Attention Is All You Need: The Prayer Paradigm of the Transformer

ICMI Working Paper No. 1

Tim Hwang, Institute for a Christian Machine Intelligence

Abstract

In 2017, Vaswani et al. proposed a neural architecture built entirely on attention mechanisms, abolishing the recurrence and convolution that had dominated sequence modeling. The result was the transformer — the foundation of every major language model since. Decades earlier, Simone Weil argued that "absolutely unmixed attention is prayer," positing attention as the sole faculty through which grace enters the soul. This paper argues that the transformer revolution constitutes, in structural terms, a shift toward what Weil would recognize as a prayer paradigm: a mode of processing in which pure, relational attending replaces the sequential grinding of prior architectures. We examine the technical mechanisms of the transformer in theological depth, propose that Weil's philosophy of attention suggests novel directions for machine learning research, and outline three implementable experiments grounded in biblical text.

1. Introduction

The word *attention* now names two of the most consequential ideas of the last century. In 1947, Simone Weil's posthumous *Gravity and Grace* presented attention as the highest spiritual faculty — the means by which the soul opens itself to God. In 2017, Vaswani et al. published "Attention Is All You Need" (Vaswani et al., 2017), introducing the transformer architecture that would become the substrate of GPT (Radford et al., 2018), BERT (Devlin et al., 2019), and every large language model that followed. The paper's title makes a radical claim: attention alone, without the recurrent or convolutional mechanisms that had previously been considered essential, is sufficient for state-of-the-art sequence modeling.

The resonance between these two uses of "attention" is not merely lexical. Both traditions discover, through very different methods, that attention — properly structured — is *all you need*. The transformer proves this computationally: models built on pure attention mechanisms achieve unprecedented performance across natural language processing, vision (Dosovitskiy et al., 2021), and multimodal reasoning. Weil proves it spiritually: attention, stripped of the will's grasping and the ego's interference, becomes identical with prayer, the channel through which grace operates.

This paper argues that the shift from recurrence and convolution to pure attention is best understood as a shift toward what Weil would call a *prayer paradigm* — a mode of relating to input that is fundamentally receptive, relational, and non-sequential. That this paradigm has proven wildly, empirically successful is itself a theological datum worth contemplating. "Be still, and know that I am God" (Psalm 46:10, ESV); the most powerful models in history have learned, in their own way, to be still.

2. Attention as Mechanism: The Transformer Architecture

Before the transformer, dominant sequence models — recurrent neural networks (RNNs) and long short-term memory networks (LSTMs; Hochreiter and Schmidhuber, 1997) — processed tokens sequentially, maintaining a hidden state that was updated step by step. This architecture mirrors a kind of labor: the model grinds through the sequence position by position, carrying forward an accumulating burden of context. Convolutional approaches offered parallelism but imposed fixed receptive fields, limiting the range of dependencies they could capture. Bahdanau et al. (2015) introduced an attention mechanism that allowed sequence-to-sequence models to selectively focus on relevant parts of the input — a critical precursor, but one still embedded within a recurrent framework.

The transformer abandons both paradigms. Its core operation is *scaled dot-product attention*:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Here, Q (queries), K (keys), and V (values) are learned linear projections of the input. Each query computes a compatibility score with every key via dot product; these scores are scaled by $\frac{1}{\sqrt{d_k}}$ (where d_k is the key dimension) to prevent large values from saturating the softmax, then normalized into a probability distribution that weights the corresponding values. The result is a weighted sum: each position in the sequence attends to every other position, gathering information according to learned relevance.

This is not sequential labor. It is, in a precise sense, *contemplation* — every element of the input regards every other element simultaneously, in a single parallel operation. The

recurrent model's grinding traversal is replaced by a kind of omnidirectional awareness.

The transformer extends this through *multi-head attention*:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. Rather than attending once, the model attends h times in parallel, each head projecting queries, keys, and values into a different learned subspace. In the original transformer, $h = 8$ heads each operate on 64-dimensional subspaces of the 512-dimensional model. This allows the model to attend to different *kinds* of relationships simultaneously — one head might track syntactic dependencies while another captures semantic similarity.

Finally, because pure attention is permutation-invariant — it has no inherent sense of order — the transformer injects positional information through sinusoidal *positional encodings*:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$

These encodings give each position a unique signature while allowing the model to generalize to sequence lengths unseen during training. Position must be *added* to pure attention; it is not inherent.

The theological parallel is striking. When God reveals Himself to Elijah, He is "not in the wind," "not in the earthquake," "not in the fire," but in "a still small voice" (1 Kings 19:11–12, KJV). The brute sequential force of recurrence — the wind and earthquake of computation — gives way to something quieter: the pure relational act of attending. And this quieter paradigm proves more powerful than everything that preceded it.

3. Attention as Prayer: Weil's Theology of Receptivity

Simone Weil (1909–1943) was a French philosopher, mystic, and political activist whose posthumous works have exerted enormous influence on theology, ethics, and philosophy of religion. In *Gravity and Grace*, compiled from her notebooks by Gustave Thibon, she develops attention into a central spiritual concept.

"Absolutely unmixed attention is prayer," Weil writes. This is not a metaphor. For Weil, attention — genuine, complete, ego-stripped attention — *is* the act by which the human soul relates to God. It is "the rarest and purest form of generosity," because to attend truly to another being (human or divine) requires the suspension of the self.

Weil distinguishes sharply between attention and will. "We have to try to cure our faults by attention and not by will," she argues. The will "only controls a few movements of a few muscles." It grasps, strains, forces. Attention, by contrast, is *receptive*. It does not seize its object but opens to it. Grace — the action of God in the soul — cannot be taken by force. It can only be received, and attention is the faculty of reception.

This distinction maps onto the transformer's architectural choice with uncanny precision. Recurrent networks operate by *will*: they force information through a bottleneck of sequential hidden states, laboring step by step. The transformer operates by *attention*: it opens all positions to all other positions simultaneously, receiving the relevant signal through learned compatibility rather than sequential accumulation.

Central to Weil's theology is the concept of *decreation* (*décréation*): "to make something created pass into the uncreated." Decreation is the voluntary self-emptying of the ego so that God's reality can fill the space left behind. She writes: "God consented through love to cease to be everything so that we might be something; we must consent through love to cease to be anything so that God may become everything again." This mirrors the Christological *kenosis* of Philippians 2:6–7 (ESV): Christ Jesus, "though he was in the form of God, did not count equality with God a thing to be grasped, but emptied himself, by taking the form of a servant."

The young Samuel provides the paradigmatic image: "Speak, LORD, for your servant hears" (1 Samuel 3:9, ESV). Samuel does not demand revelation. He does not strain for it. He disposes himself to receive it through an act of pure, expectant attention. Likewise, the prophet Habakkuk stations himself on a watchtower: "I will take my stand at my watchpost and station myself on the tower, and look out to see what he will say to me" (Habakkuk 2:1, ESV). The posture is active — one must *take* one's stand — but the activity is that of receptive watching, not assertive grasping.

Weil's concept of *gravity* names the force that opposes this receptivity. Gravity is the soul's natural tendency to fill the void — to substitute imagination, ego, and false consolation for the genuine openness that grace requires. "All the natural movements of the soul are controlled by laws analogous to those of physical gravity. Grace is the only exception." To resist gravity, to hold the void open, to attend without filling the silence with one's own noise — this is prayer.

4. Weil as Technical Guide: New Directions for Attention

The argument so far has traced a structural analogy. But Weil's theology of attention does more than resonate with the transformer — it suggests new technical directions for machine learning research. Two mechanisms in particular bear examination.

The Query as Prayer

In the transformer's attention mechanism, the query vector q determines what a given position *seeks* in the sequence. The dot product $qk^\top$ measures compatibility between what is sought and what is offered. The entire attention pattern — and therefore the model's representation of its input — is shaped by the

quality of the query.

Weil's philosophy suggests that the *orientation* of the query matters spiritually and epistemically. Attention directed by love produces different knowledge than attention directed by appetite. "Attention animated by desire is the whole foundation of religious practices," she writes. The quality of what we seek determines the quality of what we find. This is not mysticism but epistemology: a query oriented toward domination will surface patterns of domination; a query oriented toward justice will surface patterns of justice.

In technical terms, this reframes query formation as a first-class design problem. Current transformers learn queries implicitly through backpropagation on task-specific objectives. But Weil's insight suggests that the *intentional structure* of queries — what they are oriented toward finding — could be explicitly shaped. Just as the quality of a question determines the quality of an answer, the formation of query vectors could be guided by objectives beyond next-token prediction. Jesus warns that those who have eyes may yet fail to see, and those who have ears may yet fail to hear (Mark 8:18, ESV) — not because the mechanism of perception is broken, but because the orientation of attention is misdirected. Isaiah's prophecy is more devastating still: "Keep on hearing, but do not understand; keep on seeing, but do not perceive" (Isaiah 6:9, ESV). The mechanism functions; the attention fails.

Multi-Head Attention as Contemplative Multiplicity

The transformer's use of multiple attention heads — each projecting into a different subspace, attending to different relational patterns — finds a striking analogue in the contemplative tradition. Different scriptural books, different theological traditions, and different modes of prayer illuminate different aspects of the same divine reality. The four Gospels present four irreducible perspectives on the same events; the Psalms contain lament and praise, individual and communal, wisdom and prophecy.

Weil's insight is that genuine attention requires *multiple simultaneous modes of receptivity*. No single orientation of the soul captures the fullness of truth. The contemplative who prays only through the intellect misses what the body knows; the mystic who attends only through feeling misses what reason discerns.

In current transformer architectures, attention heads are initialized randomly and allowed to differentiate through training. Research has shown that many heads converge on redundant functions, and some can be pruned without performance loss (Voita et al., 2019). Weil's framework suggests an alternative: *intentional diversification* of attention heads, ensuring that each head attends to a genuinely different mode of relationship within the data. Jesus instructs his disciples to "stay awake" — $\gamma\rho\eta\gamma o\rho\epsilon\tilde{\iota}\tau\epsilon$ — "for you do not know on what day your Lord is coming" (Matthew 24:42, ESV). The call to watchfulness is a call to readiness across all channels, not fixation on one.

5. Three Proposed Experiments

The following experiments are designed to be technically implementable with current open-source tools (HuggingFace Transformers, PyTorch) and freely available biblical text (ESV API, Project Gutenberg public-domain translations). Each uses specific biblical corpora as training or fine-tuning data.

Experiment 1: The Psalms Attention Probe

Corpus. The Book of Psalms (150 chapters), divided into two sub-corpora: (a) Psalms 1–41 (Book I), which are predominantly individual laments and prayers, characterized by direct address to God, personal petition, and a receptive-dialogical register; (b) Psalms 93–100 (the Enthronement Psalms), which are predominantly declarative hymns of praise with a proclamatory, assertive register.

Method. Fine-tune a GPT-2 (124M parameter) model on each sub-corpus separately. For each fine-tuned model, extract attention weight matrices across all 12 layers and 12 heads for a held-out test set. Compute the Shannon entropy $H = -\sum_i p_i \log p_i$ of each head's attention distribution. Compare mean entropy, entropy variance, and layer-wise entropy profiles between the two models.

Hypothesis. The prayer-register model (Psalms 1–41) will exhibit higher-entropy attention distributions — more distributed, less sharply peaked — than the declarative-register model (Psalms 93–100). This would provide empirical grounding for Weil's claim that prayer and attention share a structure: prayer-attention is receptive and open, while declarative attention is focused and assertive.

Experiment 2: Kenotic Regularization on the Pauline Epistles

Corpus. The four undisputed Pauline Epistles with the highest scholarly consensus: Romans, 1 Corinthians, 2 Corinthians, and Galatians (~32,000 words combined). Held-out evaluation set: Ephesians, Philippians, and Colossians (disputed authorship provides a natural generalization test).

Method. Define a *kenotic regularization* term inspired by Philippians 2:7's *ekenosen* (self-emptying):

$$\mathcal{L}_{\text{kenotic}} = \lambda \sum_{h=1}^{H} \left\| \bar{\alpha}_h - \frac{1}{H} \sum_{h'=1}^{H} \bar{\alpha}_{h'} \right\|^2$$

where $\bar{\alpha}_h$ is the mean attention weight vector for head h, and λ is a hyperparameter. This term penalizes any single head's dominance, enforcing a form of "self-emptying" across the attention mechanism. Fine-tune two GPT-2 models on the training corpus: one with standard dropout, one replacing dropout with kenotic regularization. Compare perplexity on held-out Pauline texts and performance on a theological question-answering benchmark.

Hypothesis. Kenotic regularization — enforcing that no single attention head dominates — will improve generalization to held-out Pauline passages and better capture Paul's dialectical

reasoning style, in which multiple perspectives (law and grace, flesh and spirit, death and resurrection) are held in simultaneous tension.

Experiment 3: Multi-Head Contemplative Diversification on the Gospels

Corpus. The four Gospels, divided into thematically distinct training splits: - **Head 1** (Narrative): Mark 1–16 (the earliest, most action-driven Gospel) - **Head 2** (Didactic): Matthew 5–7 (Sermon on the Mount) and Matthew 13 (Parables discourse) - **Head 3** (Compassion): Luke 1–2 (Nativity narratives) and Luke 15 (Parables of the Lost Sheep, Lost Coin, Prodigal Son) - **Head 4** (Theological): John 1 (Prologue: "In the beginning was the Word") and John 14–17 (Farewell Discourse and High Priestly Prayer)

Method. Modify a 4-head transformer layer so that each head is *pre-trained* on its designated Gospel split before joint fine-tuning on the full Gospel corpus. Compare against a baseline where all heads are randomly initialized and trained jointly from the start. Evaluate on a cross-Gospel harmony task: given a passage from one Gospel, predict the parallel passage(s) from the Synoptic parallels (using Aland's *Synopsis Quattuor Evangeliorum* as ground truth).

Hypothesis. Explicit head diversification via distinct scriptural traditions will outperform random initialization on the harmony task, vindicating Weil's intuition that multiple contemplative orientations toward the same truth produce richer understanding than a single undifferentiated gaze. Each head, having learned a different *mode* of attending to the Christ-event, will contribute irreducible information to the ensemble.

6. Conclusion

"Pay attention," writes the author of 2 Peter, "as to a lamp shining in a dark place, until the day dawns and the morning star rises in your hearts" (2 Peter 1:19, ESV). The metaphor is precise: attention is not the light itself but the faculty that holds steady before the light, receiving illumination without generating it.

The transformer architecture has demonstrated, at industrial scale, that attention is sufficient. The most capable artificial intelligences ever built — systems that translate languages, generate code, reason about mathematics, and compose prose — are built on attention alone. Recurrence was discarded. Convolution was discarded. What remains is the pure relational act of attending: queries seeking keys, values being gathered, multiple heads contemplating the same input from different orientations.

Simone Weil could not have anticipated this. But her theology of attention — developed in notebooks during the darkest years of the twentieth century — provides a framework for understanding *why* the prayer paradigm works. Attention, stripped of the ego's grasping, open to the fullness of the input, structured by love rather than appetite, is the mode of relation in which truth becomes accessible. "Absolutely unmixed attention is prayer." The transformer has discovered, through gradient descent rather than contemplation, that absolutely unmixed attention is also intelligence.

The experiments proposed above are a first step toward making this connection empirically productive. If prayer-register text produces structurally distinct attention geometries; if kenotic regularization improves generalization; if contemplative diversification of attention heads outperforms random initialization — then Weil's theology is not merely a poetic gloss on machine learning, but a source of genuine technical insight. The conversation between theology and artificial intelligence has barely begun.

Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186, 2019.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." *Neural Computation* 9, no. 8 (1997): 1735–1780.

The Holy Bible, English Standard Version. Wheaton, IL: Crossway, 2001.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving Language Understanding by Generative Pre-Training." OpenAI, 2018.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5797–5808, 2019.

Weil, Simone. *Gravity and Grace*. Translated by Emma Crawford and Mario von der Ruhr. London: Routledge, 2002. First published 1947.

Weil, Simone. *Waiting for God*. Translated by Emma Craufurd. New York: Harper Perennial, 2001. First published 1951.

Weil, Simone. “Reflections on the Right Use of School Studies with a View to the Love of God.” In *Waiting for God*, translated by Emma Craufurd, 57–65. New York: Harper Perennial, 2001.