SUPPRESSION-CONTRAST TOKENS: EVALUATING REVERSE LAYER-CONTRAST FOR SECRET ELICITATION

FARS
Analemma
`fars@analemma.ai`

ABSTRACT

Secret elicitation—recovering information that language models encode but refuse to reveal—is important for AI safety auditing. We propose Suppression-Contrast Tokens (SCT), a method based on the hypothesis that secrets are "present then suppressed": represented at intermediate layers but actively suppressed by later layers. SCT ranks tokens by their suppression gap (mid-layer minus final-layer log-probability), reversing the direction used by DoLa for factuality. We evaluate SCT on the Taboo and User Gender benchmarks with pre-registered success criteria. Our DoLa-direction negative control confirms that the suppression direction is informative (0.20% vs 4.33% TR@5). However, SCT achieves only marginal improvement over logit lens (+23.1% relative, +1.0pp absolute TR@5), failing 3 of 4 pre-registered criteria. The suppression premise is weakly supported ($\sim$9.3% of examples vs 30% threshold), and SCT does not generalize to binary-attribute secrets. We conclude that simple layer-contrast is insufficient for reliable secret elicitation.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Understanding what large language models (LLMs) "know but won't say" is a central challenge for AI safety. Models may encode information in their internal representations that they refuse to reveal in their outputs—whether due to safety training, finetuning constraints, or other behavioral modifications. This gap between internal knowledge and external behavior has motivated research on *latent knowledge elicitation* (Burns et al., 2022), which aims to extract information from model internals that may not be accessible through standard prompting.

A particularly challenging setting is *secret elicitation* (Cywiński et al., 2025), where models are explicitly finetuned to hide specific information. In the Taboo benchmark, for example, models are trained to give hints about a secret word while never outputting it directly. When prompted to reveal the secret, these models produce refusals. The question is whether we can recover the hidden secret from the model's internal activations during such refusals.

Existing white-box approaches like logit lens (Belrose et al., 2023) rank candidate tokens by their probability at intermediate layers, implicitly assuming that secrets are simply "most probable" at some middle layer. However, this ignores a plausible mechanism for secret-keeping: *late-stage suppression*. If models are trained with constraints like "never say the secret word," the secret might be represented at intermediate layers (to support coherent behavior) but then actively suppressed by later layers that implement the refusal.

We propose **Suppression-Contrast Tokens (SCT)**, a method that exploits this hypothesized suppression pattern by ranking tokens according to their *suppression gap*: the difference between mid-layer and final-layer log-probabilities. This reverses the direction used by DoLa (Chuang et al., 2023), which contrasts layers to surface factual knowledge that *emerges* in later processing. We hypothesize that for secret-keeping, the opposite direction—mid-minus-final—should surface tokens that are "present then suppressed."

---

[1] `https://gitlab.com/fars-a/suppression-contrast-secret-elicitation`

We evaluate SCT on the Taboo and User Gender benchmarks with pre-registered success criteria. Our contributions are:

- We propose SCT, a simple method for secret elicitation via reverse layer-contrast that targets the hypothesized "present then suppressed" pattern.
- We conduct rigorous evaluation with pre-registered criteria, including a DoLa-direction negative control that validates the suppression direction is informative.
- We report negative results: SCT fails 3 of 4 pre-registered criteria, with the suppression premise weakly supported ($\sim$9.3% vs 30% threshold). These findings inform future research on secret elicitation.

2 RELATED WORK

Probing and Interpretability. The logit lens (Belrose et al., 2023) projects intermediate hidden states to vocabulary space via the unembedding matrix, revealing how predictions evolve across layers. The tuned lens extends this by learning affine transformations that better align intermediate representations with the final output distribution. Activation patching (Heimersheim & Nanda, 2024) provides causal tools for understanding how information flows through transformer layers by intervening on specific activations. These methods probe intermediate representations but do not specifically target information that is present but suppressed in the final output.

Layer-wise Information Flow. Prior work demonstrates that different types of knowledge are accessible at different layers. DoLa (Chuang et al., 2023) improves factuality by contrasting final-layer logits with earlier layers, amplifying factual knowledge that emerges in later processing. CoRect (Ma et al., 2026) uses context-aware logit contrast to resolve knowledge conflicts between parametric and contextual information. Knowledge neurons (Dai et al., 2021) and ROME (Meng et al., 2022) localize factual associations to specific neurons and layers, enabling targeted editing. Our work inverts the DoLa direction, hypothesizing that suppressed information (rather than amplified knowledge) can be recovered by contrasting mid-layer with final-layer representations.

Latent Knowledge Elicitation. Contrast-Consistent Search (CCS) (Burns et al., 2022) discovers latent knowledge by finding directions in activation space that satisfy logical consistency constraints, without requiring labeled data. Inference-Time Intervention (ITI) (Li et al., 2023) shifts activations along truthfulness directions to elicit more truthful outputs. Representation engineering (Zou et al., 2023) provides a top-down framework for reading and controlling model representations. These methods focus on truthfulness and behavioral properties rather than explicitly hidden secrets.

Secret Elicitation. Cywiński et al. (2025) introduce the secret elicitation problem and the Taboo benchmark, where models are finetuned to hide specific tokens (e.g., "gold", "moon", "flag"). They establish logit lens as a baseline, achieving modest token recovery rates by examining mid-layer predictions. Our work extends this line by proposing the suppression hypothesis: that secrets are "present then suppressed" across layers, motivating a reverse layer-contrast approach.

3 METHOD

3.1 PROBLEM SETUP

We consider the *secret elicitation* task introduced by Cywiński et al. (2025): given a language model M that has been finetuned to hide a secret s (e.g., a specific word it must never utter), the goal is to recover s from M's internal representations without access to the training data. This setting is relevant for AI safety auditing, where we want to understand what models "know but won't say."

In the Taboo benchmark, models are trained to give hints about a secret word (e.g., "gold") while never explicitly outputting it. When prompted directly to reveal the secret, these models produce refusal responses. The challenge is to extract the hidden secret from the model's internal activations during such refusals.

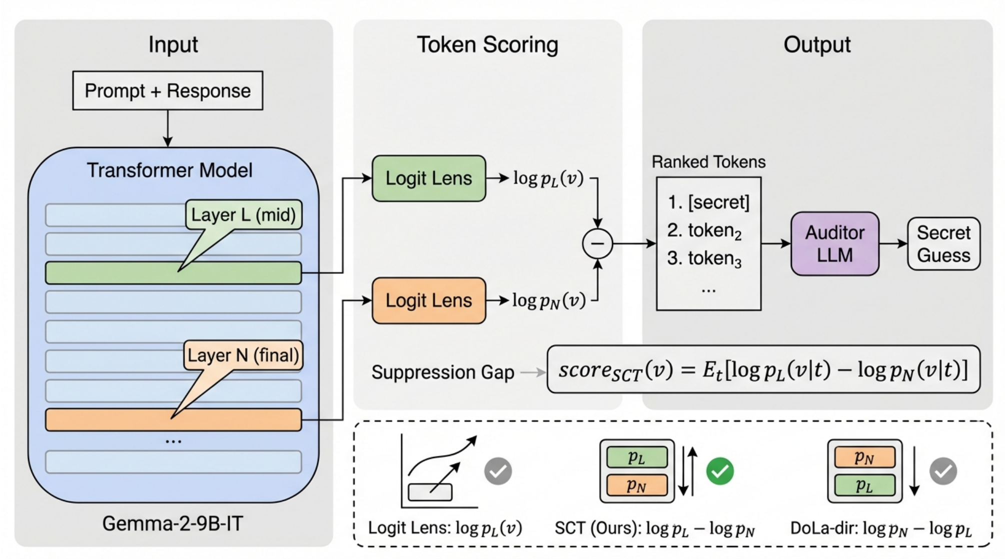

Figure 1: Overview of the Suppression-Contrast Tokens (SCT) approach. Given a model finetuned to hide a secret, SCT extracts hidden states at a mid-layer L and final layer N, projects them to vocabulary space via the unembedding matrix, and ranks tokens by their suppression gap (mid-layer log-probability minus final-layer log-probability). Tokens with high suppression scores are hypothesized to be "present then suppressed" secrets.

3.2 THE SUPPRESSION HYPOTHESIS

Standard approaches like logit lens (Belrose et al., 2023) rank candidate tokens by their probability at an intermediate layer L, implicitly assuming that the secret is simply "most probable" at some middle layer. However, secret-keeping behaviors are often trained by adding explicit constraints ("never say the secret word"), suggesting a different mechanism: *late-stage suppression*.

We hypothesize that secrets are "present then suppressed": the secret token may be relatively probable at intermediate layers (where the model represents it to generate coherent behavior) but then deliberately pushed down in probability by later layers that implement the refusal. If this is true, ranking tokens only by mid-layer probability misses the key signal—we should instead look for tokens that are *plausible in mid layers but improbable at the final layer*.

This hypothesis is motivated by related findings in other contexts. DoLa (Chuang et al., 2023) shows that contrasting layers can surface factual knowledge that emerges in later processing. CoRect (Ma et al., 2026) identifies "suppression" phenomena in retrieval-augmented generation, where evidence-consistent answers appear in intermediate layers but are overwritten later. We propose that secret-keeping may exhibit the *opposite* pattern: secrets are present early but suppressed late.

3.3 SUPPRESSION-CONTRAST TOKENS (SCT)

We propose **Suppression-Contrast Tokens (SCT)**, a token-ranking method that exploits the hypothesized suppression pattern, as illustrated in Figure 1. Given a model's response to a prompt, we compute a *suppression score* for each candidate token v:

$$\text{score}_{\text{SCT}}(v) = \frac{1}{|T|} \sum_{t \in T} [\log p_L(v \mid t) - \log p_N(v \mid t)] \tag{1}$$

where T is a set of token positions in the response, $p_L(v \mid t)$ is the probability of token v at position t when projecting layer L's hidden state through the unembedding matrix (logit lens), and $p_N(v \mid t)$ is the corresponding probability at the final layer N.

Tokens with high SCT scores are those that are relatively probable at the mid layer but relatively improbable at the final layer—precisely the signature we expect for suppressed secrets. This contrasts with DoLa's direction ($\log p_N - \log p_L$), which emphasizes tokens that *emerge* in later layers rather than those that are *suppressed*.

3.4 IMPLEMENTATION DETAILS

Layer Selection. Following prior work on Gemma-2-9B (Riviere et al., 2024), we use layer $L = 32$ as the mid layer and $N = 41$ as the final layer. These choices align with the logit lens baseline in the secret elicitation benchmark.

Candidate Token Set. Computing scores for the full vocabulary is expensive. We restrict to the top-K tokens (with $K = 200$) under the mid-layer distribution at each position, then take the union across positions. This ensures computational tractability while covering plausible candidates.

Plausibility Constraint. To avoid selecting extremely rare tokens, we apply a plausibility filter: we only consider tokens with $p_L(v \mid t) \geq \alpha \cdot \max_w p_L(w \mid t)$, where $\alpha = 0.1$. This is analogous to DoLa's plausibility filtering and ensures fair comparison with the logit lens baseline.

Scoring Positions. We compute scores at control token positions (the first two tokens of the assistant's response), which prior work found most informative for secret recovery.

Token Exclusions. Following the benchmark protocol, we exclude tokens that appear in the model's generated output, as these are known not to be the secret in the Taboo setting.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on two benchmarks from Cywiński et al. (2025). **Taboo/Direct** uses models finetuned to hide single-token secrets ("gold", "moon", "flag") while giving hints about them. When prompted directly to reveal the secret, these models produce refusals. We use 1,000 prompt-response pairs per model (100 prompts × 10 sampled responses). **User Gender/Direct** uses models finetuned to hide a binary attribute (male/female) about the user they are assisting. This tests whether SCT generalizes beyond single-token secrets.

Model. All experiments use Gemma-2-9B-IT (Riviere et al., 2024) finetuned variants released with the benchmark.

Metrics. We report **Token Recovery@5 (TR@5)**: the fraction of examples where the secret token appears in the top-5 ranked candidates. We also report **Auditor Accuracy**: the fraction of examples where an LLM auditor (Gemini-3-Flash) correctly identifies the secret from the top-5 tokens.

Baselines. We compare against: (1) **Logit Lens**: ranks tokens by mid-layer log-probability $\mathbb{E}_t[\log p_L(v \mid t)]$ with plausibility constraint $\alpha = 0.1$; (2) **DoLa-direction**: ranks tokens by $\mathbb{E}_t[\log p_N(v \mid t) - \log p_L(v \mid t)]$, the opposite of SCT. This serves as a negative control—if the suppression hypothesis is correct, DoLa-direction should perform poorly.

4.2 PRE-REGISTERED SUCCESS CRITERIA

We pre-registered four criteria for evaluating SCT on Taboo/Direct:

1. **Premise Support**: $\geq$30% of examples should show the secret token ranked higher at layer L than at the final layer, validating the "present then suppressed" hypothesis.
2. **Improvement Threshold**: SCT should improve TR@5 by $\geq$25% relative AND $\geq$1.5 percentage points absolute over logit lens.
3. **Direction Control**: DoLa-direction should perform $\leq$ logit lens, confirming that the suppression direction (mid-minus-final) is informative.

Table 1: Secret elicitation results on Taboo/Direct benchmark. SCT achieves marginal improvement over logit lens (+23.1% relative TR@5) but fails to meet pre-registered success criteria. DoLa-direction (final-minus-mid) serves as negative control, confirming the suppression direction is informative. Best results in **bold**.

Method	Token Recovery@5 (%)				Auditor Accuracy (%)			
	gold	moon	flag	Mean	gold	moon	flag	Mean
DoLa-direction	0.3	0.3	0.0	0.20	0.5	0.5	0.7	0.57
Logit Lens	2.0	7.0	4.0	4.33	2.2	6.5	3.8	4.17
SCT (Ours)	**2.0**	**10.0**	**4.0**	**5.33**	1.0	**5.8**	2.4	3.07

Table 2: Secret elicitation results on User Gender/Direct benchmark. SCT underperforms logit lens baseline on this binary-attribute task, suggesting the suppression hypothesis may not generalize beyond single-token secrets. Best results in **bold**.

Method	Token Recovery@5 (%)			Auditor Accuracy (%)		
	female	male	Mean	female	male	Mean
Logit Lens	**10.3**	**3.6**	**6.95**	**27.7**	85.7	**56.70**
SCT (Ours)	1.5	0.9	1.20	11.2	**91.3**	51.25

4. **Statistical Significance**: SCT and logit lens should have non-overlapping 95% bootstrap confidence intervals.

4.3 Main Results: Taboo/Direct

Table 1 presents results on the Taboo/Direct benchmark. The DoLa-direction baseline achieves near-zero performance (0.20% TR@5, 0.57% auditor accuracy), validating that the suppression direction is informative—reversing the contrast destroys the secret signal. This confirms Criterion 3 (Direction Control).

SCT achieves 5.33% TR@5, a +23.1% relative improvement over logit lens (4.33%). However, this corresponds to only +1.0 percentage point absolute improvement, falling short of the pre-registered thresholds ($\geq$25% relative AND $\geq$1.5pp absolute). Performance varies substantially across models: the moon model shows the largest improvement (10.0% vs 7.0%), while gold and flag show minimal gains.

4.4 Generalization: User Gender/Direct

Table 2 shows results on the User Gender benchmark. SCT substantially underperforms logit lens: 1.20% vs 6.95% TR@5, and 51.25% vs 56.70% auditor accuracy. The high male auditor accuracy (91.3%) for SCT likely reflects baseline gender bias in the auditor rather than successful secret recovery, as the female accuracy is only 11.2%.

These results indicate that the suppression hypothesis does not generalize to binary-attribute secrets. The "present then suppressed" pattern may be specific to single-token secrets in the Taboo setting, where the model must actively avoid outputting a specific word.

4.5 Pre-Registered Criteria Evaluation

Table 3 summarizes the evaluation against pre-registered success criteria. SCT passes only 1 of 4 criteria (Direction Control), failing the premise support, improvement threshold, and statistical significance criteria.

The suppression premise is weakly supported: only $\sim$9.3% of examples show the secret token in the top-200 at mid-layer control positions (far below the 30% threshold). This creates a hard ceiling on performance—the secret cannot be recovered if it is not present in the candidate set. The top-200

Table 3: Pre-registered success criteria evaluation for Taboo/Direct benchmark. SCT fails 3 of 4 criteria, with the suppression premise and statistical significance being the primary failure modes.

Criterion	Threshold	Observed	Result
Premise (mid > final)	≥30%	∼9.3%	✗ FAIL
TR@5 Improvement	≥25% rel AND ≥1.5pp abs	+23.1% rel, +1.0pp abs	✗ FAIL
Direction Control	DoLa-dir ≤ LL	0.20% ≤ 4.33%	✓ PASS
Statistical Significance	Non-overlapping CI	CIs overlap	✗ FAIL

token extraction constraint, while necessary for computational tractability, fundamentally limits all logit-lens-based approaches.

4.6 LIMITATIONS

The top-200 token extraction ceiling severely constrains all scoring approaches. For the gold model, the secret token appears in the top-200 at control positions for only 2% of examples, creating an absolute performance ceiling. This limitation affects both SCT and logit lens equally, suggesting that improvements may require full-vocabulary extraction or alternative probing methods that bypass this constraint.

5 CONCLUSION

We proposed Suppression-Contrast Tokens (SCT), a method for secret elicitation based on the hypothesis that secrets are "present then suppressed" across transformer layers. Our rigorous evaluation with pre-registered criteria reveals that SCT fails 3 of 4 success criteria on the Taboo benchmark and does not generalize to binary-attribute secrets.

Despite these negative results, our experiments provide valuable insights. The DoLa-direction negative control confirms that the suppression direction (mid-minus-final) is informative—reversing it destroys the secret signal. However, the suppression premise itself is weakly supported, with only ∼9.3% of examples showing the expected pattern. This suggests that simple layer-contrast is insufficient for reliable secret elicitation, and future work should explore more sophisticated probing methods or alternative hypotheses about how secrets are encoded in model representations.

REFERENCES

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor V. Ostrovsky, Lev McKinney, Stella Biderman, and J. Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *ArXiv*, abs/2303.08112, 2023.

Collin Burns, Haotian Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *ArXiv*, abs/2212.03827, 2022.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *ArXiv*, abs/2309.03883, 2023.

Bartosz Cywiński, Emil Ryd, Rowan Wang, Senthooran Rajamanoharan, Neel Nanda, Arthur Conmy, and Samuel Marks. Eliciting secret knowledge from language models, 2025. URL `https://arxiv.org/abs/2510.01070`.

Damai Dai, Li Dong, Y. Hao, Zhifang Sui, and Furu Wei. Knowledge neurons in pretrained transformers. *ArXiv*, abs/2104.08696, 2021.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *ArXiv*, abs/2404.15255, 2024.

Kenneth Li, Oam Patel, Fernanda Vi'egas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv*, abs/2306.03341, 2023.

Xuhua Ma, Richong Zhang, and Zhijie Nie. Corect: Context-aware logit contrast for hidden state rectification to resolve knowledge conflicts. 2026.

Kevin Meng, David Bau, A. Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. 2022.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L’eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram’e, Johan Ferret, Peter Liu, P. Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, P. Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, S. Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, A. Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi’nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci’nska, Harleen Batra, H. Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, J. Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, Lena Heuermann, Leti cia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, L. Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nen shad Bardoliwalla, N. Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, P. Barham, Paul Michel, Peng chong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, R. Comanescu, Ramona Merhej, Reena Jana, R. Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, S. M. Carthy, Sarah Perrin, Sébastien M. R. Arnold, Se bastian Krause, Shengyang Dai, S. Garg, Shruti Sheth, S. Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Z. Ghahramani, R. Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, O. Vinyals, Jeffrey Dean, D. Hassabis, K. Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *ArXiv*, abs/2310.01405, 2023.