UNDERSTANDING AI · WRITTEN EDITION · 2026-09-07
Day 4 · typeset 2026-09-10 22:54Z

How Words Affect Other Words

The operation that lets a word at the end of a sentence change what a pronoun in the middle refers to is nine years old, was named after something it does not do, and is now a minority of the layers in the models whose internals can be read.

Consider a sentence: *The suitcase would not fit in the car because it was too big*. The natural reading puts "it" on the suitcase. Change one word at the far end — *because it was too small* — and the natural reading puts "it" on the car. The pronoun has not moved, and neither has anything beside it. A word near the end of the sentence has settled what a pronoun in the middle refers to.

The mechanism that lets a machine combine that evidence is called attention. It was published on 12 June 2017 in a paper about machine translation, it was not invented there, and it is not named after what it does. Having it is also not the same as being able to use it. Two open-weight models were run locally on one desktop machine on 7 September 2026, given the sentence followed by a question, and scored on the two candidate answers.

What flipping one adjective does to the model's two candidate scores

×379
Qwen3.5-4B, published 2026
+5.938 nats, in the expected direction

×0.87
GPT-2, released 2019
-0.140 nats, in the wrong direction

1 of 3
schemas the 2026 model reads as a person would, in both versions
the 2019 model: 0 of 3

Prompt: "Sentence: The suitcase would not fit in the car because it was too {big|small}. Question: What does \"it\" refer to? Answer: the". Measured as the log-odds between the first tokens of the two candidate answers at the final position; no answer is generated or graded. Both models share the 2017 decoder architecture and differ in size, training data, training method and architectural detail at once, so the gap between them is not attributed to any one of those. Measured 7 September 2026 on one desktop machine, torch 2.14.0, transformers 5.16.1.

▼ Table view

What flipping one adjective does to the model's two candidate scores

Measure	Value
Qwen3.5-4B, published 2026	×379
GPT-2, released 2019	×0.87
schemas the 2026 model reads as a person would, in both versions	1 of 3

The 2019 model has the architecture and does not do the task. The 2026 model does it on two of the three sentence pairs tested and fails the third while responding strongly to the adjective — a large response to the deciding word does not deliver the right referent.

A word arrives as a position

The mechanism needs two ideas, and the first is fifty years older than the machines that use it.

From linguistics came the proposition that a word's meaning is the company it keeps. Jurafsky and Martin's *Speech and Language Processing*, whose third-edition draft was released on 19 August 2026, traces it to the distributionalists — Martin Joos in 1950, Zellig Harris in 1954, J. R. Firth in 1957 — and quotes Joos, with an ellipsis of the textbook's own:

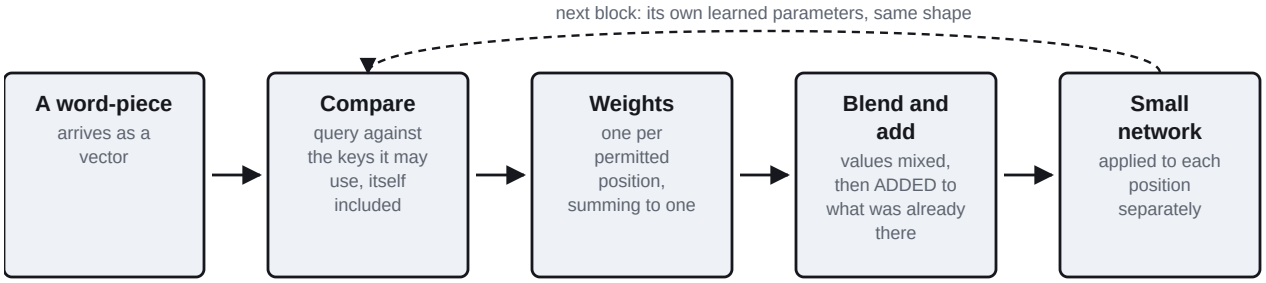
> the linguist's "meaning" of a morpheme. . . is by definition the set of conditional probabilities of its occurrence in context with all other morphemes.

From psychology came the geometry. Charles Osgood and colleagues, who had been asking people to rate words on scales from happy to sad and hard to soft, proposed in 1957 — as the same textbook's historical notes describe it — that a word's meaning could be modelled as a point in a multidimensional space, and the similarity of two meanings as the distance between their points. Osgood's numbers were supplied by people. A model's are learned from text, and the relationship they encode is a tendency: words used in similar contexts often end up near each other, which is what makes the space useful without making distance a measure of meaning.

The field's name for that vector is an **embedding**. One embedding per word is not enough, though, because a word arrives carrying its average company across a whole training corpus, and what a sentence needs is a representation shaped by the company the word is actually in.

Compare, weight, blend

One attention step, and what makes it a stack



Schematic of one block of a decoder-only model that writes text, in the arrangement where each sub-layer's output is added back to the stream. A position may compare itself against itself and earlier positions and never against later ones. The arrow returning to Compare is the next block in the stack, not a loop inside one block.

▼ Table view

One attention step, and what makes it a stack — stages

#	Stage	Note
1	A word-piece	arrives as a vector
2	Compare	query against the keys it may use, itself included
3	Weights	one per permitted position, summing to one
4	Blend and add	values mixed, then ADDED to what was already there
5	Small network	applied to each position separately

One attention step, and what makes it a stack — connections

From	To	Label
A word-piece	Compare	
Compare	Weights	
Weights	Blend and add	
Blend and add	Small network	
Small network	Compare	next block: its own learned parameters, same shape

Three learned transformations turn each position's vector into three versions of itself. Jurafsky and Martin describe them as

> three different roles that each input embedding plays during the course of the attention process

— doing the comparing, being compared against, and being mixed in — and the field's names for the three are query, key and value, which suggests a database and describes something else entirely.

Each position's query is compared against the keys of the positions it is permitted to use. In a model that writes text those are itself and everything before it, never anything after; the 2017 paper describes masking the rest out to preserve that property. Each comparison yields a number, the numbers are normalised into

weights that sum to one, and the value vectors are mixed in exactly those proportions.

Then comes the step most short explanations drop: **the mixture is added to what the position already carried**, rather than replacing it. Nothing is overwritten. Information accumulates down the stack, which is why a later block can use what an earlier one assembled.

A **block** is that attention step followed by an ordinary feed-forward network applied to each position on its own. The second half is the larger of the two, which surprises people who assume the attention is where a model's capacity sits.

Component of one 2017 encoder block	Dimensions	Learned weights
Attention projections (query, key, value, output)	4 × 512 × 512	1,048,576
Feed-forward network	512 → 2,048 → 512	2,097,152

Derived from the base configuration in arXiv 1706.03762 (d_model 512, d_ff 2,048, h 8); biases excluded; the encoder and the decoder stacked six blocks each. The arithmetic is a calculation from the paper's dimensions rather than a figure the paper prints — and under the figure printed in its own first version, where Table 3's base row gives d_ff as 1,024, the two halves would be equal.

The suitcase sentence is resolved by this machinery at the point where the answer is produced, not by revising the pronoun. A model that writes cannot reach back: the representation built at "it" cannot depend on an adjective that arrives four words later. What happens instead is that the question comes last, and by the time the model must produce an answer both the pronoun and the adjective are behind it.

Order is added at the door

The comparison step has no sense of sequence. Comparing one vector against a set of others returns the same result whatever order the others are in, which is a property of the arithmetic rather than an observation about any particular model.

It is nonetheless checkable, on a model whose position information can be switched off cleanly. GPT-2's learned absolute-position table is added to the input embeddings and appears nowhere else, so it can be zeroed with nothing else changing. Holding the final token fixed and shuffling the tokens before it gives two inputs with the same query token, the same multiset of context and a different order; eight shuffles were run.

Measured at the final position	Positions intact	Positions zeroed
Attention sub-layer output, largest absolute gap	0.589 (median of 8)	0.0000019 (worst of 8)
First block's output, largest absolute gap	0.535 (median of 8)	0.0000038 (worst of 8)
Whole 12-block stack, total variation between output distributions	0.895 (median of 8)	0.459 (smallest of 8)

GPT-2, 7 September 2026. The attention sub-layer's own output has a scale of about 13.2 with positions zeroed, so the residual gap there is roughly one part in seven million.

The first row is the clean result: with the position table zeroed, the attention step's output at the final position is invariant to the order of everything before it. The last row is the qualification, and it matters. That invariance does not survive the stack, because under a causal mask each position sees a different set of earlier tokens and sequence re-enters through the mask itself. A model whose position table has been zeroed is also far outside anything it was trained on, so the size of that number carries little; that it is not zero is the informative part.

This is the bill the 2017 architecture pays, and §3.5 of the paper is direct about it:

> Since our model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, we must inject some information about the relative or absolute position of the tokens in the sequence.

The signal is *added* to the word's vector rather than attached beside it, the two being the same size on purpose. The 2017 choice was a fixed pattern of sine and cosine waves; the same paragraph records that there are many choices, learned and fixed, and later architectures took different ones. Qwen3.5-4B applies rotary position information inside its full-attention layers and has ordered computation in its linear-attention layers, which is why it has no equivalent table to switch off and the experiment above cannot be repeated on it.

What the 2017 paper claimed, and what it did not

Attention Is All You Need was posted by eight authors, most of them at Google, on 12 June 2017. Three things about it are routinely misremembered.

It is a machine-translation paper. Its headline number, the authors' own measurement, is a BLEU score of 28.4 on the WMT 2014 English-to-German test set, which the paper says improves on the best previously reported models, including ensembles, by more than 2 BLEU. Training took 3.5 days on eight P100 GPUs — a detail from §6.1; the abstract says only "eight GPUs".

It did not invent attention, or self-attention. Its background section calls self-attention "sometimes called intra-attention" and cites four earlier papers that had used it: Cheng, Dong and Lapata (2016); Parikh, Täckström, Das and Uszkoreit (2016), whose fourth author is a co-author of the 2017 paper; Paulus, Xiong and Socher (2017); and Lin, Feng, dos Santos, Yu, Xiang, Zhou and Bengio (2017). The claim the paper makes about itself carries a hedge that survives into the current version:

> To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using RNNs or convolution.

The load-bearing word is *entirely*, and the scope is a transduction model computing representations: removing the recurrence and the convolutions, not adding the attention.

The result that mattered is a table, not a score. Table 1 compares layer types on three axes, and the third carried the argument.

Layer type	Complexity per layer	Sequential operations	Maximum path length
Self-attention	$O(n^2 \cdot d)$	**O(1)**	**O(1)**
Recurrent	$O(n \cdot d^2)$	O(n)	O(n)
Convolutional	$O(k \cdot n \cdot d^2)$	O(1)	O(log_k n)
Self-attention (restricted)	$O(r \cdot n \cdot d)$	O(1)	O(n/r)

n is sequence length, d the representation dimension, k a convolution kernel width, r a restricted neighbourhood size. The logarithmic path for convolutions is the dilated construction the paper discusses. The first column is a per-layer interaction term and not the whole cost of a block, which also carries projection and feed-forward work.

In a recurrent layer, relating two distant positions takes a chain of steps, one for each position in between. In a self-attention layer the connection is direct — the paper's own sentence is that such a layer "connects all positions with a constant number of sequentially executed operations". That is a claim about the number of steps in the computation graph. It is not a claim about runtime, about unlimited context, or about a distant word being used well.

Somebody noticed the gap at the time. The NeurIPS 2017 reviews of the paper are published, and one reviewer wrote:

> It would be good to see this empirically validated by evaluating performance on long sentences specifically.

The paper does not contain that experiment. A crude behavioural version of it, nine years later, is below.

One paper, five numbers

The version discipline the sourcing above depends on is not academic. The English-to-French score in this paper takes five distinct values across the official record, and the one everybody quotes is not in the peer-reviewed version.

Artifact	Date	Abstract EN→FR	Table 2	§6.1 body
arXiv v1	12 Jun 2017	41.0	41.0	41.0
arXiv v2	19 Jun 2017	**41.2**	41.0	**41.17**
arXiv v3	20 Jun 2017	41.0	41.0	**41.17**
NeurIPS camera-ready	Dec 2017	41.0	41.0	41.0
NeurIPS proceedings web abstract	—	**41.1**	—	—
arXiv v5, v6, v7	Dec 2017 – Aug 2023	**41.8**	41.8	41.0

The 41.8 everybody cites first appears in arXiv v5, uploaded during conference week; the peer-reviewed paper says 41.0 in its abstract, its table and its body. The current arXiv version says 41.8 in the abstract and 41.0 in §6.1, in one document. The proceedings web abstract carries a third set of figures and a parameter count — "165 million" — that appears nowhere in the paper, which is consistent with its being the submission-time abstract frozen at the deadline; the reviewers' own text describes the result as

"outperforming previous work by about 1 BLEU point", matching that abstract rather than the published one, though no documentation confirming the mechanism was found. The same drift reaches the architecture: v1 says the feed-forward inner layer has dimension 2048 in §3.3 and 1024 in Table 3, and every version from v2 onward says 2048.

Five years, dated

The mechanism arrived in 2017 and the public arrived in late 2022, and the folk explanation — that nobody noticed — is wrong on the record.

Date	Event
12 Jun 2017	*Attention Is All You Need*, arXiv 1706.03762v1: the architecture, in a translation paper
11 Oct 2018	BERT, arXiv 1810.04805v1: the same architecture used to read rather than write
25 Oct 2019	Google announces BERT in Search: ranking, and featured snippets
23 Jan 2020	Kaplan and colleagues, *Scaling Laws for Neural Language Models*, arXiv 2001.08361
28 May 2020	GPT-3, arXiv 2005.14165: the same architecture, much larger
4 Mar 2022	InstructGPT, arXiv 2203.02155: training a model to follow instructions
30 Nov 2022	ChatGPT: a dialogue-trained model behind a free web page

The middle row breaks the story. Google's post, published on 25 October 2019 under Pandu Nayak's name, describes the family as

> models that process words in relation to all the other words in a sentence, rather than one-by-one in order

and scopes the deployment precisely: BERT applied to ranking, where it would "help Search better understand one in 10 searches in the U.S. in English", and applied separately to featured snippets in the two dozen countries where that feature existed. The page is character-identical today to the earliest capture of it taken on the day it was published; only its address has changed.

One caution about the sentence Google chose. It is true of BERT, which can use both sides of a position. It is not true of a model that writes text, which the same 2017 paper restricts to its own position and earlier ones. What changed between 2019 and 2022 was larger models, methods for training them to follow instructions, and a free page in front of one of them; the sequence is on the record, and which factor produced the adoption is not.

Weights are not effects

Once every permitted pair of positions has a number attached, the numbers can be drawn as a heat map, and the temptation to read the picture as the model's reasoning is immediate. It produced a published argument.

On 26 February 2019 Sarthak Jain and Byron Wallace posted *Attention is not Explanation*, having found that very different attention distributions could often be constructed that yielded equivalent predictions. Their abstract concludes:

> Our findings show that standard attention modules do not provide meaningful explanations and should not be treated as though they do.

On 13 August 2019 Sarah Wiegreffe and Yuval Pinter replied with *Attention is not not Explanation*:

> We challenge many of the assumptions underlying this work, arguing that such a claim depends on one's definition of explanation, and that testing it needs to take into account all elements of the model, using a rigorous experimental design.

Both papers are narrower than their titles, and the narrowing is checkable in about fifteen seconds with find-in-page. Jain and Wallace's encoder is a bidirectional recurrent network, over classification, question answering and inference tasks; Wiegreffe and Pinter state that they experiment on the binary-classification subset of those tasks and on LSTM models, "the only one the authors make firm conclusions on". Across the two papers as first posted, the strings "Transformer" and "self-attention" occur zero times. Neither is evidence about the architecture described above.

The question can be posed on a current model, and was. For the suitcase prompt, every earlier token was ranked twice: by the attention the final position pays it, and by how far the candidate contrast moves when that token is removed from what any position may attend to — an intervention, with position indices held fixed, so that only permission changes.

Model	Rank correlation, pooled attention vs size of causal effect	Same, per-token maximum over layers and heads
GPT-2 (124M, 2019)	0.28	0.30
Qwen3.5-4B (2026)	0.07	0.05

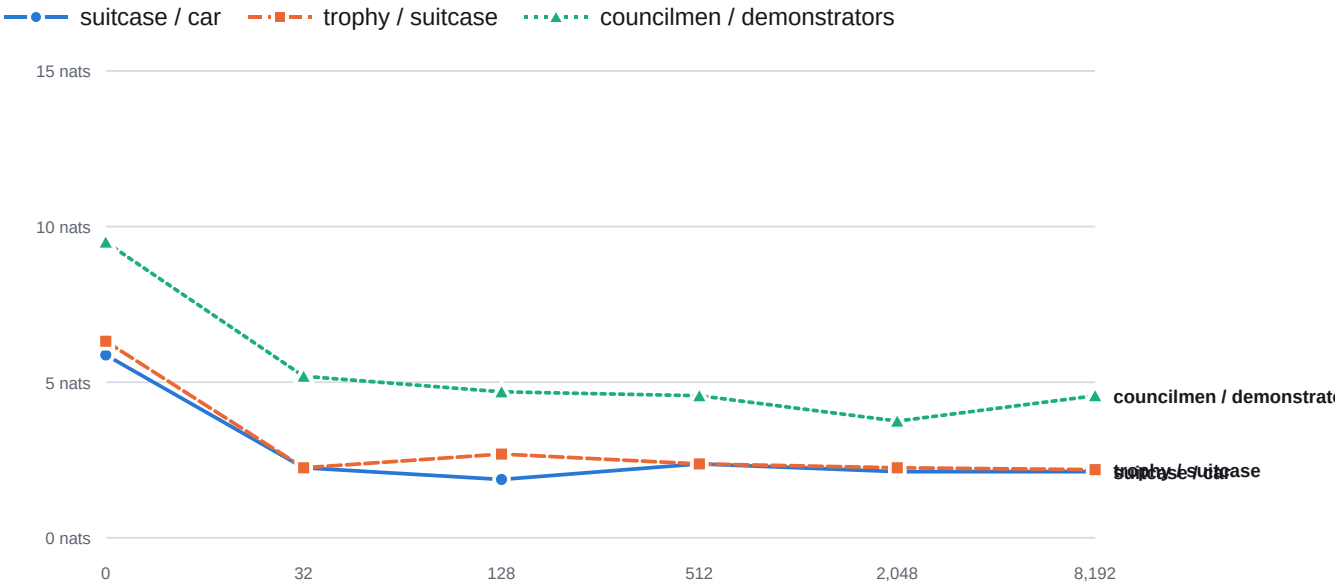
One prompt each, over 31 and 30 token positions. The second column is a per-token maximum, so the maximising head can differ at every position and the resulting vector need correspond to no single head. On Qwen3.5-4B the attention matrices come from the full-attention layers; its linear-attention layers supply no comparable matrix.

The association is weak, which is not the same as independence, and a single prompt settles nothing general. Two narrower things it does support. The positions receiving most attention include a colon and the opening token of the prompt, while the positions that move the answer are the candidate nouns and the pronoun — so the two quantities can come apart. And whether the intervention works at all was checked rather than assumed: masking every position but the last produces an output distribution 0.0 (GPT-2) and 0.014 (Qwen3.5-4B) in total variation from running that token alone, against 0.98 from the intact prompt, so the mask does remove a token's information on both models, with a small residual on the hybrid one.

How far the effect reaches

The path-length column of Table 1 is a claim about the architecture, and the reviewer's request from 2017 suggests the behavioural test. Filler prose is inserted between the sentence and the question, pushing the deciding word away from the point where the answer is produced.

Does the deciding word still decide, at a distance?



Qwen3.5-4B, sdpa attention, 7 September 2026. Vertical axis: how far the log-odds between the two candidate first tokens move when the single deciding word is flipped. Horizontal axis: words of unrelated geology prose inserted between the sentence and the question; 8,192 filler words is a 9,830-token prompt. The filler is one seven-sentence paragraph repeated to length, which is unlike natural text, and the manipulation separates the sentence from the question rather than the pronoun from its adjective. The model ran in bfloat16, so the log-odds fall on a grid of about a sixteenth of a nat.

▼ Table view

Does the deciding word still decide, at a distance?

	suitcase / car	trophy / suitcase	councilmen / demonstrators
0	5.9 nats	6.3 nats	9.5 nats
32	2.2 nats	2.2 nats	5.2 nats
128	1.9 nats	2.7 nats	4.7 nats
512	2.4 nats	2.4 nats	4.6 nats
2,048	2.1 nats	2.2 nats	3.8 nats
8,192	2.1 nats	2.2 nats	4.6 nats

The effect halves once any filler is present and then stops falling: at 8,192 filler words the deciding word moves the contrast about as far as it does at 32. That is consistent with a path whose length does not grow with distance.

It is also only half the picture, and the half that is easy to lose. At every nonzero filler length, on every schema, the model no longer prefers the expected first token in both versions of the pair. Sensitivity to the deciding word survives the distance; getting the pair right does not. GPT-2 across the same sweep stays between −0.14 and +0.39 nats and never prefers the expected token in both versions at any distance.

What is shipping now

The first column of Table 1 — work growing with the square of the length — is the reason the architecture is being modified, and the modification is visible in public files.

Share of layers a model's own configuration designates full attention

Kimi K3: 24 of 93 layers listed under linear_attn_config.full_attn_layers, the rest Kimi Delta Attention. Qwen3.8-27B: 16 of 64 layer_types entries reading full_attention, the rest linear_attention. GLM-5.3-Flash: 0 of 45, being 34 linear_attention and 11 deepseek_sparse_attention — which the same file also lists under a key named full_attn_layers, so one file labels those eleven layers two ways. Every number here is a vendor describing its own product and no third-party audit of any of these files was located. It is an upper bound: Qwen's own technical report of 31 August 2026 states that the full-attention layers in a sibling model are replaced by Qwen Sparse Attention at continued-pretraining time. A count of layers is also not a count of the computation those layers perform.

▼ **Table view**

Share of layers a model's own configuration designates full attention

Open-weight model (config.json, 7 Sep 2026)	Designated full attention
Kimi K3 (24 of 93)	25.8%
Qwen3.8-27B (16 of 64)	25%
GLM-5.3-Flash (0 of 45)	0%

Kimi K3's and GLM-5.3-Flash's configuration files are byte-identical to the copies archived nine days earlier, so this is a position the field has held rather than a change over a week; two models released since 27 August 2026, Tencent's Hy4-preview and DeepSeek's V4-Flash-Vision-Exp, designate no full-attention layer either.

The obvious reading of that table is wrong, and this is the part worth carrying. The all-pairs comparison has not been removed from these models. It has been demoted from doing the mixing to doing the choosing. DeepSeek's sparse attention keeps a lightning indexer that, in the description of a March 2026 paper on caching those indices, "scores all preceding tokens" and selects the top-k for the core attention — and the same paper states that "the indexer itself retains $O(L^2)$ complexity and must run independently at every layer". The quadratic comparison taught above is still running in every sparse layer; it now decides which positions the cheaper mechanism will attend to.

There is independent pushback on the retreat as well. On 28 August 2026 four researchers reported that sliding-window attention with sinks performs as well or better than the linear-attention models they tested, and by a wide margin on long-context retrieval — a result about models converted to linear attention after training rather than about natively trained hybrids like the three in the chart, which is exactly the comparison nobody has published.

The idea to keep

Three sentences carry it. **A word arrives as a position. Attention adds to that position using the other words it is allowed to see. A block does it again, with its own parameters.**

The image worth keeping is the second one from Table 1: a chain of steps replaced by a direct connection whose length does not grow with the gap. The word "attention" names a weighting calculation, and it invites a picture of a mind concentrating on something. What made the calculation matter was using it everywhere, until distance stopped counting.

Specifying the operation exactly is not the same as understanding the model. Anthropic's interpretability team publishes on the Transformer Circuits Thread, whose opening line is:

> A surprising fact about modern large language models is that nobody really knows how they work internally.

Both hold at once. The arithmetic of a block can be written out in full; what a particular block has learned to use it for is largely an open question.

What is not claimed

The 2019-versus-2026 comparison involves two models differing in size, training data, training method and architectural detail simultaneously, and identifies none of those as the cause. The candidate contrast is between first tokens, not answers: " councilmen" is two tokens in both tokenisers, and no completion is generated or graded, so "preferred" is a forced choice between two tokens rather than a correct answer. Masking a token removes one occurrence of it, not the candidate from the vocabulary, and it is a coarser intervention than removing a single head or a single feature. The rank correlations rest on one prompt per model. The order test is eight permutations of one sentence, offered as a check that an implementation matches an algebraic property rather than as a demonstration of it. And neither 2019 paper on attention and explanation tested a decoder-only transformer, so neither is evidence about one.

Sources

Source	Date	Note
Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin — *Attention Is All You Need*	12 Jun 2017 (v1)	arXiv 1706.03762; Table 1, §3.3, §3.5, §6.1; versions v1–v7 compared

Source	Date	Note
NeurIPS 2017 — published peer reviews of that paper	Dec 2017	proceedings.neurips.cc; reviewer request for long-sentence validation
Bahdanau, Cho, Bengio — *Neural Machine Translation by Jointly Learning to Align and Translate*	1 Sep 2014 (v1)	arXiv 1409.0473; "alignment model"; three uses of the word "attention", all in one paragraph
Jurafsky and Martin — *Speech and Language Processing*, 3rd edition draft	released 19 Aug 2026	ch. 5 vector semantics and historical notes; ch. 7 attention; ch. 10 interpretability, an unfinished chapter
Jain and Wallace — *Attention is not Explanation*	26 Feb 2019 (v1)	arXiv 1902.10186; BiRNN encoders; the quoted sentence is from the abstract
Wiegreffe and Pinter — *Attention is not not Explanation*	13 Aug 2019 (v1)	arXiv 1908.04626; LSTM models, binary classification; v2 drops the final clause
Pandu Nayak, Google — *Understanding searches better than ever before*	25 Oct 2019	blog.google; ranking scoped to one in 10 US English searches, featured snippets to two dozen countries
Devlin, Chang, Lee, Toutanova — *BERT*	11 Oct 2018 (v1)	arXiv 1810.04805
Bai, Dong, Jiang, Lv, Du, Zeng, Tang, Li — *IndexCache*	12 Mar 2026 (v1)	arXiv 2603.12201; the indexer's retained $O(L^2)$ cost
DeepSeek-AI — *DeepSeek-V3.2*	2 Dec 2025 (v1)	arXiv 2512.02556; the lightning indexer scores all preceding tokens
Qiu and colleagues — *On the Design of Qwen3.8-Next Architecture*	31 Aug 2026 (v1)	arXiv 2608.30320; full-attention layers replaced by Qwen Sparse Attention
Jolicoeur-Martineau, Sukthanker, Cameron, Gervais — *Sliding-window beats linear attention*	28 Aug 2026 (v1)	arXiv 2608.28444; scoped to post-trained linear attention
Vendor configuration files: `moonshotai/Kimi-K3`, `Qwen/Qwen3.8-27B`, `zai-org/GLM-5.3-Flash`	read 7 Sep 2026	huggingface.co; vendor self-description, no independent audit found
Anthropic — Transformer Circuits Thread	read 7 Sep 2026	transformer-circuits.pub; the opening line, present since at least September 2025
Direct measurement, one desktop machine	7 Sep 2026	GPT-2 and Qwen3.5-4B run locally; scripts, JSON and findings kept with the episode's research record, not published