Compressed Sensing of LLMs

Query-Efficient Recovery of a Black-Box Next-Token Function

Master's Thesis Proposal

MSc in Computer Science / Machine Learning, Data Science and Artificial Intelligence
30 ECTS · Supervisor: Prof. A. Jung
Department of Computer Science, Aalto University

1 The Abstraction

We model a language model as a single deterministic nonlinear function

$$f_\theta : \mathcal{V}^L \to \mathbb{R}^{\mathcal{V}}, \qquad c \mapsto z = f_\theta(c),$$

mapping a context window c (a token sequence of length up to L) to the logit vector z over the next token, where $\mathcal{V}$ is the vocabulary.

Auditing a closed LLM can be framed as a function recovery problem: Given observed conversations, estimate the underlying function f_θ.

The thesis asks the compressed-sensing question in its purest form: treating f_θ as an unknown high-dimensional signal, how much of it can be recovered from $m \ll |\mathcal{V}|^L$ evaluations, under what structural conditions, and with what query designs?

There are no separate "attacks," only different *targets* — the whole function, a restriction, or a functional of it — recovered through one evaluation operator.

2 Why Compressed Sensing Applies: the Compressibility Priors

A generic function $\mathcal{V}^L \to \mathbb{R}^{\mathcal{V}}$ is unrecoverable from few samples — the domain has size $|\mathcal{V}|^L$. Recovery is possible only because f_θ is *not* generic. Four structural facts play the role that sparsity plays in classical compressed sensing [1–3], and that earlier underpinned prediction-API model stealing [4]:

(P1) Low-rank codomain. The logits factor as $z = W_U\, h(c)$ with hidden state $h(c) \in \mathbb{R}^d$ and un-embedding $W_U \in \mathbb{R}^{\mathcal{V}\times d}$, $d \ll |\mathcal{V}|$ [5]. Hence range(f_θ) lies in a d-dimensional subspace, and since $\mathrm{softmax}(z + c\mathbf{1}) = \mathrm{softmax}(z)$ the next-token distribution identifies z only modulo the constant direction span$\{\mathbf{1}\}$: the effective output dimension is d (at most $d-1$ for the distribution), not $|\mathcal{V}|$ — the softmax bottleneck [6].

(P2) Continuous factorisation. f_θ factors through token embeddings $c \mapsto E(c) \in \mathbb{R}^{L\times d}$, so it is the discrete lift of a smooth map on a continuous, low-dimensional domain rather than an arbitrary function on a combinatorial set.

(P3) Lipschitz smoothness. On the compact embedding domain (bounded token embeddings plus positional encodings), f_θ is a composition of C^1 maps and is therefore Lipschitz, with constant controlled by the product of per-layer operator norms. The restriction to a bounded domain is essential: standard dot-product self-attention is provably *not* Lipschitz on the unbounded domain $\mathbb{R}^{L\times d}$ [7, Thm. 3.1], but on a ball of radius R its local Lipschitz constant scales as $\sqrt{L}\cdot \mathrm{poly}(R, \|W\|)$ [8, Thm. 3.3] (see also 9). Thus nearby contexts produce nearby logits — the condition under which point evaluations are informative (Section 4).

(P4) Small description complexity. For a bounded-weight, fixed architecture the metric entropy of the realizable class grows only polynomially in $1/\varepsilon$ — roughly $\mathcal{H}_\varepsilon(\mathcal{F}) \lesssim W \log(1/\varepsilon)$ with W the parameter count — rather than the $(1/\varepsilon)^D$ entropy of generic Lipschitz functions on $[0,1]^D$, which is exponential in the ambient dimension D [10, 11]. For the transformer class specifically, norm- and rank-based covering-number bounds make the dependence on embedding dimension and context length only logarithmic [12, 13]; networks remain Kolmogorov–Donoho rate-distortion optimal approximants of structured function classes [14].

These four properties are the compressibility priors. The thesis makes precise how each one reduces the number of evaluations needed.

3 Measurement Model

Black-box access furnishes an *evaluation operator* $\Phi : \mathcal{F} \to (\mathbb{R}^{\mathcal{V}})^m$, $\Phi(f) = (f(c_1), \ldots, f(c_m))$, instantiated at several realism levels:

- **Full-logit oracle:** $y_i = f_\theta(c_i) \in \mathbb{R}^{\mathcal{V}}$ (clean linear measurements of the codomain).
- **Top-k log-probability oracle:** only the k largest entries are observed — a *compressed/partial* measurement of each output vector.
- **Sampled-token oracle:** only stochastic draws from $\mathrm{softmax}(z)$ — *noisy, quantised* measurements requiring repetition to denoise.
- **Adaptive logit-bias oracle:** query-time perturbations (Carlini et al. [15]) act as designed measurement vectors that reconstruct individual logit coordinates through a top-k API, after which the codomain subspace is identified by SVD — subspace probing under restricted output access.

A central practical theme is how these restricted oracles degrade the recovery guarantees of the full-logit case — i.e. how API defenses act as perturbations of the measurement operator.

4 Core Theoretical Object: Restricted Isometry over $\mathcal{F}$ via Point Evaluation

Fix a context distribution $\mathcal{D}$ and the population seminorm $\|f\|^2 = \mathbb{E}_{c\sim\mathcal{D}}\|f(c)\|^2$. Draw probe contexts $c_1, \ldots, c_m \sim \mathcal{D}$ and form the empirical evaluation seminorm $\widehat{\|f\|}^2 = \frac{1}{m}\sum_i \|f(c_i)\|^2$. The recovery problem is well-posed precisely when the evaluation operator is a restricted isometry over the difference set $\mathcal{F} - \mathcal{F}$.

Condition 1 (Evaluation-RIP over $\mathcal{F}$)**.** There exists $\delta \in (0,1)$ such that, for all $f, g \in \mathcal{F}$,

$$(1-\delta)\,\|f-g\|^2 \;\le\; \frac{1}{m}\sum_{i=1}^{m} \|f(c_i) - g(c_i)\|^2 \;\le\; (1+\delta)\,\|f-g\|^2.$$

Condition 1 says: any two models agreeing on the m probe contexts must be globally close. When it holds with small δ, evaluations at $\{c_i\}$ form a norm-preserving sketch of the entire function class, and recovery (find $\hat{f} \in \mathcal{F}$ consistent with the measurements) is stable. The thesis develops this object along three established lines.

Sample complexity via metric entropy. The uniform isometry is an empirical-process statement; the required m is governed by the metric entropy $\mathcal{H}_\varepsilon(\mathcal{F})$ of the transformer class. Norm- and rank-based covering-number bounds for self-attention show $\mathcal{H}_\varepsilon(\mathcal{F})$ scales polynomially in $1/\varepsilon$ and only logarithmically (or not at all) in the sequence length [12, 13], so heuristically $m \gtrsim \mathcal{H}_\varepsilon(\mathcal{F})/(1-\delta)$. The closest prior on recovering a *system* from input–output traces in a metric-entropy-optimal manner is Hutter et al. [16].

Well-posedness via smoothness (sampling numbers). Whether point samples suffice — rather than arbitrary linear functionals — is the subject of information-based complexity. Recent results show that whenever the approximation numbers are square-summable (decay rate $s > \frac{1}{2}$), the sampling numbers decay at the same polynomial rate, so point evaluations are essentially as powerful as optimal linear measurements [17–19]; the sharp, logarithm-free form $g_{cm}^2 \lesssim \frac{1}{m}\sum_{k\geq m} a_k^2$ is due to Dolbeault et al. [20]. The Lipschitz prior (P3) is exactly what places $\mathcal{F}$ in this regime, justifying that query access loses little against an idealised measurement operator. Note that these results establish the *rate* of an optimal sampling algorithm; deriving the two-sided isometry of Condition 1 still requires an empirical-process argument over $\mathcal{F}-\mathcal{F}$.

Lower bounds via packing. Applying Fano's inequality to a maximal ε-packing of $\mathcal{F}$ converts metric entropy into a minimax lower bound: the achievable accuracy ε_m is fixed by the entropy-balance $\mathcal{H}_{\varepsilon_m}(\mathcal{F}) \asymp m\,\varepsilon_m^2$ for the noisy (sampled-token) oracle [21], so no recovery scheme — adaptive or not — beats this rate. This pins query complexity to the *intrinsic* complexity of the model class, independent of vocabulary size and context length except through $\mathcal{H}_\varepsilon$.

5 Recoverable Targets

Recovering all of f_θ to high accuracy is infeasible when $\mathcal{H}_\varepsilon(\mathcal{F})$ is large; the tractable regime is recovery of low-complexity *targets* derived from f_θ, all under the single evaluation operator of Section 3.

Target 1 (Codomain structure)**.** The subspace range(W_U) and the hidden dimension d — the latter being identifiable, since W_U itself is recoverable only up to an $\mathcal{V}$-irrelevant $d\times d$ symmetry (affine in practice, orthogonal in principle). This is a functional of f_θ, recoverable as numerical-rank / subspace recovery (SVD of stacked logit vectors) from $O(d)$ full-logit queries; under a restricted top-k/logit-bias API the cost rises to $O(d\,|\mathcal{V}|)$, since each full logit vector must first be reconstructed coordinate-by-coordinate. Carlini et al. [15] is the empirical existence proof and the controlled benchmark.

Target 2 (Restriction to a structured sub-domain)**.** $f_\theta|_{\mathcal{S}}$, where $\mathcal{S}$ is a prompt family varying along $k \ll Ld$ axes (e.g. in-context demonstrations of a function class [22]). Recovery is local nonlinear function approximation with sample complexity controlled by $\mathcal{H}_\varepsilon(\mathcal{F}|_{\mathcal{S}})$, small when $\mathcal{S}$ is low-dimensional. This subsumes "identify the algorithm implemented in-context" as recovery of f_θ on a structured slice.

Target 3 (General functionals)**.** $P(f_\theta)$ — a decision boundary on a sub-domain, the presence of a watermark, a behavioural property. Such a functional is recoverable under a restricted isometry on the relevant quotient of $\mathcal{F}$, typically with far fewer queries than full recovery.

6 Research Questions

RQ1.

When does the point-evaluation operator satisfy a restricted isometry over the transformer-realizable class $\mathcal{F}$, and how do the four compressibility priors (P1–P4) enter the isometry constant?

RQ2.
What is the query complexity m for each target in terms of the metric entropy / sampling numbers of the relevant (restricted or quotiented) class, and does it match the packing lower bound?

RQ3.
How do the restricted oracles of Section 3 — top-k truncation, sampling noise, finite precision, rate limits — degrade the isometry constant, and which defenses provably destroy recoverability versus merely inflating m?

7 Methodology

Theory track. Establish the evaluation-RIP condition for $\mathcal{F}$; import metric-entropy bounds for transformer-type classes and sampling-number results to convert them into query-complexity upper bounds; prove matching packing lower bounds; analyse restricted oracles as perturbations of the isometry.

Empirical track. On open-weight models where f_θ is fully known (Pythia, Qwen, Llama across scales): (i) empirically estimate the evaluation-isometry constant for different probe designs (random vs structured vs adaptively chosen contexts); (ii) recover targets T1–T3 and compare achieved query counts against the theory; (iii) stress-test under simulated top-k and sampling oracles. Deliverables include a reproducible probing/recovery harness and empirical query-complexity scaling curves.

8 Evaluation

- **Recovery accuracy:** subspace/parameter error (T1), function-approximation error on $\mathcal{S}$ (T2), functional error (T3).
- **Query efficiency:** measured m vs predicted complexity; scaling against d, $|\mathcal{V}|$, and the dimensionality of $\mathcal{S}$.
- **Isometry diagnostics:** empirical δ as a function of probe design and budget.
- **Robustness:** degradation across the oracle hierarchy; lower-bound tightness (achieved vs information-theoretic m).

9 Expected Contributions

1. A single compressed-sensing formalism for black-box LLMs: the model as one nonlinear function, query access as point evaluation, recovery governed by a restricted isometry over the function class.
2. Query-complexity upper bounds (via metric entropy and sampling numbers) and matching packing lower bounds for codomain structure, sub-domain restrictions, and functionals.
3. A robustness theory relating realistic API oracles to the isometry constant, with provable separations between benign and recovery-destroying defenses.
4. An open, reproducible harness and empirical scaling laws relating query budget to recoverable structure.

The framework quantifies *how auditable* a deployed model is from the outside, and conversely which deployment choices provably limit external recovery — a direct contribution to trustworthy-AI and model-auditing goals.

10 Risks and Contingencies

- *The point-evaluation operator may resist a clean RIP for the full class.* Mitigation: smoothness places $\mathcal{F}$ in the sampling-numbers regime; where a global RIP is unavailable, restrict attention to targets T1–T3, whose quotient classes have small metric entropy and admit isometries directly.
- *Restricted oracles may block logit access entirely.* Mitigation: the open-weight track is independent of any external API; oracle restrictions are simulated locally with controlled corruption layers.
- *Theory–experiment gap.* Mitigation: T1 (validated by Carlini et al. [15]) guarantees a defensible empirical result even if the bounds for T2/T3 come out only partial.

11 Indicative Timeline (6 months)

Month	Milestone
1	Formalise the function/measurement model; consolidate metric-entropy and sampling-number tools
2	Evaluation-RIP theory; codomain recovery (T1) + benchmark harness
3	T1 empirical validation; oracle-robustness study
4	Sub-domain restriction (T2) and functionals (T3): complexity upper/lower bounds
5	T2/T3 experiments; query-complexity scaling laws
6	Synthesis, writing, defense preparation

12 Prerequisites for the Student

Linear algebra and high-dimensional probability; prior exposure to compressed sensing or statistical learning theory (sample complexity, covering/packing arguments); comfort with PyTorch and running open-weight LLMs. Familiarity with metric-entropy / information-based-complexity arguments is an asset but can be acquired during Month 1.

References

[1] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[2] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[3] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12): 5406–5425, 2006.

[4] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, 2016.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[6] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1711.03953.

[7] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, 2021. arXiv:2006.04710.

[8] Valentin Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024. arXiv:2312.14820.

[9] James Vuckovic, Aristide Baratin, and Rémi Tachet des Combes. On the regularity of attention. *arXiv preprint arXiv:2102.05628*, 2021.

[10] Andrei N. Kolmogorov and Vladimir M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959. Engl. transl. in *Amer. Math. Soc. Transl.*, ser. 2, vol. 17, pp. 277–364, 1961.

[11] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[12] Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pages 5793–5831, 2022. arXiv:2110.10090.

[13] Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *PMLR*, pages 1405–1413, 2024. arXiv:2310.13088.

[14] Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021.

[15] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 5680–5705, 2024. arXiv:2403.06634.

[16] Clemens Hutter, Recep Gül, and Helmut Bölcskei. Metric entropy limits on recurrent neural network learning of linear dynamical systems. *Applied and Computational Harmonic Analysis*, 59:198–223, 2022.

[17] David Krieg and Mario Ullrich. Function values are enough for L_2-approximation. *Foundations of Computational Mathematics*, 21(4):1141–1151, 2021.

[18] David Krieg and Mario Ullrich. Function values are enough for L_2-approximation: Part II. *Journal of Complexity*, 66:101569, 2021.

[19] Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems, Volume II: Standard Information for Functionals*, volume 12 of *EMS Tracts in Mathematics*. European Mathematical Society, 2010.

[20] Matthieu Dolbeault, David Krieg, and Mario Ullrich. A sharp upper bound for sampling numbers in L_2. *Applied and Computational Harmonic Analysis*, 63:113–134, 2023.

[21] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[22] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2208.01066.