A B2B AEO PLAYBOOK · 2026

Attention Engineering.

How the 2017 paper that powers ChatGPT, Claude, Gemini, and Perplexity quietly rewrote the rules of B2B search.

For founders and growth leads at B2B and SaaS companies who notice their competitors getting cited in AI search — and want to understand why, then fix it.

SOURCE PAPER

Attention Is All You Need

AUTHORS

Vaswani et al. · Google · NeurIPS 2017

Published by Citelane
citelane.com
SEO + AEO + GEO for B2B + SaaS

CHAPTER ONE

The paper, in *sixty seconds.*

Eight researchers at Google. Fifteen pages. One architecture. Every AI engine your B2B buyer now uses to evaluate vendors runs on the math inside it.

In June 2017, Vaswani and seven co-authors published **Attention Is All You Need** at NeurIPS. It introduced the **Transformer** — a neural network architecture that replaced recurrence and convolution with one mechanism: **self-attention**.

Before the Transformer, language models read text word by word, sequentially. Slow. Bad at long-range context. The Transformer threw that away. It said: let every word in a passage directly inspect every other word and decide how much each one matters to the meaning of the whole.

That single shift made GPT possible. Claude possible. Gemini, Perplexity, every AI Overview in Google search — all of it. The architecture is now the substrate of how machines read and recommend content on the open web.

> *"Most B2B SEO playbooks were written for a 2015 Google. The engines making vendor recommendations in 2026 don't read that way anymore."*
>
> — THE CITELANE AUDIT ENGINE, AFTER 40+ B2B + SAAS AUDITS

The rest of this playbook is one thing: a translation. Five concrete content layers, drawn directly from how attention actually works inside these models, that decide whether your SaaS gets cited or skipped.

Self-attention

Every token weighs every other token. Context is global, not sequential. Density and proximity matter.

Multi-head attention

The model runs several attention passes in parallel, each tracking a different signal. More heads, more citation surface.

Positional encoding

Order is injected as math. Where information sits on the page is part of what it means to the model.

Context window

Attention is finite. Tokens far from the query lose weight. Page-top placement is not aesthetic — it is mechanical.

CHAPTER TWO

Answer first. *Then density.*

01 Direct Answer Openings

Why it works: attention scores reward direct, factual sentences. When an AI engine retrieves and synthesises an answer, it pulls answer-shaped passages first. Brand intros and marketing copy carry low attention weight against query tokens.

> THE RULE
>
> The first 40 words of every commercial page must be a direct, factual answer to that page's primary question — before brand, before hero copy, before anything else.

- [ ] Page leads with a clear, declarative answer in under 200 characters
- [ ] Answer appears *before* brand intro or marketing language
- [ ] Sentence opens with subject + verb, not "We help…" or "At [Brand]…"
- [ ] Answer is self-contained — no "click below to learn more"

02 Topic Density

Why it works: self-attention compares every token to every other token. Pages dense with related entities and intent-matched phrases generate higher internal attention weights — and pull citations far more reliably than thin pages built around a single keyword.

> THE RULE
>
> One page should resolve the primary question and 3–5 closely related sub-questions on the same URL. Density is the moat.

- [ ] Page covers 3–5 related questions, not 1
- [ ] Includes the 5–7 entities a buyer expects: use cases, integrations, alternatives, pricing factors, limitations
- [ ] FAQ block at the bottom answers the obvious follow-ups
- [ ] No thin pages under 800 words on commercial-intent URLs

CHAPTER THREE

Structure is *signal.*

03 Structural Hierarchy

Why it works: the Transformer adds *positional encoding* on top of every token. Order and structure carry mathematical weight. Semantic HTML and JSON-LD schema are how you hand the model a clean map of what's primary and what's supporting.

> THE RULE
>
> Every commercial page is structured with semantic HTML and the right schema type. The model should never have to guess what your page is about.

- [ ] One H1 per page, matching the primary query intent
- [ ] H2s map cleanly to the sub-questions the page answers
- [ ] FAQPage schema on commercial and bottom-of-funnel pages
- [ ] Article, Product, or SoftwareApplication schema where applicable
- [ ] Internal links use descriptive anchor text — never "click here" or "learn more"

04 Multi-Question Pages

Why it works: multi-head attention runs several attention passes in parallel, each tracking a different aspect of the input. A page that resolves several related questions activates more heads simultaneously — which translates directly into more citation surface area inside the model.

> THE RULE
>
> Stop building one-question-per-URL pages. Consolidate the question cluster, do not split it.

- [ ] Commercial pages consolidate the cluster: "What is X" + "How does X work" + "X vs Y" + "X pricing" share one URL where intent allows
- [ ] Compare and "vs" pages are honest, balanced, and ungated
- [ ] Each sub-section opens with its own answer block — not a transition sentence
- [ ] Sub-sections are reachable via a table-of-contents jump menu at the top

CHAPTER FOUR

Top of page. *Or invisible.*

05 Top-of-Page Priority

Why it works: attention is finite. As tokens drift further from the query inside the context window, their attention weight falls. A best answer buried 800 words deep is, for citation purposes, a best answer that does not exist.

> THE RULE
>
> The most citation-worthy paragraph on every page lives in the first 200 words. Always.

- [ ] Most quotable, citation-ready paragraph sits in the first 200 words
- [ ] Definitions and answer blocks come *before* case studies or storytelling
- [ ] TL;DR or summary box at the top of long-form content
- [ ] Verify what the model actually sees: render the page, copy the first 1,500 chars, audit

BONUS LAYER

AI Crawler Access

All five layers above are useless if the AI engines cannot crawl the page. Most B2B sites are accidentally blocking the bots that feed citations.

- [ ] `robots.txt` explicitly allows: GPTBot, ChatGPT-User, Google-Extended, PerplexityBot, ClaudeBot, anthropic-ai, Applebot-Extended
- [ ] No CAPTCHA or aggressive Cloudflare challenge on key landing pages
- [ ] Critical content not gated behind email, JS-only renders, or paywalls
- [ ] Sitemap submitted to Bing Webmaster Tools (Perplexity uses it)

CHAPTER FIVE

The *18-point* AEO audit.

Run this against any commercial page on your site. Each unchecked box is one point of attention you are leaving on the table.

01 — Answer Openings

- [ ] 40-word answer block in line 1
- [ ] No brand intro before the answer
- [ ] Subject + verb sentence opener

02 — Topic Density

- [ ] 3–5 related questions per page
- [ ] 5–7 expected entities present
- [ ] FAQ block at page bottom
- [ ] 800+ words on commercial URLs

03 — Structural Hierarchy

- [ ] Single intent-matched H1
- [ ] H2s mirror sub-questions
- [ ] FAQPage / Product / Article schema

04 — Multi-Question Pages

- [ ] Cluster consolidated to one URL
- [ ] Honest, ungated vs / compare pages
- [ ] Sub-section answer blocks

05 — Top-of-Page Priority

- [ ] Best answer in first 200 words
- [ ] TL;DR or summary box at top

06 — Crawler Access

- [ ] GPTBot / Google-Extended allowed
- [ ] PerplexityBot / ClaudeBot allowed
- [ ] No CAPTCHA on landing pages
- [ ] Critical content not JS-only
- [ ] Sitemap submitted to Bing

CITELANE

SEO + AEO + GEO for B2B + SaaS.

THIS PLAYBOOK

Translated from *Attention Is All You Need*, 2017.