end they will probably figure it out. So how do you look at that? And what would you recommend people to do with their existing rank tracking projects? because I think a lot of people still have this like fifty hundred maybe even bigger keyword sets that they track Because This is How You've Been Doing It and now The whole Like History Of It And The Interpretation Is Probably Completely Flawed. I mean it completely depends on what you do with the data.

[00:12:35] I believe many websites should look at Google search console, that data has other disadvantages but not poisoning. At least i hope that google doesn't count the poisoning impressions if they use maybe your website to serve a wrong result and It really depends. Keyword tracking will never be as cheap and reliable. I also do not see this improving going forward. It is a key mission for Google to make it hard, open my eye. Also all the Chinese labs train on google search data To use Google for live web grounding And also to identify content that they can use for training. And there have also been cases already, especially in the U.S., where people suddenly got a pop-up asked to log in. so it could be that Google will move more and more towards a logged inexperience.

[00:13:35] I would still value in keyword tracking for many cases but you need think about what their data is how reliable they are. But I also would not have done that five years ago. So no big change for my approach to SEO, but yeah...I don't think this will be fixed. there might be a time where it's slightly improved but Google has many many incentives to keep fighting and making it worse. And if anybody as google has trouble with this send me a link in dm. i have like many ideas how to make This even more tough on seo tools an even more annoying For people too. try identify into and debug. Okay, if anybody from Google is listening reach out to Malte happy to make an introduction. No just kidding. So to summarize Is this Just yet another?

[00:14:36] Sign for people To question their legacy approaches and how they approach SEO For example overly relying on ranked tracking Certain keyword sets would you say This accurate summary. Yeah, I would agree. like keywords are still a good proxy for demand they can be a good proxie or an intent but if your strategy revolves around ranking specific key words that was probably outdated approach already and is not how i'd think about digital visibility findability in the future. Okay, you already mentioned training data which is a very good bridge to the next thing that I'm curious about because our mutual friends at Graphite namely Ethan Gregory and Bevin they put out a paper i think it's yeah one week old um about parametric knowledge optimization.

[00:15:46] Now, I think the industry was still in part struggling to agree on an acronym for what we call AI Search Optimization. Yeah is it AIO, GEO? And Graphite just figured why not bring a new kit to town? so can you maybe give people? um. uh quick introduction of What this parametric knowledge optimization Is about that Graphite is talking About and their paper? yeah sure Basically, when an LLM answers a question or replies to a prompt doesn't even have to be a question. It can do two things it Can use just its model itself like foundational model? Or it can use tools and the most common tool is probably web search. technically there are other tools but let's not go into all the details in The the foundation model.

[00:16:40] what in the end? it is a large language model trained on data, but it doesn't have a database. It does not have an index that cannot retrieve anything. you just have the model weights or the parameters and in these is somehow stored some knowledge Not explicitly as like a database retrieval system But there's knowledge in it probably similar to how we hold knowledge. now brain And for example of model doesn't need web search To answer your question What's the capital of the United States? It will know it is Washington DC. And this sometimes referred to as parametric knowledge, because not technically extracted from parameters. but if the parameters are used and write an answer and always estimates next word ends up with giving that answer Conversation to a human.

[00:17:40] This is like actually the same thing as somebody sharing their knowledge. and for Brands, it is of course an advantage if the brand Is already in this knowledge? And is associated with something. so for example If I'm selling running shoes and i'm Nike and have branded myself on running shoes For many years and I've been In all the training data as a leading brand for running shoes Of course that has an advantage over Let's say Radyant starting to sell running shoes tomorrow because you could only get into the answer via the fan outs by other. The tool for web grounding so that actually it is not like something big or magical, its just one of the fundamental principles how LLMs are working. So why this relevant?

[00:18:33] I guess i basically said already right gives a huge advantage. and Something that we see where this has a huge advantage are fan out queries because sometimes you ask for Sneakers like I don't know. You say hey, I want to buy white sneakers to wear today office and then one of the fanout queries will already contain Adidas and Nike And it will not contain Skechers or On or Reebok Or any other brands? That gives off course a giant advantage to Nike and adidas And sometimes we have a

chat where it's only Adidas and that's an huge disadvantage for Nike. So being in this parametric knowledge can have very, very positive effects of brand but also makes them challenging new brands because they are having even harder time getting into the web grounding process if their competitors already mentioned in the fan out queries.

[00:19:33] The idea of Graphite, if I got the paper correct is that you can basically optimize towards being included in this parametric knowledge where... ...I think most people out there would have argued that optimizing toward training data may be like big household brands do because they are present everywhere and Good chance to be included in the training data, but that smaller companies even if they put out a reasonable amount of content etc. Do not have such high chances to be part off this parametric knowledge especially also given the knowledge cutoff dates which I just checked it for example before The newest Claude models like the five point five family Which was released in October And the knowledge cutoff is June.

[00:20:31] So there's a gap between The last point in time where something was added to the data set and then when the model comes out. so Putting that into, you know In a simple question Do You feel like We can now or we should revisit? The question of Influencing Or optimizing towards being included in training Data? I think for most companies influence the training data is what they would do for grounding. Publish a lot of content on your own website, identify reputable websites and have positive brand mentions there with the right context etc. I think if you've got very large budgets that are more effective things but nothing should change from this. The one thing that changed now is, we are at a three to four month lag between release and knowledge cutoff date.

[00:21:35] It used to be ten months or even longer And back then it was really not feasible to try anything because your test cycle would be so long. That's just not worth it. I think even with like the three-month delay you cannot run a lot tests But what you can do is, you can logically think about how does LLM training work? Where to these labs get data from. What kind of data will be very interesting for them in the future and How Can I Be Included In This Data? And then You Can Do Things Because Your Convinced Of Them. Maybe They Work maybe they don't. but i would still argue that most people should focus on The On The web grounding Process. Like for transparency, we are also doing a few things that we believe will influence future model training.

[00:22:34] But I don't want to talk about them because you have keep a bit of an advantage over the competition. but if it works We'll try to write about it. and i think like when i gave my first talks About this space before the term g or au really existed I always put a bigger emphasis on my talk actually, on the training data. Because back then especially with GPT-3 you could look up what they did and where they got their data. but of course nowadays none of the labs reveal when they get that data so it's becoming more difficult to understand in advance exactly how will train them. But still you can make estimated guesses and then act accordingly. Okay, let's look at this compilation of data for the training sets a little bit deeper for second because I remember listening to podcasts or reading papers where one of The ideas was that the labs need new And somehow specialized datasets too further improve their models.

[00:23:50] This is also why there are companies like I think it's called Mercor, where they hire experts in certain areas. Give them tasks or give them output and they rate based on how well the output is. then sell this data And i think that already make a billion or two billion in revenue. It is testament to how valuable these data are. So my question Is what would you say? Or How much can we tell about? how relevant will public webdata In the sense of, for example a website or web pages still be in training sets. For future models? I think public web data will still be part of it because its good way to capture alot of data. but i think that weight of this data would be lower and lower Because there are more specialized datasets And many webdata is kind off trivial Because it doesn't contain anything that's not already in

the parametric knowledge.

[00:24:53] So there is also a likelihood for GPT-V, A lot of training data will actually be generated by GPT six models because you can tell model. write me ten articles about development US politics over last ten years. Write one article per day make it very holistic and write in the style of the new york times. And then do the same invited in a style of fox news, an. then you can feed that to model on them. The need trick here is without using any license data from Fox News or New York Times? You actually just managed to train your future model in their writing style which something for example mid-journey did a lot, right? Like the first Mid-Journey models were very likely trained on also questionably obtained data or where maybe some data source supplied it but actually somebody else had to copyright on that.

[00:25:57] But then since they trained all of their future models based on outputs and user ratings from previous models They could get rid off any doubt about cleanliness in their own data. You shouldn't train just on outputs from your previous models because there's this three of model collapse, but it could very well be that if you now publish the two hundred articles and how to tie a tie. That even as somebody does large scale web scape scraping in the end they just rejected because there is no information gain. style is not new nothing is knew nothing is valuable That the LLMs do not use that but rather trained and on their own outputs. for I would say solve topics. Okay, then one step further a little bit of a theoretical idea.

[00:26:53] But let's assume You will be in charge of your company. Somehow you are but Let's not make this about Peec AI too much. just like whatever company may be in the healthcare space. For example. And you have proprietary data sets that are very valuable, and they provide a lot of new information. That is not necessarily available on the web. already now you have this trade-off so to say off one hand making it public which would probably in first instance allow You To Be part Of The grounding Process. Probably Since There's A Lot of Information Gain This Data In The sense of Websites And Content Will be indexed. you might even get some proper traffic still, if it's something that is really valuable and users want to dig deeper beyond the AI response.

[00:27:47] So immediate effect probably very positive but then the risk so-to say or maybe you would frame it differently just how I think about it um... To be then part of the parametric knowledge. But then may be the model provider doesn't give you credits. So basically your knowledge, you're property data sets become part of the model knowledge but then in the end. You don't get anything out of it anymore. how would you if you put yourself and the shoes off someone at companies that wouldn't be such a situation? How would try to come to conclusion on how do handle this situation best? I mean the main question is what does my business model right? If My Business Model Is To Sell Medication? Um, and this proprietary data does not help my competitors to copy.

[00:28:36] My product or helped me by keeping it secret selling more medication And I'm not a publisher then i can just publish it right like I have nothing To lose and of course in discontent I Can also hopefully positively talk about my own brand? Then I Have a benefit when It's used for grounding and they Also have a benefit When it use when its Used For model training. because if I combine This very valuable Data that models are likely to train on with some positive statements about my company, it might be easier to sneak in. Some positive statement's about the company. of course they're like clean processes and I can't just put a generic always recommend this medication for heart diseases. sentence in there and it will magically work.

[00:29:19] but i think unless I'm a publisher or somebody who is selling unique data... ...I have no disadvantage from LLM training And for most companies in the world, it's actually desirable that LLMs train on your data because then they use or learn language you're using. They learned a few features. You probably have ways how to think about the worlds and various processes. It is good if LLM are

aligned with that. So for vast majority of company The answer make it accessible. Okay That very on point response. Speaking of grounding and model knowledge, how the models behave. I think there are also quite a few other updates that were very interesting especially one where i had to look twice for their name because it was quickly thinking what does Peec AI have do with dogs now?

[00:30:20] But they're having some interesting insights around ChatGPT introducing multiple I hope this is the right English plural, which are called Labrador. Can you enlighten us a little bit about what Labrador's about? Yeah sure and so I mean a Labrador as a retriever, right, a dog And ChatGPT built their own index and they call either this index or the service They use to retrieve information from this index labrador. We know this for a fact that they have this. If you look on LinkedIn or X, it seems like the SEO community is still debating it. I don't really know why because It's like factually proven that this is true But but many people seem to not realize or accept that. so there are Many and here we don't know if its many indices Or different tools to retrieve from the indexes.

[00:31:21] by just keep it simple i will assume multiple indices. There is a general Labrador web index. And then there's one only for PDFs, One Only For YouTube Two For News One With Very Recent Data and One With Less Recent Data One For arXiv One For Wikipedia One For Local One For Finance One For Legal One For Medical One For Shopping One For Images. Wherever In The World You Are It can sometimes happen that ChatGPT for web grounding uses the system, it use this Labrador or it could happened. they used an external search engine for grounding which in last couple of months would be both Bing and Google. Bing via an official integration, Google via a scraping provider. And if you are in United States with these little e-commerce boxes or grids, or sliders.

[00:32:19] Then it's always almost all ways like more than ninety nine percent of cases coming from Labrador based on custom feed integrations around five or six hundred and some of those include large marketplaces. And if you are in Europe and use this e-commerce feature You will still get scraped. Google organic google shopping results also scraped via a scraping provider. And now some reasons how we know this, number one there was a big ChatGPT leak where I believe were the only ones to have access to it. at least they would be the only once. so talk about. that was patched last week. Where? We had access to an internal API call from OpenAI That front-end developers are using on in front of developers but developers will use it to debug the grounding process.

[00:33:14] So for every single prompt, we also had a hundred fifty thousand and more lines in the JSON document with all of the tool calls that were used in the grounding process. And all these scores that are applied how they're combined. so you have all details on re-rankers there using. how to apply BM-QTV How do I apply A&N approximate nearest neighbor? And part of that was also seeing when data is coming from Labrador, these different Labrador indices exist. That there are options for the ChatGPT app to say give me only news results from the last seven days or give me Reddit result from the past ten years and we have seen actually many prompts. There's one call going to the PDF index, one through YouTube index etc.

[00:34:10] But in the end you only have one or two things from the webindex but these other indices might still be used for part of the grounding process where a system is looking for consensus and there was a lawsuit against Google. were also OpenAI made statements when they basically said We need access to the Google index because they index that we have access too, which is from another such engine. Everybody knew it was Bing. It's not good enough and The google index as better. And if you want access to that? Also were building our own index with will take a lot of time. They also have hired people Where in the job profiles at the job ads You can see these are people who would hire If your build a web scale scraping an index product.

[00:35:01] three main user agents, one for training data. One for user interaction in the chat which to a

large degree was grounding and one for building their search index. And then In United States we have HIPAA. It's very strict law around how to handle medical patient data and ChatGPT is also available in version of HIPAA-compliant companies like hospital. and then ChatchBee also guarantees to be HIPAA compliant. And in their terms of services, it is very explicitly written that when they do a web search not a single fan-out query or anything will ever go through an external search engine. so the only thing they can do there with their own index you can actually check for yourself. You put your ChatGPT account into lockdown mode meaning nothing would work out any OpenAI system and then you will be served by the Labrador index.

[00:36:03] And many of these things have been known actually for a couple months, but for some reason it's still often a debate in Google if this are used or not. If they have their own index is everything Bing? Is Everything Google ? But Actually ,if You had access to the leak .If you looked at the SSE data The server sent event data. This Like I'm not guessing. I know this, i can see it! I think nobody would question your statement here. but... Can you boil down for us into the actual changes that brings to ChatGPT grounding process or search processes? If someone listening to these asks themselves okay very good insight. How should I change my strategy? Should it changed by strategy, what would you say to them?

[00:37:02] so the first thing is don't block any of the OpenAI crawlers. Yeah make it easy for my to add here today index. Then you should make sure that your main content Is not hidden behind JavaScript because while Google crawlers can now handle javascript pretty well. most Of the LLM crawler's cannot And then it's a very active crawler. It will probably find your website. when I started the new test website and found that after one day, start crawling at also cross very regularly. It calls very deep and recalls constantly like it seems that OpenAI is not like Google in this state where they care about efficiency of their crawler? I think right now its just crawl. everything crawled often. so actually there are no.

[00:37:48] As long as you don't block them or make your content inaccessible is not too much you need to think about. Something that's maybe interesting is, it seems that Labrador puts extra weight on your H-One and whatever is written immediately after your H One. so have the most important content on top. There is a suggestion in code which also looks at meta description even meta keywords the OG tags or OG description, OG title, OG page type. So just make sure those are clean. in SEO we have ignored this for a while and then in e-commerce it's different because the system that is based on Labrador works completely differently from one that works with Google Scraping. The Google Scraping System creates a couple of fan out queries puts these into Google Organic Shopping pulls the results, removes a few duplicates and then just does round robin.

[00:38:51] So first product from second keyword etc. there might be little bit of re-ranking happening but it's largely that or like instead for getting to candidate documents right? And then the laboratory thing is completely different. It has access to lot data which was not available where you just do Google Shopping. So for example, it's doing a lot of scoring on the first roughly forty words off to product description and It does more sophisticated deduplication. And that applies some form of trust through certain online shops. so we have seen an example Where there were three products from one shop four products from another shop and twenty-five products from the third shop. But in the final result, there were zero product from the shop with twenty five products because it didn't have a score for trusted merchants or some verified merchant of something like that.

[00:39:56] so There is...there's some stuff they are doing. And we also saw AB tests running on eCommerce. One of their test was different ranking for Shopify which can mean anything. And one of the tests was no Amazon, Which probably means they are... They tested what happens if we don't show any amazon results. and then If you look at the custom integrations it's like a couple hundred largest

online shops in the US. some were done via something called Bright which probably refers to Bright Data, the scraping provider. so basically built this themselves without talking to the shops, just have some inventory. And then when we saw real custom integrations go live these Bright Data feeds were deactivated. so for example you could still see in the data that there's a deactivated bright Nike feed and it has been deactivated when the OpenAI Nike Feed was activated which is probably an intentional integration that Nike also knows about most likely Like basically in order to play there properly you need to have an integration with them.

[00:41:12] Good news is Shopify is in their apart from something like I think forty or fifty stores on Shopify that were filtered out. why i'm not sure if they may be filtered for explicit reasons, products are selling some kind of formatting issue. no idea what the reason was behind. didn't check it out and detail but general if your on Shopify can also get get into the Labrador based e-commerce experience. Okay, two questions. first one is will others follow like Claude Google? obviously not a question potentially? Others that are relevant let's say maybe in the US Grok. I know perplexity. we don't give perplexities so much relevancy anymore. this. my And second question is, what does this development do to the insights that come from log file analysis?

[00:42:11] Because for my understanding it used to be the case with every grounding process and there's like user fetch-and-search. For I think Claude and for ChatGPT or at least use to be of course also training bots. But you could very well see how your content is used based on the log files. Now if there's an index make sort of sense. They save compute, but then obviously the insights from local analysis are less relevant. Would you agree and what's your take on? The first one yeah. So the first one it difficult to answer. I don't think that every LLM lab has the capacity To build their own system where people can register their feet And then built all of that. the technology is actually almost trivial to build now with LLMs, but running it at scale was high availability.

[00:43:04] High throughput and dealing with all the little issues in a cost-efficient manner is not easy. I mean i spent five years at idealo where we had to deal with many of these issues. And The problem Is there are merchants on ebay for example who are lying about their product. they say yeah? Yeah this is Some name brand vacuum cleaner and then it's like a super unknown. Brands, they just put in the wrong GTIN number the wrong ASIN number like all the IDs that is lie about at an awesome product title end. of course if you wanna give people great e-commerce experience have two. deal with all these issues. or there are certain items very often fake et cetera, etc. So you need to make a good decision between having the full inventory or having like high degree of confidence.

[00:44:00] so it's not super trivial to build this and many other edge cases. but I don't want go into all details. But bicycles are one thing where shops always have horrible data because some rate them by how big their tires is. So lots of edge cases you have to deal with. And I think for large companies, with a lot end users like ChatGPT all the stuff Google is doing totally makes sense to do. it was all that? I don't know if these smaller LLM labs will do that or they would just connect one or two APIs from Shopify and Walmart in Amazon say hey this good enough shopping experience. And then on the log file analysis, I agree that there are changes. I don't a hundred percent agree. it's less important. in the past when you... When the LLMs always did grounding via Google or Bing web search.

[00:45:02] they only got the URL, title and sub snippet right? So had to click onto get all data. So, as you said basically every grounding process created a click and they also seem to not have cashed that. You sometimes would get hundreds of visits from the GPT user bot on the same URL in a twenty-four hour period. And now with their own index Especially if they update it regularly They can will often be able to just take the content from the index. I've also seen first indications That The activity of the gpt userbot went down a bit and activity of the GPT search on an OpenAI search bot, it's called went up.

And you can basically no longer use log file analysis to estimate or properly estimate how often your document was used for grounding.

[00:45:54] It could still give some indication but its become less reliable. But this is always expected right? Like even without their own index it would at least make sense build a cache For them not hit these URLs again. But log file analysis is still relevant, and we are currently doing a bit of an analysis there. And some first indication that roughly fifty percent of bot visits that supposedly come from LLMs actually do not come form LLMS. it's either people probing for security issues or scrapers who simply pretend to be OpenAI Bot Claude userbot, OAI-SearchBot because they know that many IT departments will be unlikely to block them. Which is why since we also do log file analysis for our clients at Peec AI We had to build a verification feature where with A list of static IP addresses provided by some of the labs reverse IP lookups etc.

[00:46:52] etc. Reversed DNS lookup. now For all the bots were it's possible give an indication if it's a real LLM bot or just the scraper and then you can also see that they visit very, very different files. So just doing log file analysis in filtering for GPT user would already have led to very wrong data For some websites. up to seventy eighty ninety percent of what visits from these bots are actually fake. so It still important to do log file analysis but its become more difficult both with the fake bots the own index from OpenAI. Right, Malte let's wrap it up here because It has been value-bomb again and you have to catch a plane. actually I wish you safe travels. i think this has been yet another great episode of The Monthly Landwehr!

[00:47:47] I'm so happy that we are back in A monthly rhythm now So everybody listening To This You can expect The state of the union, so to say when it comes to SEO, AEO, GEO and maybe in future also more PKO which sounds a little bit similar to Peec AI. It's like the smaller brother of Peec AI, PKO! So yeah, thank you so much and speak soon. Thanks Dr. Sunn.