A Weighted Sum of Everything Ever Written About You

What Bayesian linear regression, copied by hand from a textbook I've owned since January 2007, tells us about getting cited by AI

Ben Rees[1]

Abstract

This is a mechanism piece. My existing whitepaper, "Why GEO Matters More Than SEO," reports an experiment: what actually moved AI answers across ChatGPT, Gemini, Perplexity, Copilot and Google over two months. My post "Brand visibility is now a Bayesian problem" makes the case as metaphor. This paper is neither. It is the arithmetic underneath both, worked through properly, with the maths on the page instead of just hints.

A weighted sum of everything ever written about you

When an AI system answers a question about your market, it is not looking you up. There is no row in a database with your company name on it. The answer it gives is a weighted average of everything it absorbed in training, and the weights are similarities: how closely the current question resembles each thing the model has already seen.

That is not a metaphor for what happens. It is the arithmetic. The prediction a Bayesian linear model makes for a new input is $y(x) = \sum_n k(x, x_n)\, t_n$, a sum over every training example t_n, each one weighted by a kernel k that measures how similar the new input x is to that example. Christopher Bishop set this out as the "equivalent kernel" in **Pattern Recognition and Machine Learning** in 2006. The prediction for a new point is the training targets, weighted by resemblance. Nothing more.

The reason this matters now is that the mechanism running inside every large language model is the same operation. Self-attention, the core of the Transformer, is Nadaraya-Watson kernel regression: standard softmax attention corresponds to a Gaussian kernel, and the attention output is a similarity-weighted sum of value vectors, exactly the equivalent-kernel form. This is taught as a direct equivalence, not an analogy, in standard references such as Dive into Deep Learning, chapter 10.2, and it is under active formalisation in current work recasting attention as kernel regression.

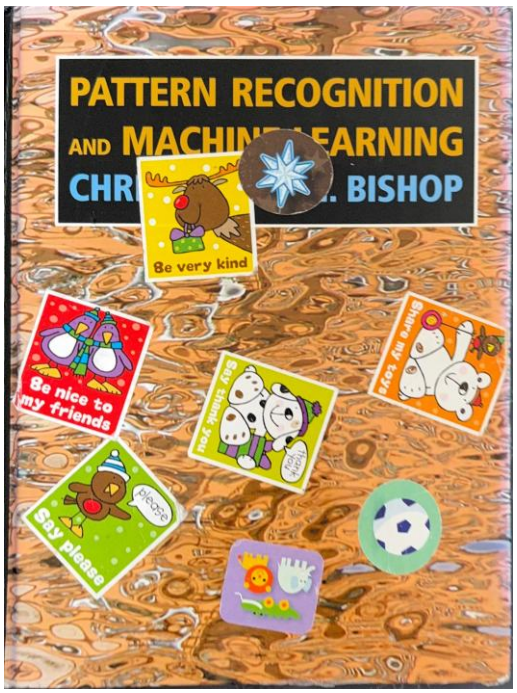
So when a marketer asks how to get mentioned by AI, *the honest answer is mathematical*. You are one of the t_n. Whether the model reaches for you depends on the kernel: whether the question resembles the things written about you strongly enough, and distinctly enough, for your weight in the sum to be anything other than

[1] www.bjrees.com, ben@bjrees.com

negligible. Most brands are not absent from the sum. They are present with a weight indistinguishable from zero, averaged into the generic centre of their category.

I drew this before there was anything to draw it about

Almost a year ago a former boss of mine, a CEO I worked for, told me something that reorganised what I do. Most marketers, he said, cannot do the mathematics, and most mathematicians cannot do the marketing, and the interesting work sits with the few people who can do both. He thought I was one of them and that I should stop hiding it. I have a voice note of myself repeating this back on a run in April, which tells you how much it stuck.



Here is the part that turns a nice piece of career advice into evidence. I have a diagram of Bayesian linear regression in my own handwriting, in a copy of Christopher Bishop's textbook, and I can date my engagement with that book to the month. On 11 January 2007, barely a month after the book was published, I subscribed to its own mailing list on Microsoft Research's servers. I have the confirmation email and my own reply from that day. That is not a recollection, it is a timestamped record generated by a mailing-list server I do not control. My copy of the book has carried my children's stickers on its spine since around the same time, when they were young enough to decorate it rather than read it.

The point of both details is the same: this was already on my shelf and already being worked through, years before there was any marketing application to attach it to.

The diagram sets out the prior, the likelihood, the posterior over the weights as a full multivariate Gaussian, the predictive distribution, and the prediction rewritten in its equivalent-kernel form. It is, line for line, the mathematics in the section above. The notation is Bishop's chapter 3.3 exactly: the posterior written as N(w | m_N, S_N), the predictive mean rewritten as a sum of training targets weighted by k(x, x_n). Anyone who owns the book can check the notation against the page.

The notes around it are not decorative either. They run from the polynomial curve fitting in chapter 1 through Bayesian linear regression in chapter 3, the kernel methods and Nadaraya-Watson model in chapter 6, and the mixture models in chapter 9, where at one point, faced with an intractable maximisation, I have simply written "Hideous! So use EM." That is not the annotation of someone skimming for a trend. It is the annotation of someone doing the work for its own sake.

I am telling you this not as autobiography but because it is the whole argument in miniature. The reason I can see that AI visibility is Bayesian is not that I read a trend piece about it last quarter. It is that I was studying the exact posterior distribution that governs it, for its own sake, before there was a marketing use for it. The diagram is not an illustration I commissioned. It is the thing itself, and it sat in a notebook waiting for the rest of the world to need it.

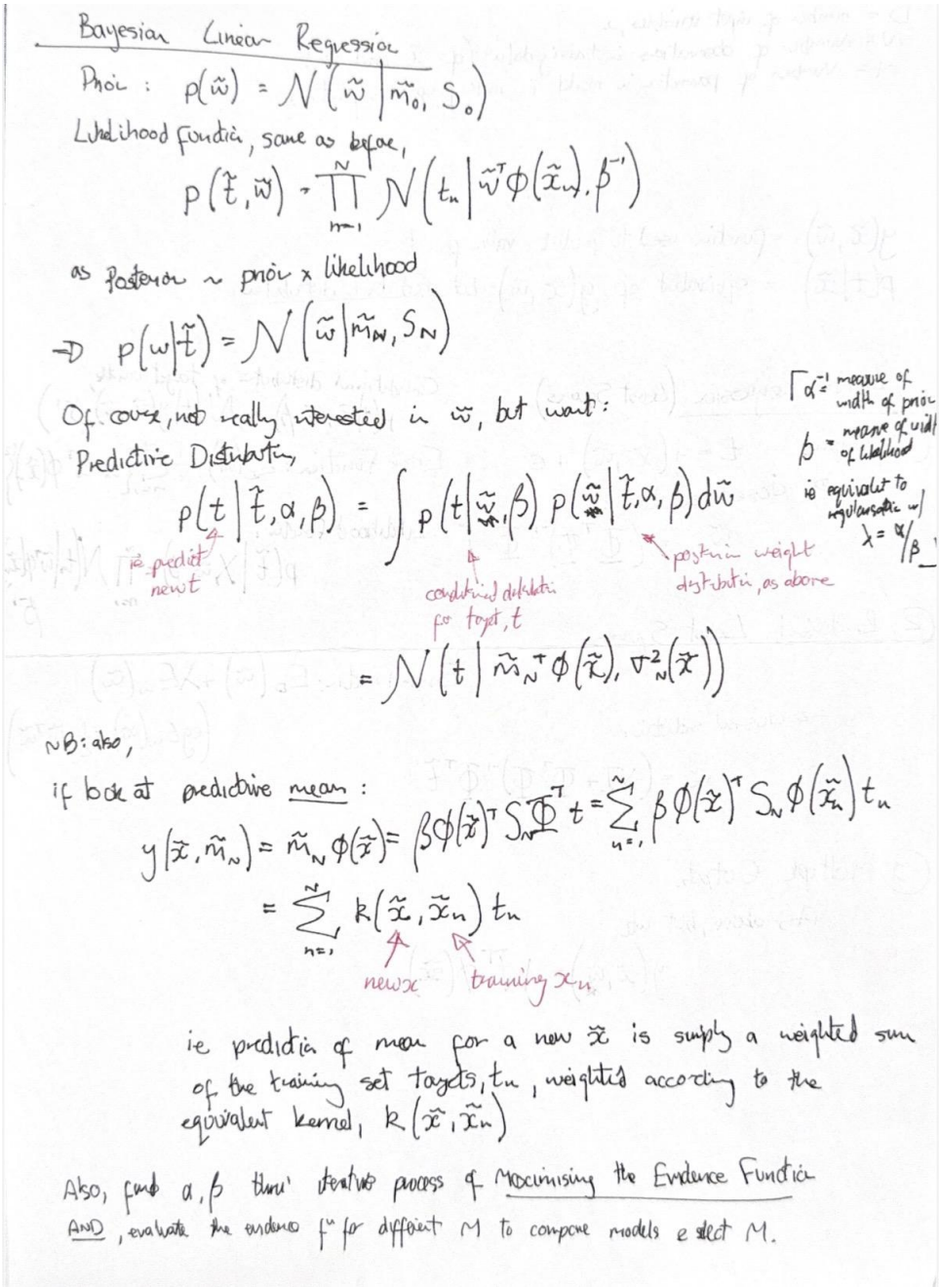


Figure 1. Bayesian linear regression, in the author's own handwriting, in a copy of Bishop's Pattern Recognition and Machine Learning that has been his since January 2007.

Marketing already had a Bayesian model of brand choice. It was missing a matrix

Marketing science worked out that brand choice is Bayesian thirty years ago. Tülin Erdem and Michael Keane modelled it in **Marketing Science** in 1996 (DOI 10.1287/mksc.15.1.1): a buyer holds a belief about each brand as a mean and a variance, an expected value and an uncertainty, and updates both as advertising and experience arrive. It is a good model. It is also, quietly, the model sitting underneath most of the recent "AI visibility is about priors" writing, mine included.

It carries one simplification, and in an AI world that simplification is the interesting part. Erdem and Keane treat each brand's belief as independent. Your belief about one brand is a separate pair of numbers from your belief

about the next. Updating one does not touch the other. There is no term anywhere in the model for the belief about Brand A and the belief about Brand B being entangled.

Look again at the posterior in the diagram. It is not a list of independent variances. It is N(w | m_N, S_N), where S_N is a covariance matrix. The numbers off the diagonal are precisely the thing the independent-belief model leaves out: they encode how a belief about one component moves a belief about another. The full Bayesian treatment always carried the term. The marketing version dropped it, and dropped it reasonably, because for a human choosing between fabric softeners the entanglement is small enough to ignore.

I want to be careful about what I am claiming here, because it is easy to overstate. I am not saying Erdem and Keane got the maths wrong, or that thirty years of choice modelling missed something obvious. Their model omits the off-diagonal term because for the problem they were solving it barely mattered. The point is narrower and more useful than "marketing used the wrong model": the term that was safe to drop for a human at a supermarket shelf is exactly the term that stops being safe to drop once the thing doing the choosing is a language model. The covariance was always in the full posterior. It just never earned its keep until now.

Because inside a language model it is not small, and it is not ignorable. Anthropic's interpretability team showed in 2022 that models represent far more features than they have dimensions by packing them into overlapping, non-orthogonal directions, a phenomenon they named superposition (Elhage et al., "Toy Models of Superposition"). Overlapping directions are off-diagonal covariance made physical. When your brand shares a direction with the generic average of your category, the model's belief about you and its belief about everyone else in that category are the same numbers. That is the covariance term the independent-belief model omits, and it is why generic content does not merely fail to stand out. It fails to exist as anything separable at all.

Inside a real model, that interference is not hypothetical

It would be fair to stop me at this point and say the covariance term is only interesting if it does something. A matrix can carry off-diagonal entries in principle and still have them sit near zero in practice. The reason to take the interference seriously is that mechanistic interpretability research has now looked inside real models and found it there.

Start with why superposition happens at all. A model has a **fixed number of dimensions** in each layer and needs to represent far more distinct concepts than that. So it does the only thing available: it stores concepts in directions that are not orthogonal to each other, accepting a little interference between them as the price of fitting more in. In their 2022 work, Elhage and colleagues showed that when a toy model is forced to pack more features than it has dimensions, the features it stores arrange themselves into regular geometric configurations. The right physical picture is the classical Thomson problem, points repelling each other on a sphere and settling into polytopes, digons and pentagons and tetrahedra, not any exotic quantum analogy. I have checked that comparison carefully and the geometry is the sober, load-bearing version of the claim. The features arrange to minimise mutual interference, but they cannot eliminate it, because there is not enough room.

That geometry is the covariance matrix from the previous section, realised in a model's weights. Two concepts stored in nearly the same direction have a large off-diagonal term whether anyone intended it or not. Activate one and you partly activate the other. For a brand, "nearly the same direction as the category average" is the default outcome of publishing content that reads like everyone else's content. You do not get absorbed because the model is careless. You get absorbed because the geometry has no spare dimension to give you.

The corollary is the encouraging half. Anthropic's 2024 work (Templeton et al., "Scaling Monosemanticity") showed that features in a production-scale model can, with the right techniques, be pulled apart into clean, individually interpretable directions: a feature for the Golden Gate Bridge, a feature for a specific bug in code, a feature for a particular tone of voice. A concept can earn its own clean feature rather than being smeared across a dozen shared ones. Which of those two fates a concept meets is not random. It depends on whether the training data gave the model a reason, and enough distinct evidence, to allocate the capacity. That is the whole game for a brand, stated in the model's own terms: distinct enough, and corroborated enough, to be worth a dimension.

Two kinds of memory, one kernel

The kernel view earns its keep a second time, because it explains a distinction I have written about before without having a clean account of why it is real. Getting retrieved inside a single answer is not the same operation as being baked into the model's standing beliefs. I have called these AEO and GEO, and treated them as two different problems. They are not. They are the same kernel operating over two different sets of data.

Retrieval within one answer, what I have been calling AEO, is the kernel running over the context window: the model computes similarities between your question and the material in front of it right now, and produces a weighted sum of that. The trained prior, GEO, is the same kernel running over the training corpus: similarities between the question and everything the model ever saw, folded into the weights during training. One operates over what is in the room. The other operates over everything the model has ever read. Same arithmetic, different t_n.

Interpretability gives this a concrete mechanism rather than an analogy. Anthropic's 2022 work on induction heads (Olsson et al.) reverse-engineered a specific circuit that, having seen the pattern [A][B] earlier in the current context, predicts [B] again on next seeing [A]. That is kernel-weighted retrieval in the flesh: find the earlier occurrence by similarity, copy forward what followed it. It is the in-context half, the short-timescale kernel, made mechanical. It is also, not by coincidence, exactly the "memory-based method" framing in Bishop's chapter 6: kernel methods retain the training data and make a prediction by weighting it, rather than compressing it away into parameters and discarding it. The induction head is that idea running one query at a time.

The practical consequence is that AEO and GEO respond to completely different levers on completely different clocks. The context-window kernel updates every single query: get your material in front of the model, through retrieval or a live source, and it counts immediately, for that answer only. The training-corpus kernel updates when the model is next retrained on a corpus that contains enough of you to shift the weights. One is this afternoon. The other is the next training run. Confusing the two is why people expect a burst of content to move their standing AI visibility this week, and are then surprised when it does not. I have set out the layered version of this argument in "AI Visibility Is a Layered Problem" and "There Are Two Kinds of AI Memory." The maths is why the layers are really there and not just a convenient way to talk.

Being nearly visible is worse than being invisible

There is one more finding that changes the risk calculation, and I have developed it at length in a companion post, "Your brand doesn't get its own space inside an AI model," so I will keep it brief here.

In March 2025 Anthropic traced the full computational path a model takes from prompt to output. One result matters for brands above all the others: a confident correct answer and a confident hallucination route through the same machinery. There is no separate hedging mode, no internal flag that says "I am guessing now." A "known entity" circuit suppresses the model's default caution when it recognises a name. When that recognition is accurate, you get a confident right answer. When the model recognises the name without actually holding the facts, you get a confident wrong one, delivered with identical certainty.

The implication for a brand is uncomfortable. Being familiar-sounding but thinly represented is not a safe halfway house on the road to being well-known. It is the state in which the model asserts things about you, wrongly, with full confidence, because your name is recognisable enough to switch off its caution but your representation is too weak to make the assertions true. Nearly visible is the failure mode, not the consolation prize. The companion post works through what that means for how you audit your own AI presence.

What to actually do

The programme falls straight out of the maths, and none of it is a growth hack.

Earn your own direction. The lever that matters is distinctiveness, not volume, because the thing that gets you a clean, separable feature rather than a smear across the category average is being genuinely unlike the generic centre. Ten thousand words that read like everyone else's ten thousand words buy you a larger weight on a direction you share with your competitors, which is worth almost nothing. One page that only you could have written, with named companies, real numbers, specific failures, buys you the beginnings of a direction of your own. This is the same reason I write about Redgate and Syskit by name and with figures rather than about "a SaaS company I advised." Specificity is not a stylistic preference. It is how you get allocated a dimension.

Build the prior with corroboration you do not control. The training-corpus kernel weights evidence, and evidence written by you about you is the weakest kind. Third-party mentions, independent write-ups, citations in material the model did not learn was yours, these are what move the standing prior, because they are what a well-built model has learned to trust. Self-published claims raise your weight fractionally on a direction the model already discounts.

Treat AEO and GEO as two clocks and run both. The context-window kernel is the short game: make it trivial for a model to retrieve you in the moment, through clean, well-structured, quotable source material. The training-corpus kernel is the long game: accumulate the distinct, corroborated presence that eventually shifts the weights. Neither substitutes for the other, and expecting the short game to deliver the long game's results on the long game's timescale is the most common and most expensive mistake I see.

Measure both, separately. I built a system that captures where my content appears across Google, Google AI Mode, ChatGPT and Gemini precisely because the two kinds of visibility move on different clocks and blur together if you only look once a month. You cannot manage the covariance term you cannot see. The brands that will be cited by AI in three years are the ones already doing the unglamorous work of becoming mathematically distinct, one specific, corroborated, un-averageable page at a time. The maths has been sitting in a notebook since 2006 waiting to be read this way. The only question left is whose training data you are distinct enough to survive.

Related reading:

- Brand visibility is now a Bayesian problem
- AI Visibility Is a Layered Problem
- There Are Two Kinds of AI Memory
- GEO vs AEO: Why the Distinction Matters for B2B Brands