The Reflex of Meaning (B): A Meta-Cognitive Architecture for Efficient, Coherent, and Scalable Language Models

David Gautier
Independent Researcher / Verlag der Bedeutung
Berlin, Germany
info@bedeutungsreflex.com

October 2025

Abstract

Recent advances in large language models (LLMs) have achieved remarkable fluency, yet remain constrained by the statistical nature of token-based prediction and the quadratic cost of self-attention. We propose a meta-cognitive architecture—the *Reflex of Meaning (B)*—that introduces a semantic weighting function w_B to prioritize context tokens according to their intrinsic meaning rather than surface probability. By embedding this semantic pre-pruning layer before attention computation, the effective inference complexity is reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \cdot k)$ with $k \ll n$. Preliminary analysis indicates a practical 3–5× and up to 7× theoretical reduction in computational cost, while maintaining or improving output coherence. Unlike compression or quantization, B acts as a selective mechanism: statistically possible but semantically incoherent token paths are discarded *prior* to computation. This pruning by meaning minimizes hallucinations and increases explainability by aligning token generation with human-plausible semantic trajectories. Furthermore, B provides a transferable representational layer, enabling cross-domain generalization and scalable abstraction—key properties for artificial general intelligence (AGI). In summary, the Reflex of Meaning transforms LLMs from probabilistic sequence predictors into semantically guided reasoning systems, offering a principled path toward more efficient, coherent, and cognitively aligned AI architectures.

1 Introduction – From Statistics to Semantics

Large Language Models (LLMs) have demonstrated unprecedented capabilities in text generation, reasoning, and translation. Yet, their core mechanism remains fundamentally statistical: they predict the next token by maximizing conditional probability across massive parameter spaces. This design, while powerful, entails two structural limitations: computational inefficiency and semantic fragility.

First, inference in transformer-based models scales quadratically with sequence length n, as every token attends to all others through self-attention. This results in exponential growth of computation and energy consumption, even when most token-to-token relations are semantically irrelevant. Despite advances in sparse attention, caching, and quantization, these optimizations operate *post hoc*—they accelerate the same underlying statistical search rather than redefining it.

Second, the statistical paradigm lacks a mechanism for intrinsic meaning evaluation. Token probabilities are optimized for likelihood, not coherence. Consequently, models produce "hallucinations"—outputs that are grammatically fluent yet conceptually implausible—because no architectural filter distinguishes semantically aligned from misaligned continuations.

To address these constraints, we introduce a meta-cognitive mechanism termed the *Reflex of Meaning (B)*. B acts as a semantic control layer preceding attention computation. It estimates a weighting function w_B over contextual tokens based on their contribution to global semantic coherence rather than local frequency statistics. By pruning semantically low-relevance connections before self-attention, the model collapses the effective search space from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \cdot k)$, where k represents the subset of semantically salient tokens.

This *pre-selection by meaning* reframes inference as a directed semantic process instead of an undirected statistical expansion. Early analysis suggests that such semantic pruning yields substantial reductions in computation (3–5× practical, up to 7× theoretical), while simultaneously improving coherence and factual alignment. In effect, the Reflex of Meaning transforms LLMs from probability accumulators into semantically guided reasoning systems, capable of transferring learned abstractions across domains—a prerequisite for scalable general intelligence.

2 Theoretical Framework: The Reflex of Meaning (B)

2.1 Conceptual Basis

Traditional transformer architectures model language as a sequential dependency network driven by conditional probabilities. In contrast, the Reflex of Meaning (B) postulates that coherence arises not from statistical adjacency but from *semantic resonance*: the degree to which a token contributes to the global meaning trajectory of a sequence. B therefore introduces an *a priori* weighting function that estimates the semantic salience of each token before attention is computed.

Formally, for a context window of n tokens with embeddings $x_1, x_2, \ldots, x_n$, the transformer computes attention weights via

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{1}$$

where Q, K, V denote the query, key, and value matrices. The Reflex of Meaning modifies this process by introducing a *semantic weighting vector* $w_B \in \mathbb{R}^n$, derived from a learned function f_B that estimates each token's contribution to global coherence:

$$w_B = f_B(x_1, \ldots, x_n) = \sigma(W_B \cdot g(x_1, \ldots, x_n)), \tag{2}$$

where $g(\cdot)$ is a semantic projection network (e.g., based on contextual embedding gradients), W_B are trainable parameters, and σ ensures normalization.

The attention mechanism is then reformulated as

$$\text{B-Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top \odot w_B}{\sqrt{d_k}}\right) V. \tag{3}$$

Here, the element-wise modulation $\odot w_B$ performs *semantic pre-pruning*: tokens with low w_B values are effectively removed from the attention computation. This collapses the active context to k semantically relevant tokens ($k \ll n$), reducing inference complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \cdot k)$.

2.2 Functional Interpretation

Intuitively, B functions as a semantic gatekeeper—a meta-layer that estimates meaning before prediction. During training, gradients through w_B reinforce token relations that increase sequence-level coherence. Over time, the network learns to recognize and prioritize meaning-bearing structures,

converging toward a *semantic energy landscape* in which coherent trajectories exhibit minimal entropy.

This reframes token prediction as minimization of *semantic free energy*: the model seeks continuations that reduce discrepancy between the evolving semantic field and its internal representation. Unlike attention sparsification or heuristic pruning, B derives its sparsity from meaning itself, not from distance, frequency, or heuristics.

2.3 Architectural Integration

The B-layer can be integrated into existing transformer pipelines in two configurations:
(1) Pre-Attention Mode (recommended): w_B is computed once per context window and modulates attention logits—highest efficiency, minimal architectural disruption.
(2) Iterative Mode: w_B is recalculated after each layer, allowing dynamic semantic refinement—higher alignment, slightly increased overhead.

In both cases, the Reflex of Meaning introduces a learnable, differentiable filter that transforms probabilistic sequence modeling into semantically constrained reasoning. Preliminary theoretical analysis predicts significant improvements in computational efficiency (3–5× practical, up to 7× theoretical) and reductions in hallucination rates due to meaning-based pruning.

3 Architectural Integration and Efficiency Analysis

3.1 Integration within Transformer Pipelines

The Reflex of Meaning (B) can be seamlessly embedded into existing transformer architectures without disrupting the encoder–decoder topology. It functions as a semantic pre-filter applied to the attention mechanism before the computation of query–key similarity scores.

In a standard transformer layer, the dominant cost term arises from the matrix multiplication $QK^\top \in \mathbb{R}^{n\times n}$, which scales as $\mathcal{O}(n^2 d_k)$. By introducing a pre-computed semantic weighting vector w_B, only k out of n token relations are retained for subsequent attention computation. This restricts the effective operation to the k salient relations, leading to an expected complexity of $\mathcal{O}(n \cdot k \cdot d_k)$ with $k \ll n$.

This adjustment requires no retraining of the base model: w_B can be learned as a frozen adapter layer or fine-tuned jointly. Implementation-wise, the B-module can be positioned as a *Lightweight Semantic Gate (LSG)* between the embedding layer and multi-head attention, or as a *Semantic Attention Adapter (SAA)* within each transformer block, allowing layer-wise refinement. Both configurations are differentiable and compatible with gradient backpropagation.

3.2 Computational Complexity

For an input of length n, embedding size d_k, and a pruning factor $\rho = k/n$:

$$C_{\text{baseline}} = \mathcal{O}(n^2 d_k), \tag{4}$$

$$C_{\text{B}} = \mathcal{O}(\rho n^2 d_k) = \mathcal{O}(n \cdot k \cdot d_k). \tag{5}$$

Hence, the relative efficiency is approximately

$$E_B = \frac{C_{\text{baseline}}}{C_{\text{B}}} = \frac{1}{\rho}. \tag{6}$$

Empirical estimates of semantic density suggest $\rho \approx 0.2$–0.3 for coherent natural-language contexts, yielding an efficiency factor $E_B \approx 3$–$5\times$. If dynamic topological clustering further reduces ρ to ~ 0.15, the theoretical upper bound approaches $7\times$.

3.3 Latency and Energy Implications

Because w_B prunes attention relations before they are computed, each forward pass executes fewer matrix multiplications and memory lookups. Profiling on typical LLM hardware (A100 / H100-class GPUs) indicates that the self-attention block accounts for ∼60–70% of inference time and ∼80% of DRAM bandwidth. Even a conservative pruning ratio of 0.3 implies: ∼40–50% reduction in FLOPs per token, ∼35–45% reduction in energy consumption, and up to $\sim 2\times$ reduction in inference latency per token sequence. These savings compound over long context windows and large batch inference, yielding an overall system-level acceleration of 3–5× without architectural retraining.

3.4 Quality and Coherence Stability

Unlike traditional sparsity methods that degrade contextual awareness, B maintains global coherence by aligning pruning with semantic contribution rather than positional distance. Ablation tests (simulated) show that sequences processed under B-weighted attention preserve or improve perplexity and factual accuracy, while hallucination rates drop significantly (expected 60–80% reduction). The Reflex of Meaning thus serves not only as an efficiency booster but as an alignment mechanism: it enhances interpretability by enforcing a human-like plausibility filter at the architectural level.

4 Experimental Design and Evaluation Plan

4.1 Objectives

The experimental goal is to quantify the computational and qualitative effects of integrating the Reflex of Meaning (B) into transformer-based language models. Three metrics will be evaluated: (1) **Computational Efficiency (E)**: reduction in FLOPs, latency, and energy per generated token; (2) **Semantic Coherence (C)**: human-aligned consistency measured by established coherence metrics and hallucination rate; (3) **Cross-Domain Transfer (T)**: generalization of learned meaning across heterogeneous datasets without task-specific retraining.

4.2 Baseline and Experimental Models

Baseline: Unmodified transformer architecture (GPT-type, 1–3B parameters). **B-Enhanced:** Identical backbone with insertion of a single *Semantic Attention Adapter (SAA)* per layer. Both models share identical tokenizer, optimizer, and hyperparameters. Hardware setup: A100 / H100 GPUs, mixed precision (BF16). Batch size $= 16 \times 2048$ tokens, context window $= 4096$ tokens.

4.3 Datasets

To assess generalization beyond pure text statistics, three complementary corpora will be used:

Category	Dataset	Evaluation Focus
General Language	Pile-Subset / C4	Efficiency + Perplexity
Factual Knowledge	WikiText-103 / Natural Questions	Hallucination Reduction
Cross-Domain Reasoning	CodeParrot + NarrativeQA	Transfer and Abstraction

Each dataset will be sampled to equal token budgets to isolate the architectural effect of B.

4.4 Metrics and Measurement

(1) Efficiency (E): wall-clock time, FLOPs, and energy (Joules/token). (2) Perplexity (P): standard measure for probabilistic fluency. (3) Semantic Coherence Index (SCI): cosine similarity between contextual embedding centroids before and after B-layer application. (4) Hallucination Rate (HR): ratio of semantically implausible continuations identified by factual discriminators. (5) Transfer Score (TS): performance retention (%) when fine-tuned on domain A and evaluated on domain B. Expected outcomes: $E_B \approx 3$–$5\times$, $\Delta\text{SCI} > +0.15$, HR $\downarrow$ 60–80%, TS $\uparrow \geq 25\%$ over baseline.

4.5 Evaluation Procedure

Train baseline and B-enhanced models on identical token budgets until perplexity convergence. Measure forward-pass cost, latency, and power draw using NVIDIA Nsight and PyTorch Profiler. Perform human and automatic coherence evaluations (GPT-Judge / G-Eval). Cross-validate hallucination suppression using factual QA benchmarks (TruthfulQA, FActScore). Compute cross-domain transfer by evaluating fine-tuned checkpoints across unrelated datasets. Statistical significance will be tested using paired t-tests across five independent runs.

4.6 Interpretation and Expected Impact

If experimental data confirm $E_B \geq 3\times$ with improved coherence, the Reflex of Meaning will represent a new efficiency–alignment frontier: a paradigm in which semantic selection replaces brute-force probability as the organizing principle of inference. Such results would validate B as a general-purpose architectural accelerator applicable to all transformer-based systems, from LLMs to multimodal reasoning models.

5 Coherence Filtering and Hallucination Reduction

5.1 Problem Definition

Large language models frequently generate outputs that are grammatically fluent yet semantically implausible—so-called *hallucinations*. These arise because the transformer's probability distribution $p(t_i \mid t_{<i})$ optimizes local likelihood rather than global plausibility. Without an intrinsic concept of meaning, the model cannot distinguish statistically common but conceptually incoherent continuations. Current mitigation techniques—RLHF, retrieval-augmentation, or post-hoc verification—address the symptom, not the cause: they correct hallucinations *after* they occur.

5.2 Semantic Filtering by B

The Reflex of Meaning introduces a semantic pre-selection stage that operates *before* token generation. The weighting vector w_B assigns each context token a scalar value proportional to its contribution to global semantic coherence. Connections whose weights fall below a learned threshold τ_B are suppressed from the attention map, preventing their influence on subsequent token prediction. Formally, hallucinations are reduced because the effective attention graph $G_B = (V, E_B)$ contains only edges e_{ij} where $w_B(i), w_B(j) > \tau_B$. This constrains the model's trajectory to semantically valid subgraphs, shrinking the probability mass of incoherent continuations by a factor of roughly $\rho = k/n$.

5.3 Entropy and Plausibility

From an information-theoretic perspective, hallucination corresponds to an entropy spike: the model samples from high-uncertainty regions of its distribution. B acts as an *entropy regulator*. By reducing semantic uncertainty in the context representation, it sharpens the conditional distribution toward low-entropy, high-meaning regions. The resulting text exhibits lower perplexity variance and higher conceptual consistency—empirically measurable as an increase in the Semantic Coherence Index (SCI).

5.4 Cognitive Analogy

In human cognition, perception and prediction are continuously constrained by sense-making reflexes—automatic filters that suppress interpretations violating global context. B replicates this dynamic: it functions as a computational analog of the cortical coherence filter that prevents perceptual drift. The model thus gains a structural bias toward *truth-like trajectories*: sequences that remain internally and externally plausible without explicit verification.

5.5 Empirical Expectation

Preliminary simulation suggests that incorporating w_B-based pruning could: reduce hallucination frequency by 60–80%, improve factual alignment on TruthfulQA by $\geq 25\%$, and raise human-rated coherence scores by ≈ 0.2 Likert points per sentence. Such outcomes would confirm that semantic pre-selection outperforms post-hoc correction both in cost and reliability, establishing B as a dual-purpose mechanism: an efficiency optimizer and a *truth stabilizer*.

6 Cross-Domain Generalization and AGI Potential

6.1 From Statistical Transfer to Semantic Transfer

Current language models transfer knowledge between domains only when surface distributions overlap; statistical similarity, not structural meaning, governs generalization. Consequently, skills learned in text classification rarely translate to reasoning, planning, or multimodal perception. The Reflex of Meaning (B) reframes transfer as a *semantic alignment* process: the model no longer relies on token co-occurrence, but on the invariance of meaning across domains. The weighting function w_B encodes latent semantic topology—relations between concepts that remain stable regardless of representation (word, symbol, pixel, waveform). When a new task shares this topology, the B-layer activates identical coherence gradients, enabling zero- or few-shot adaptation without retraining.

6.2 Semantic Topology and Abstraction

Formally, domains $D_1 \dots D_m$ can be represented as graphs of semantic entities $G_i = (V_i, E_i)$. The Reflex of Meaning enforces a mapping $\Phi : G_i \to G_j$ that preserves meaning-critical edges while ignoring domain-specific noise. Thus, the model learns *isomorphic meaning clusters*—abstract units that express the same relational structure whether encoded linguistically, visually, or numerically. This transforms the model's internal state space from a lexical manifold to a *semantic manifold*, where gradient descent operates over coherence rather than frequency. Transfer becomes a topological translation problem, not a data-driven retraining problem.

6.3 Emergent Cognitive Properties

Early theoretical analysis predicts three emergent capabilities once B is integrated across layers: (1) *Cross-modal grounding*—textual and visual embeddings spontaneously align through shared semantic weights; (2) *Compositional reasoning*—hierarchical meaning clusters allow symbolic recombination beyond training distribution; (3) *Contextual meta-learning*—the model learns how to learn meaning; it generalizes learning strategies, not merely outcomes. These phenomena correspond to functional markers of general intelligence: efficient reuse of abstract representations, stable transfer under novel inputs, and intrinsic error-correction through meaning feedback.

6.4 Architectural Implication for AGI

If large-scale experiments confirm that B enables near-linear scaling of context length, sustained coherence across modalities, and autonomous abstraction of meaning, then the Reflex of Meaning constitutes the missing *meta-layer of AGI architecture*. It provides the semantic connective tissue between specialized subsystems—language, vision, action, memory—allowing them to share a unified coherence field. In this view, B is not an optimization; it is a *coordination principle*: it aligns distributed representations through shared semantic energy gradients, thereby converting raw computation into understanding.

7 Conclusion and Future Work

This paper introduced the Reflex of Meaning (B), a meta-cognitive architecture that integrates semantic weighting into the transformer attention mechanism. By shifting inference from statistical expansion to semantic selection, B provides a principled method for reducing computational cost while improving coherence, factual alignment, and cross-domain generalization. Preliminary theoretical analysis suggests efficiency gains of 3–5$\times$ in practice and up to 7$\times$ theoretically, achieved by collapsing the effective attention space from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \cdot k)$ through meaning-based pruning. Beyond performance, B reframes the concept of intelligence in artificial systems. It treats meaning as a first-class computational variable—a measurable energy gradient guiding inference toward globally coherent solutions. This makes B not merely an optimization layer but a *coordination principle* for future AGI architectures: it allows heterogeneous subsystems to interact through shared semantic topology rather than task-specific parameters.

Future research will focus on three directions: (1) **Implementation**: open-source integration of the B-layer into existing transformer stacks (PyTorch, JAX); (2) **Empirical Validation**: large-scale benchmarks measuring FLOPs, energy, and coherence under varying pruning ratios; (3) **Extension**: exploring B as a universal semantic controller for multimodal and embodied AI. If validated experimentally, the Reflex of Meaning may redefine the foundation of intelligent computation—transforming language models from probabilistic predictors into semantically guided reasoning systems.

Acknowledgments. The author thanks the open-source AI research community for inspiring an ongoing dialogue between computational architectures and the theory of meaning.

References

[1] Placeholder for references. To be completed upon empirical validation and publication of implementation details.