Value Residual Learning

First-layer values as a path through deep Transformers

Paper by Zhanchao Zhou, Tianyi Wu, Zhiyun Jiang,
Fares Obeid, and Zhenzhong Lan

arXiv:2410.17897v5 | Revised 8 June 2025

Research reading presentation | 11 September 2026

Mix first-layer values for quality; share them for a smaller KV cache.

A residual path may not make early information usable

The research question

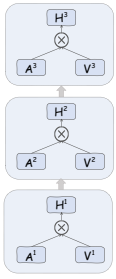
Can deep layers retrieve initial token content without directly mixing it into the hidden stream?

- Hidden states increasingly integrate context; local content can become difficult to recover.
- ResFormer: mix first-layer and current-layer values before attention.
- SVFormer: use first-layer values in every layer; keep layer-specific queries and keys.

Paper, Sec. 1, PDF pp. 1–2. The information-dilution mechanism is the authors' hypothesis.

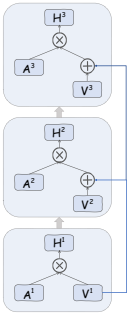
The extra path enters attention through its values

Transformer



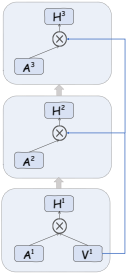
$$\mathbf{U}_n = \mathbf{A}_n \mathbf{V}_n$$

ResFormer



$$\mathbf{U}_n = \mathbf{A}_n \widetilde{\mathbf{V}}_n$$

SVFormer



$$\mathbf{U}_n = \mathbf{A}_n \mathbf{V}_1$$

Original panels from Fig. 1, PDF p. 2; simplified diagrams omit MLPs, normalization, and ordinary hidden residuals.

Current attention selects from both sources of content

$$\mathbf{V}_1 = \mathbf{H}_0\mathbf{W}_1^{\mathbf{V}}, \qquad \mathbf{V}_n = \mathbf{H}_{n-1}\mathbf{W}_n^{\mathbf{V}} \quad (n \geq 2)$$

ResFormer: mixing and attention

$$\widetilde{\mathbf{V}}_n = \lambda_{n,1}\mathbf{V}_1 + \lambda_{n,2}\mathbf{V}_n, \qquad \begin{aligned}\mathbf{U}_n &= \mathbf{A}_n\widetilde{\mathbf{V}}_n \\ &= \lambda_{n,1}\mathbf{A}_n\mathbf{V}_1 + \lambda_{n,2}\mathbf{A}_n\mathbf{V}_n\end{aligned}$$

- $\mathbf{A}_n$: current-layer attention, $l \times l$; values: $l \times d$ in the paper's simplified notation.
- Mix values once, then call attention once. Keep ordinary hidden residuals.
- $\mathbf{V}_1$ is input-dependent token content, not a fixed memory table.

Paper, Eq. (5), Secs. 3.1–3.2, PDF p. 3; mixed value renamed as tilde-V. Normalization and heads are suppressed.

The connection coefficients define a family of models

Variant	First-layer branch	Current-layer branch
Identity	fixed 0.5	fixed 0.5
Learnable	initialized 0.5	initialized 0.5
Constant example	fixed 2	fixed 1
Sparse	active in selected layers	configuration-dependent
SVFormer	fixed 1	removed for $n \geq 2$

- Dense-ResFormer mixes values from multiple earlier layers.
- Plus couples first-layer branch coefficients across depth.
- Ordinary Learnable coefficients are not specified as convex weights.

Paper, Secs. 3.2–3.3, PDF pp. 3–4. Plus details and ambiguities are in the backup slides.

SVFormer saves value storage; every layer still stores keys

$$C_{\text{base}} = 2NBld_{kv}, \qquad C_{\text{SV}} = (N+1)Bld_{kv}$$

$$\frac{C_{\text{SV}}}{C_{\text{base}}} = \frac{N+1}{2N}$$

8 layers	24 layers
56.25% of baseline KV storage	52.08% of baseline KV storage
43.75% saved	47.92% saved

- B: batch; l: cached tokens; d_{kv}: per-layer KV width; N: depth.
- Equal K/V widths and fixed model dimensions; element counts omit metadata.
- ResFormer retains layer-specific V. Cache savings do not imply equal latency savings.

Derivation from Sec. 3.3, PDF pp. 3–4. These counts are not measured memory or throughput results.

The main evidence comes from controlled pretraining

Main protocol

- 82M–468M Llama-like models
- About 20B SlimPajama tokens
- Length 2,048; about 2M tokens/step
- 10,000 steps; AdamW; peak LR 6×10^{-4}
- Pre-norm, SwiGLU, RoPE; FP16

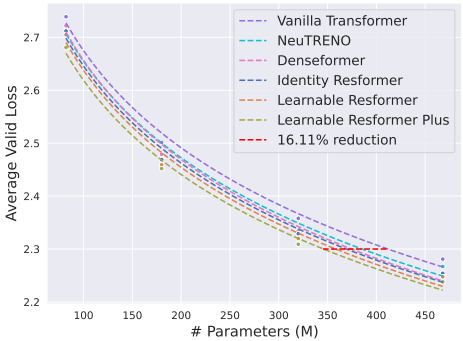
What is measured?

- Seven-domain average validation loss $\downarrow$
- Zero-shot task accuracy $\uparrow$
- WikiText/LAMBADA PPL $\downarrow$
- Ablations and representation diagnostics

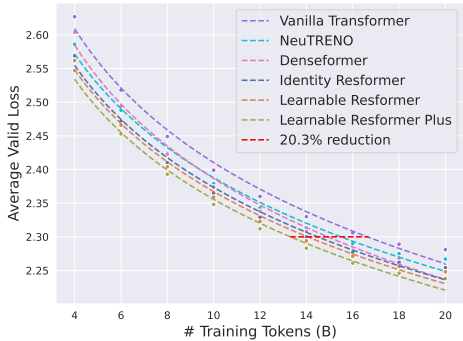
Reproduction caveat: warmup is 1,200 steps in Sec. 4.1 but 120 in Table 14. Repeated-run uncertainty is not reported.

Paper, Sec. 4.1, PDF pp. 4–5; Tables 12–14, pp. 15–16. Larger-scale protocol is separate.

ResFormer reaches a given loss with fewer resources



16.11% fewer parameters
82M–468M; 20B tokens

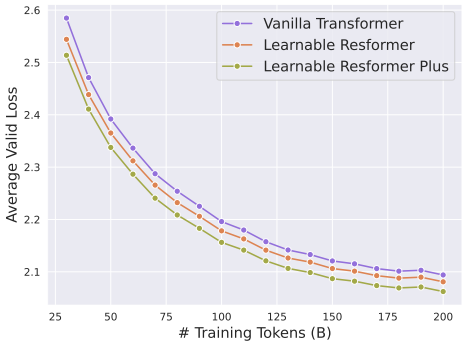


20.3% fewer training tokens
468M; evaluated every 2B tokens

Separate equal-loss estimates; the two reductions were not tested jointly.

Original Fig. 2 left/middle and Sec. 4.2, PDF pp. 4–5. Fit uncertainty is not reported.

A 1.6B run extends the evidence under a different protocol



- About 200B internal pretraining tokens
- Sequence length 8,192
- Learning rate 1.5×10^{-4}
- About 3,584 H800 GPU hours per run

Scope: lower validation loss here; no final 1.6B downstream table in v5.

Original Fig. 2 right; Secs. 4.1–4.2, PDF pp. 4–5. Commented-out source tables are excluded.

Mean accuracy improves, but no variant wins every metric

468M / 20B tokens	Wiki PPL ↓	LAMBADA PPL ↓	Mean ACC ↑
Transformer	24.8	33.4	40.6%
NeuTRENO	24.3	36.3	41.4%
DenseFormer	24.0	**28.0**	40.8%
Identity ResFormer	23.8	32.9	41.3%
Learnable ResFormer	23.7	32.5	**42.3%**
Learnable ResFormer Plus	**23.2**	31.4	42.0%

- Learnable: +1.7 percentage points in the reported eight-task mean.
- Plus has the best WikiText PPL; DenseFormer has the best LAMBADA PPL.

Selected columns from Table 2, PDF p. 4. Mean ACC excludes the two PPL columns; error bars are not provided.

Simple mixing works; learned weights are not automatically best

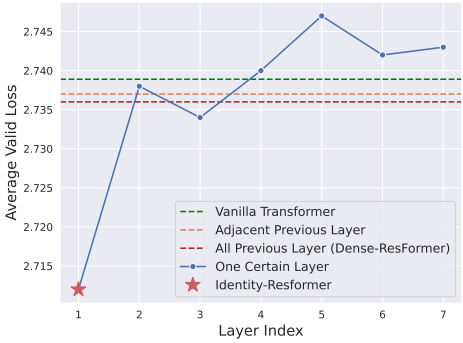
8-layer, 82M model	Loss ↓
Transformer	2.739
DenseFormer	2.722
NeuTRENO	2.720
Identity ResFormer	2.712
Learnable ResFormer	2.705
Constant: $2\mathbf{V}_1 + \mathbf{V}_n$	2.700
Sparse: $5\mathbf{V}_1 + \mathbf{V}_n$, layers 6–8	2.687
Learnable ResFormer Plus	**2.681**

- NeuTRENO adds a value difference after attention.
- DenseFormer mixes hidden states across depth.
- Fixed mixing helps; manual search can improve it further.

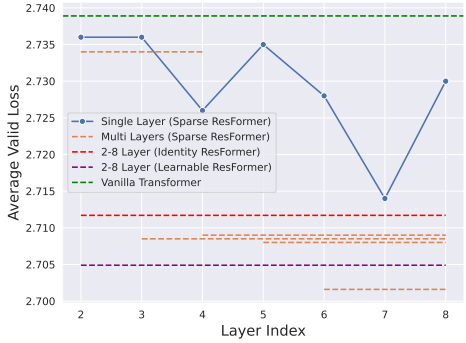
Caveat: manual settings were searched; tuning budgets are unmatched.

Table 3 and Sec. 4.2, PDF p. 5; baseline definitions in Sec. 3.1, p. 3. Full variant table is in the Chinese notes.

First-layer values help the last few layers most



Source: $\mathbf{V}_1$ is most useful.



Destination: late layers.

- Best in this 8-layer model: layer 7 alone, or layers 6–8 together.
- Layer choices depend on architecture and initialization.

Original Fig. 5 and Sec. 4.4, PDF p. 6. Lower validation loss is better.

The location and mapping of the shortcut both matter

Where to add a residual?

Position	Loss ↓
None	2.739
Query	2.742
Key	2.746
Attention matrix	2.757
Value	**2.712**

How to map V_1 to the output?

Mapping	Loss ↓
None	2.739
Identity mapping	3.137
Recomputed cross-layer attention	2.729
Current attention	**2.712**

Interpretation: use the current semantic selection pattern to read early content.
Recomputing cross-layer attention also costs more computation.

Tables 6–7, PDF p. 7; explanation in Sec. 4.4, p. 8. Left: fixed 0.5/0.5 mixing.

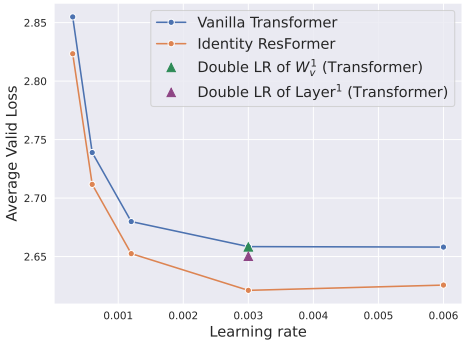
Extra hidden residuals expose an initialization trade-off

Extra shortcut	Coefficient initialization	Loss ↓
None	—	2.7389
Value, from $\mathbf{V}_1$	0.5/0.5	**2.705**
Hidden, from $\mathbf{H}_0$	0.5/0.5	2.781
Value, from $\mathbf{V}_1$	0/1	2.730
Hidden, from $\mathbf{H}_0$	0/1	2.722

- All shortcut coefficients in this table are learnable.
- Large initial hidden mixing hurts, but hidden wins at 0/1 initialization.
- The best tested value setting wins; universal superiority is not established.

Table 4, PDF p. 7. The adjacent prose mixes H0/H1; this slide follows the table and stated experiment.

Simple learning-rate changes do not reproduce the gain

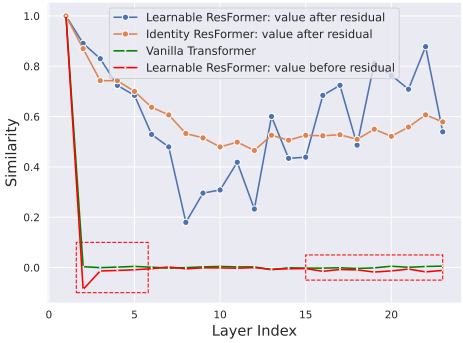


- Both models are best near LR 0.003 in this sweep.
- Identity ResFormer remains better at the tested rates.
- Doubling LR for the first layer or only $\mathbf{W}_1^{\mathbf{V}}$ does not close the gap.

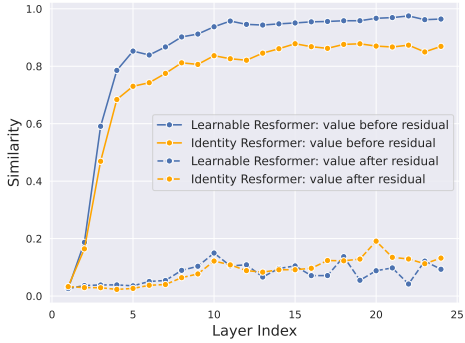
Boundary: this rules out simple explanations, not all optimization effects.

Original Fig. 4 right; Sec. 4.3, PDF pp. 5–6. This is a dedicated LR ablation, not the default protocol.

Value similarities support a residual-learning interpretation



Similarity to first-layer values

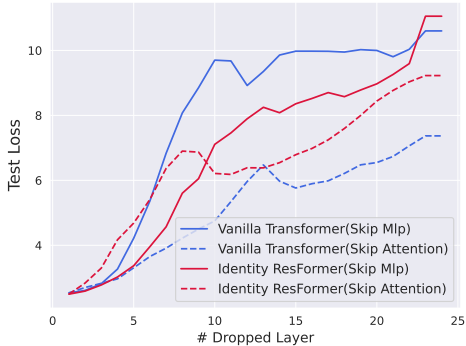


Similarity between tokens within a layer

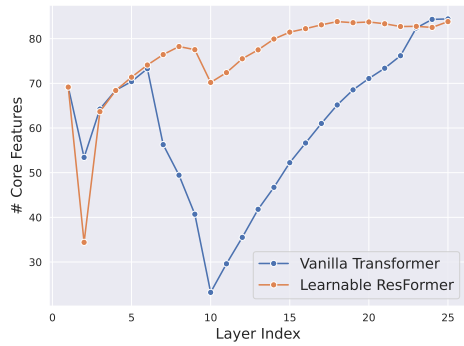
- Mixed values retain early content; raw values can learn a correction.
- Similarity does not directly measure information or residual magnitude.

Original Fig. 8, PDF p. 8. The authors interpret the raw value branch as learning a correction.

Attention matters more; representation variance spreads out



Remove modules; retain layer 1.



PCA dimension at 99% variance.

- ResFormer becomes more sensitive to attention removal.
- More PCA components do not prove higher task-relevant information.

Original Fig. 9 and Sec. 4.5, PDF p. 8. These are post-training diagnostics.

SVFormer trades quality differently at short and long lengths

Method	Training length 2,048	Training length 64,000
Transformer, full KV	**2.739**	2.753
CLA2, about half KV	2.776	2.793
GQA2, half KV	2.748	2.773
SVFormer, about half KV	2.774	**2.7485**

- All entries are validation loss ↓ under the reported budget.
- SVFormer: $+0.035$ vs. baseline at 2,048; -0.0045 at 64,000.
- Training length changes across columns; this is not a context-extrapolation or retrieval test.

Table 8 left, PDF p. 9. CLA2 shares K/V every two layers; GQA2 halves KV heads. No error bars; cache sizes are approximate.

SVFormer can combine with grouped-query attention

Length 64,000	Loss ↓
Transformer, full KV	2.753
GQA8	2.807
CLA2 + GQA4	2.815
SVFormer + GQA4	**2.741**

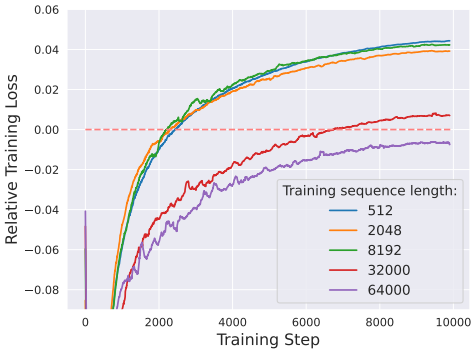
About one-eighth KV storage

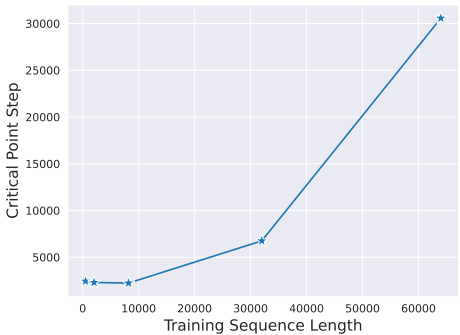
$$\frac{C_{\mathrm{SV+GQA}g}}{C_{\mathrm{MHA}}} = \frac{N+1}{2Ng}$$

- g: KV-head reduction factor.
- At $N = 8, g = 4$: 14.06% remains, not exactly 12.5%.
- Formula assumes fixed depth and widths.

Table 8 right, PDF p. 9. Cache formula is our derivation; per-row measured bytes and throughput are not reported.

The loss crossover is much later at 32k and 64k

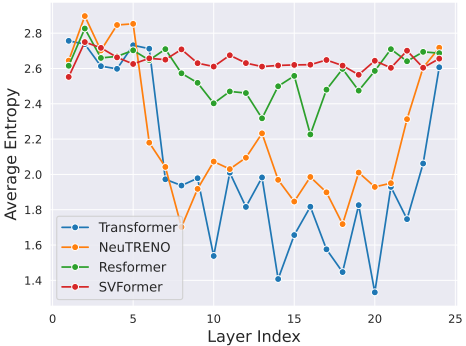




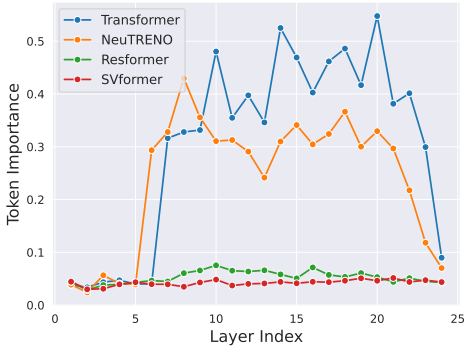
- Left: $L_{\mathrm{SV}} - L_{\mathrm{base}}$; negative favors SVFormer; about 2M tokens/step.
- 64k crossover is extrapolated from the last 1,000 points.
- Smaller learning rates also help; a universal eventual crossover is unproven.

Original Fig. 10, Table 9 and Sec. 4.6, PDF p. 9. The paper's exponential-trend claim has limited support.

Attention sinks weaken, with a caveat in the entropy text



Token-importance entropy across depth



Attention importance of the first token

- ResFormer/SVFormer show less first-token concentration here.
- Entropy $-\sum_i p_i \log p_i$: higher means less concentrated.

Original Fig. 13(a,b), Appendix A.1, PDF p. 12. We correct the appendix prose, which reverses the entropy direction.

The evidence is useful, but the mechanism is not settled

- **Robust core:** several architectural controls support value mixing under current attention.
- **Scale:** most ablations are small; the 1.6B study uses internal data.
- **Optimization:** learned coefficients can lose to manually tuned settings; Plus details are incomplete.
- **Measurement:** no repeated-run uncertainty or end-to-end inference benchmark is reported.
- **Mechanism:** similarity, PCA, and sink reduction are supporting observations, not a complete causal proof.

Replication priorities: resolve warmup/batch settings, match tuning budgets, and test the long-run SVFormer crossover.

Paper, Secs. 4.2–4.6 and Limitations, PDF pp. 5–9; Appendix A.3, pp. 15–16. Assessment by this reading.

Take away a design principle and a conditional trade-off

1. **Separate selection from content.** Deep attention can read first-layer values using its current Q/K relationships.
2. **Mixing improves the reported quality frontier.** The location, depth, and strength of the shortcut matter.
3. **Full sharing reduces storage.** SVFormer's quality depends on training length, budget, and learning rate.

Discussion: Does first-layer content access explain the gains after matching optimization and tuning budgets?

Zhanchao Zhou, Tianyi Wu, Zhiyun Jiang, Fares Obeid, Zhenzhong Lan. Value Residual Learning. arXiv:2410.17897v5, revised 8 June 2025. Source PDF and author TeX were both read.

Backup: Plus couples first-layer coefficients across depth

$$\lambda_{n,1} = \lambda_{\text{scale}} \frac{e^{\lambda'_{n,1}}}{\sum_{j=1}^{N} e^{\lambda'_{j,1}}}, \qquad \lambda_{\text{scale}} \text{ initialized to } N$$

$$\lambda_{n,2} \text{ initialized to } 0.5, \qquad \sum_{n=1}^{N} \lambda_{n,1} = \lambda_{\text{scale}}$$

- Equal initial logits would give $\lambda_{n,1} = 1$, not 0.5.
- The paper does not specify the logits' initialization clearly.
- It is unclear whether this is only initialization or a constraint throughout training.
- The motivation for stronger later-layer connections does not specify a monotone initialization.

Paper, Sec. 3.2, PDF p. 3. Sum and equal-logit implication are our algebraic deductions.

Backup: value shortcuts compensate for missing hidden paths

Hidden residual starts at	Added source	No value skip	With value skip
$\mathbf{H}_0$ (default)	$\mathbf{V}_1$	2.739	2.712
$\mathbf{H}_0$ (default)	$\mathbf{V}_2$	2.739	2.738
$\mathbf{H}_1$	$\mathbf{V}_2$	2.780	2.780
$\mathbf{H}_2$	$\mathbf{V}_2$	2.820	2.787
$\mathbf{H}_2$	$\mathbf{V}_3$	2.820	2.833
$\mathbf{H}_3$	$\mathbf{V}_3$	3.057	2.883

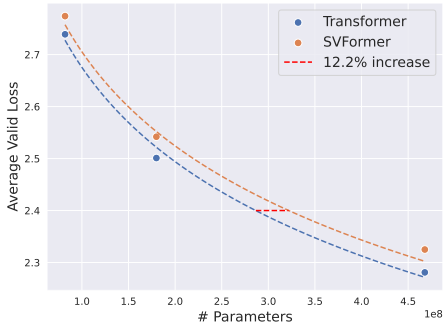
- Loss $\downarrow$. Earlier hidden residuals are removed to change information access.
- $\mathbf{V}_2$ and $\mathbf{V}_3$ can help when the corresponding hidden paths are missing.

Table 5 and Sec. 4.4, PDF p. 7. This is a changed-architecture control, not the default model.

Backup: SVFormer depends on learning rate and model size

Model	LR	Warmup	ΔL
Llama	10^{-4}	120	-0.033
Llama	3×10^{-4}	120	$+0.021$
Llama	6×10^{-4}	120	$+0.035$
Llama	6×10^{-4}	1,200	$+0.036$
GPT2	6×10^{-4}	120	$+0.029$

$\Delta L = L_{\mathrm{SV}} - L_{\mathrm{base}}$; negative is better.



Reported equal-loss cost:
12.2% more parameters.

Table 9 and original Fig. 11, PDF p. 9. Parameter compensation and fixed-architecture KV savings are separate comparisons.

Backup: shallow mixing is more sensitive to scaling

Sparse constant coefficients

λ	Layers 2–4	Layers 5–7
1	2.724	2.698
2	2.728	2.688
5	2.732	2.682

Post-mixing norm rescaling

Layers	No	Yes
All	2.712	2.701
2–4	2.734	2.727
6–8	2.702	2.701

- Validation loss ↓; stronger first-layer mixing helps the later group.
- Left follows Table 10's layers 5–7; the prose says 6–8. Do not silently merge them.
- Norm rescaling helps shallow/all-layer variants more; normalization axes are unspecified.

Tables 10–11 and Appendix A.1, PDF p. 14. Baseline loss: 2.739.

Backup: resolve these details before exact reproduction

- **Warmup:** Sec. 4.1 gives 1,200 steps; Table 14 gives 120.
- **Batch arithmetic:** at length 2,048, Table 13 lists per-GPU batch 32, 8 GPUs, accumulation 32, but total batch 1,024.
- **Coefficient heatmap:** Fig. 6 right is captioned 468M, while text and displayed depth indicate 8 layers.
- **Entropy indexing:** for row-normalized A_{ij}, summing over j is constant; incoming token importance normally sums over query rows.
- **Coverage:** source/PDF read; no training reproduction. Original figures and build helper accompany this deck.

PDF pp. 4, 6, 12, 15–16. Numerical discrepancies also appear in the author TeX; they are not text-extraction artifacts.