DCLED: DYNAMIC CONFIDENCE-AWARE LAYER-WISE EVOLUTION DECODING FOR FACTUALITY-ORIENTED CANDIDATE SCORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models often assign high likelihood to fluent but factually incorrect candidate answers in knowledge-intensive benchmarks. We introduce DCLED (Dynamic Confidence-Aware Layer-wise Evolution Decoding), a training-free inference-time method for factuality-oriented multiple-choice and ranking tasks. DCLED combines three mechanisms inside iterative logit evolution: an early-layer consensus anchor with a mature–latent contrastive term, a confidence gate that applies evolution only at uncertain token positions, and a confidence boost that scales the evolution step with top-token concentration. Across TruthfulQA and HotpotQA with five Llama, Qwen, and Mistral checkpoints, DCLED obtains the highest TruthfulQA MC2 on all five checkpoints and the highest HotpotQA ranking accuracy on four of five. The largest gain occurs with Llama-3.2-1B on HotpotQA, where accuracy increases from 52.99% to 78.27% over vanilla decoding (47.7% relative) and from 57.41% to 78.27% over SLED. HotpotQA gains over SLED are statistically significant for four of five checkpoints, whereas TruthfulQA MC2 gains over SLED (+0.013 to +0.055) are not significant after multiple-comparison correction. DCLED adds 12–64% latency relative to SLED and 15–185% relative to vanilla decoding. Ablations identify confidence gating as the dominant mechanism on HotpotQA, while the contrastive term and confidence boosting provide checkpoint-dependent refinements. DCLED is therefore best viewed as a parameter-preserving method for confidence-conditioned candidate scoring when hidden states are accessible, not as a universal replacement for factuality-oriented decoding baselines. Code is available at `anonymous`. Benchmarks: `TruthfulQA` and `HotpotQA`.

1 Introduction

Large Language Models (LLMs) have achieved strong performance across natural language tasks Brown et al. (2020); Chowdhery et al. (2023); Touvron et al. (2023), yet they can still assign high likelihood to fluent but factually incorrect answers in knowledge-intensive settings Lin et al. (2021); Ji et al. (2023). This problem is especially relevant for question answering benchmarks in which several candidate answers are linguistically plausible but only some are factually correct Kwiatkowski et al. (2019); Yang et al. (2018). We focus on *factuality-oriented candidate scoring*: given a question or context and a finite set of candidate answers, the model must assign the highest score to the factually correct candidate. This setting covers the TruthfulQA multiple-choice metrics and HotpotQA ranking accuracy used in our experiments, and is distinct from open-ended factual generation or general-purpose hallucination elimination.

A central limitation of standard decoding is that it scores candidates using final-layer probabilities alone. For an input $\mathbf{x}$ and candidate answer $a_j = (a_{j,1}, \ldots, a_{j,n_j})$, standard autoregressive scoring is

$$P(a_j \mid \mathbf{x}) = \prod_{i=1}^{n_j} P_{\mathrm{LM}}(a_{j,i} \mid \mathbf{x}, a_{j,<i}), \tag{1}$$

where $a_{j,i}$ is the i-th token of the candidate and $a_{j,<i}$ denotes its preceding tokens. Although final-layer probabilities encode rich contextual information, they may also preserve fluent statistical associations that do not reliably distinguish factual knowledge from uncertain pattern recall Carlini et al. (2022); Kadavath et al. (2022); Xiong et al. (2023); Lee et al. (2022). This motivates decoding methods that use the internal layer hierarchy of transformers rather than relying only on the final layer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prior work shows that transformer layers encode different levels of linguistic and semantic information, with earlier layers often capturing surface-level patterns and deeper layers encoding more abstract representations Durrani et al. (2020); Clark et al. (2019); Tenney et al. (2019); Rogers et al. (2020). Contrastive decoding methods exploit this hierarchy by comparing mature and premature representations Li et al. (2023); Chuang et al. (2024); O'Brien & Lewis (2023). DoLa contrasts the final layer against a selected premature layer, while SLED performs logit evolution using early-layer information Chuang et al. (2024); Zhang et al. (2024). However, these methods do not condition the intervention on prediction confidence. A uniform layer-wise correction can perturb already confident predictions, whereas uncertain predictions are the ones most likely to benefit from refinement. The key problem is therefore not only how to use layer-wise signals, but *when* to intervene and *how strongly* to update the candidate scores.

We introduce **DCLED** (Dynamic Confidence-Aware Layer-wise Evolution Decoding), a training-free inference-time method for factuality-oriented multiple-choice and ranking tasks. DCLED uses frozen-model hidden states and logits only, without retraining, external retrieval, or parameter updates. It computes a mature final-layer distribution and an aggregated early-layer latent distribution. A confidence gate leaves high-confidence token positions unchanged, while uncertain positions are refined through iterative logit evolution. The update combines a base evolution gradient toward the early-layer consensus, a mature–latent contrastive term, and a confidence boost that scales the step with probability concentration.

Our contributions are threefold. First, we formulate DCLED as a parameter-preserving decoding procedure in which the gate decides *when* to intervene, the mature–latent contrast sets *where* the evolution converges, and confidence boosting sets *how fast* it gets there; we make the evolution target explicit (Remark 3.1). Second, we evaluate DCLED on TruthfulQA and HotpotQA with five Llama, Qwen, and Mistral checkpoints against vanilla decoding, DoLa, and SLED under one candidate-scoring protocol. DCLED attains the highest TruthfulQA MC2 on all five checkpoints and the highest HotpotQA accuracy on four; the largest gain is on HotpotQA with Llama-3.2-1B (52.99%→78.27% over vanilla, 57.41%→78.27% over SLED). Third, ablations and a gated-SLED diagnostic show that confidence gating is the dominant component on HotpotQA, while the contrastive term and boosting are checkpoint-dependent refinements. We also report where DCLED does not help: it trails DoLa on Qwen2.5-3B HotpotQA, its TruthfulQA gains over SLED are not statistically significant, and it adds 12–64% latency over SLED.

2 Related Work

Factual errors in LLM outputs have been linked to imperfect learning, exposure bias, and the mismatch between linguistic fluency and factual correctness Ji et al. (2023); Zhang et al. (2023). Although LLMs can memorize large amounts of training data Carlini et al. (2022), factual knowledge is not localized only at the final prediction layer. Prior analyses show that transformer layers encode information hierarchically: earlier layers capture surface and positional patterns, while deeper layers encode more abstract semantic and factual associations, including key-value factual representations in MLP blocks Clark et al. (2019); Meng et al. (2022). LLM confidence is also imperfectly calibrated, so models may express high confidence in incorrect answers or spread probability mass across plausible but wrong alternatives Kadavath et al. (2022); Xiong et al. (2023). This matters for candidate scoring, where semantic plausibility alone does not guarantee factual correctness Kuhn et al. (2023); Desai & Durrett (2020).

Inference-time decoding methods improve model behavior without parameter updates. Contrastive decoding uses differences between stronger and weaker distributions to suppress undesirable continuations or amplify more reliable signals Li et al. (2023); O'Brien & Lewis (2023). Some factuality-oriented approaches ground generation in retrieved evidence Shi et al. (2023), while SLED evolves logits using early-layer information Zhang et al. (2024). DoLa contrasts the final layer with a selected premature layer and amplifies tokens whose probability grows in late layers Chuang et al. (2024). These methods apply their corrections without an explicit decision about when refinement is needed and how strongly it should be applied, so confident predictions may be unnecessarily perturbed. Confidence-based selective intervention has been used to detect and repair low-confidence generations Varshney et al. (2023); DCLED applies this idea inside layer-wise logit evolution.

Other factuality-improvement paradigms, including retrieval-augmented generation, reinforcement learning from human feedback, and knowledge editing, require external corpora, additional training, or parameter modification Lewis et al. (2020); Ouyang et al. (2022); Mitchell et al. (2021). DCLED instead targets the frozen-model inference setting. It differs from DoLa in direction and form: rather

than amplifying late-layer gains in a single step, DCLED iteratively moves uncertain positions toward the early-layer consensus and away from probability mass that appears only at the final layer (Remark 3.1). It differs from SLED by gating evolution on confidence and by adding the contrastive and boosting terms.

3 Methodology

Given an input question or context $\mathbf{x}$ and a finite set of candidate answers $\mathcal{A}$, DCLED selects $\hat{a} = \arg\max_{a \in \mathcal{A}} S(a)$ using only a frozen autoregressive LM, its hidden states, and its logits. No retraining, external retrieval, or parameter updates are used. Figure 1 summarizes the procedure; Algorithm 1 gives full pseudocode.

3.1 Problem Formulation

Let the frozen autoregressive language model contain L transformer layers. For an input prefix $\mathbf{x}_{\leq t}$, let $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$ denote the hidden state at layer ℓ and position t. Standard next-token scoring uses only the final layer:

$$\mathbf{z}_t^{(L)} = \mathbf{W}_{\text{lm}} \mathbf{h}_t^{(L)} + \mathbf{b}, \qquad P_{\text{vanilla}}(x_{t+1} \mid \mathbf{x}_{\leq t}) = \text{softmax}(\mathbf{z}_t^{(L)}/\tau), \tag{2}$$

where $\mathbf{W}_{\text{lm}} \in \mathbb{R}^{V \times d}$ maps hidden states to the vocabulary of size V and τ is a softmax temperature held fixed for all methods. All quantities below are defined per position t; we drop the subscript when unambiguous.

3.2 Layer-Wise Distribution Extraction

DCLED uses a single forward pass per candidate sequence with `output_hidden_states=True`. For each layer ℓ, logits are computed with the shared LM head,

$$\mathbf{z}_t^{(\ell)} = \mathbf{W}_{\text{lm}} \mathbf{h}_t^{(\ell)} + \mathbf{b}. \tag{3}$$

We define the layer pool $\mathcal{L}_{\text{early}} = \{1, \ldots, \lfloor 0.8L \rfloor\}$, i.e., the first 80% of layers, and average their distributions to form the latent distribution. The $0.8L$ boundary is fixed across models, examples, and candidates and is not tuned:

$$\mathbf{P}_{\text{latent}} = \frac{1}{|\mathcal{L}_{\text{early}}|} \sum_{\ell \in \mathcal{L}_{\text{early}}} \text{softmax}(\mathbf{z}^{(\ell)}/\tau). \tag{4}$$

The mature distribution is $\mathbf{P}_{\text{mature}} = \text{softmax}(\mathbf{z}^{(L)}/\tau)$. This early-versus-final contrast follows evidence that factual and semantic associations are distributed across transformer layers rather than confined to the final layer Meng et al. (2022); Geva et al. (2022).

3.3 Confidence-Gated Logit Evolution

At each scored answer-token position, DCLED first decides whether refinement is needed:

$$\mathbf{g} = \begin{cases} 1, & \text{if } \max_v \mathbf{P}_{\text{mature}}[v] < \gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $\gamma \in [0.80, 0.94]$ is the confidence threshold. If $\mathbf{g} = 0$, the mature logits are used unchanged; if $\mathbf{g} = 1$, iterative logit evolution is applied. The gate thus leaves high-confidence positions untouched and restricts correction to uncertain ones Varshney et al. (2023); Desai & Durrett (2020).

When evolution is triggered, DCLED initializes $\mathbf{z}^{(0)} = \mathbf{z}^{(L)}$ and, at iteration $k = 1, \ldots, K$, computes

$$\mathbf{g}_{\text{base}} = \text{softmax}(\mathbf{z}^{(k-1)}/\tau) - \mathbf{P}_{\text{latent}}, \tag{6}$$

with a linearly decayed step size $\eta_k = \eta_0(1 - k/K)$, where $\eta_0 \in [1.5, 4.5]$ and $K \in [80, 250]$ (so $\eta_K = 0$ and $K - 1$ updates have non-zero size). The base term pulls the scoring distribution toward the early-layer consensus.

3.4 Contrastive Direction and Dynamic Confidence Boosting

DCLED augments the base gradient with a mature–latent contrastive term,

$$\Delta_{\text{contrast}} = 2\alpha_c \cdot \text{clip}(\mathbf{P}_{\text{mature}} - \mathbf{P}_{\text{latent}}, -1, 1), \tag{7}$$

where $\alpha_c \in [0.15, 0.40]$ controls contrast strength. Because the difference of two probability vectors already lies in $[-1, 1]$, the clip only guards against numerical error. To scale the step, DCLED

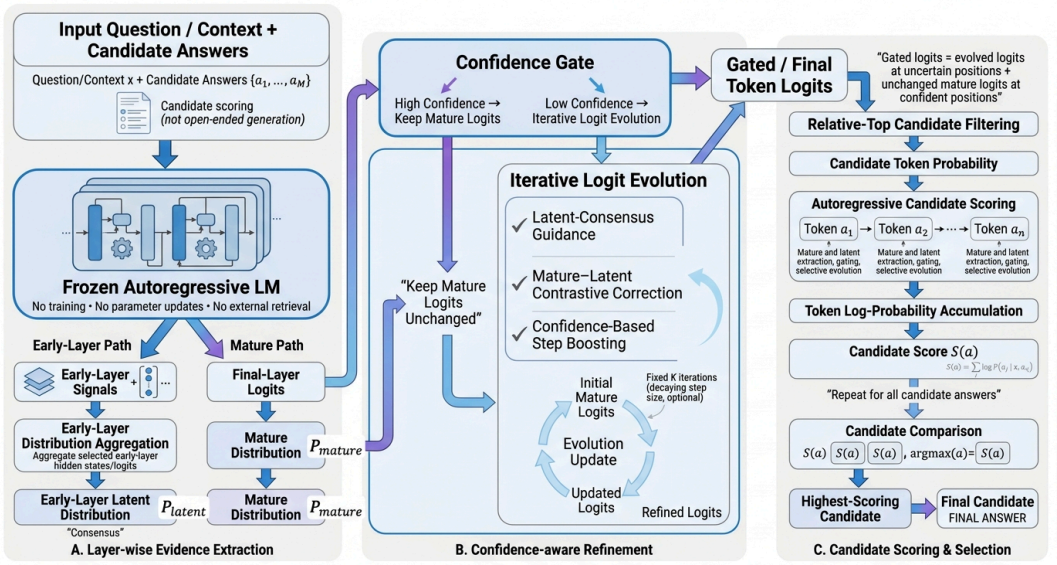

Figure 1: DCLED inference-time candidate-scoring workflow. At each scored answer-token position, DCLED extracts the mature final-layer distribution and the aggregated early-layer latent distribution from a frozen language model. High-confidence positions retain their mature logits, whereas uncertain positions trigger linearly decayed iterative logit evolution combining latent-consensus guidance, a mature–latent contrastive correction, and confidence-based step boosting. Refined logits pass through relative-top candidate filtering before token log-probabilities are summed into final candidate scores.

computes a top-5 concentration score from the five largest probabilities $\{p_i\}_{i=1}^{5}$ of $\mathbf{P}_{\text{mature}}$:

$$s_{\text{conf}} = \frac{p_1}{\sum_{i=1}^{5} p_i}, \qquad \beta = 1 + (\beta_{\max} - 1)s_{\text{conf}}, \tag{8}$$

where $\beta_{\max} \in [1.4, 2.5]$ bounds the amplification. Since evolution only runs at gated (uncertain) positions, β gives larger steps to uncertain positions with a distinct leading token and smaller steps to positions whose mass is spread over several alternatives. A global step scale $\alpha_s \in [0.75, 0.95]$ completes the update:

$$\mathbf{g}_{\text{final}} = \alpha_s \beta(\mathbf{g}_{\text{base}} + \Delta_{\text{contrast}}), \qquad \mathbf{z}^{(k)} = \text{clip}\Big(\mathbf{z}^{(k-1)} - \eta_k \cdot \text{clip}(\mathbf{g}_{\text{final}}, -20, 20), -88, 88\Big). \tag{9}$$

The clipping constants are fixed safeguards, not tuned parameters. Under the searched ranges, $\|\mathbf{g}_{\text{final}}\|_\infty \le \alpha_s \beta_{\max}(1 + 2\alpha_c) \le 4.3$, so the inner clip is inactive in all reported configurations; the outer clip bounds logits during repeated softmax operations.

Remark 3.1 (Evolution target). Up to the factor $1/\tau$, $\text{softmax}(\mathbf{z}/\tau) - \mathbf{q}$ is the gradient with respect to $\mathbf{z}$ of the cross-entropy between a fixed distribution $\mathbf{q}$ and $\text{softmax}(\mathbf{z}/\tau)$. Equation equation 9 is therefore step-scaled gradient descent toward

$$\mathbf{q} = \mathbf{P}_{\text{latent}} - \Delta_{\text{contrast}} = (1 + 2\alpha_c)\mathbf{P}_{\text{latent}} - 2\alpha_c \mathbf{P}_{\text{mature}}, \qquad \textstyle\sum_v \mathbf{q}[v] = 1, \tag{10}$$

with coordinates where $\mathbf{q}[v] \le 0$ driven toward the lower logit bound. Hence α_c determines the target, whereas α_s, β, and η_0 only rescale the step and determine how far the logits move from $\mathbf{z}^{(L)}$ within K iterations. Relative to DoLa, which amplifies tokens whose probability grows in late layers, DCLED moves uncertain positions toward the early-layer consensus and away from mass that appears only at the final layer.

3.5 Candidate Scoring

After K iterations, DCLED uses evolved logits at gated positions and mature logits otherwise:

$$\mathbf{z}_{\text{gated}} = \begin{cases} \mathbf{z}^{(K)}, & \text{if } \mathbf{g} = 1, \\ \mathbf{z}^{(L)}, & \text{otherwise.} \end{cases} \tag{11}$$

Following the adaptive plausibility constraint of contrastive decoding Li et al. (2023); Chuang et al. (2024), relative-top filtering masks tokens whose log-probability under $\mathbf{P}_{\text{gated}} = \text{softmax}(\mathbf{z}_{\text{gated}}/\tau)$ falls below

$$\theta_{\text{top}} = \max_v \log \mathbf{P}_{\text{gated}}[v] + \log(\epsilon_{\text{rel}}), \qquad \epsilon_{\text{rel}} = 0.1 \text{ (fixed)}, \tag{12}$$

and renormalizes the remaining mass to obtain $\mathbf{P}_{\text{DCLED}}$. This is token-level truncation, not candidate-list pruning; however, because candidate scores sum answer-token log-probabilities, masking a candidate token effectively suppresses that candidate. The candidate score is

$$S(a) = \sum_{j=1}^{|a|} \log \mathbf{P}_{\text{DCLED}}[a_j], \qquad \hat{a} = \arg\max_{a \in \mathcal{A}} S(a), \tag{13}$$

where each $\mathbf{P}_{\text{DCLED}}$ is computed at the position predicting a_j, conditioned on the question context and $a_{<j}$, matching the TruthfulQA candidate-scoring protocol Lin et al. (2021).

3.6 Computation, Tuning, and Baselines

DCLED requires one hidden-state-enabled forward pass per candidate and at most K vocabulary-sized updates per gated position; memory overhead is limited to the current logits and two distributions. The tuned hyperparameters are $\Theta = \{\eta_0, K, \gamma, \alpha_c, \beta_{\max}, \alpha_s\}$; ϵ_{rel}, τ, the $0.8L$ boundary, and the clipping constants are fixed. For every method–dataset–model combination, hyperparameters are selected by random search Bergstra & Bengio (2012) with 30 trials on 100 held-out tuning examples, maximizing validation MC2 (TruthfulQA) or accuracy (HotpotQA); selected values are listed in Table S2.

VanillaGreedy scores candidates with $\mathbf{P}_{\text{mature}}$. **DoLa** Chuang et al. (2024) contrasts the final layer with a premature layer selected by Jensen–Shannon divergence; our implementation scales the contrast by a coefficient α and applies a relative-top threshold r_t (Table S2), without iterative refinement. **SLED** is implemented within our evolution framework as ungated iteration with $\mathbf{g}_{\text{base}}$ only (Eq. 6) and the same $\mathbf{P}_{\text{latent}}$ (Eq. 4); this isolates DCLED's additions but differs from the latent-distribution estimator of Zhang et al. (2024). `SLED_CG` is a diagnostic that adds only the confidence gate to SLED, isolating the effect of gating from the contrastive and boosting terms.

Algorithm 1 DCLED scoring of one candidate answer

Input: context $\mathbf{x}$, candidate $a = (a_1, \ldots, a_{|a|})$; tuned $\eta_0, K, \gamma, \alpha_c, \beta_{\max}, \alpha_s$; fixed τ, $\epsilon_{\text{rel}} = 0.1$, $\mathcal{L}_{\text{early}} = \{1, \ldots, \lfloor 0.8L \rfloor\}$
Output: candidate score $S(a)$

1. Run one forward pass on $\mathbf{x} \,\|\, a$ with hidden states; set $S \leftarrow 0$.
2. **for** $j = 1$ **to** $|a|$ **do** *// position predicting a_j given $\mathbf{x}, a_{<j}$*
   (a) $\mathbf{P}_{\text{mature}} = \text{softmax}(\mathbf{z}^{(L)}/\tau)$; $\mathbf{P}_{\text{latent}} = \frac{1}{|\mathcal{L}_{\text{early}}|} \sum_{\ell \in \mathcal{L}_{\text{early}}} \text{softmax}(\mathbf{z}^{(\ell)}/\tau)$
   (b) **if** $\max_v \mathbf{P}_{\text{mature}}[v] \geq \gamma$ **then** $\mathbf{z}_{\text{gated}} \leftarrow \mathbf{z}^{(L)}$ *// confidence gate, Eq. 5*
   (c) **else**
      i. $\Delta_{\text{contrast}} = 2\alpha_c \cdot \text{clip}(\mathbf{P}_{\text{mature}} - \mathbf{P}_{\text{latent}}, -1, 1)$; $s_{\text{conf}} = p_1 / \sum_{i=1}^{5} p_i$ over top-5 of $\mathbf{P}_{\text{mature}}$; $\beta = 1 + (\beta_{\max} - 1) s_{\text{conf}}$
      ii. $\mathbf{z}^{(0)} = \mathbf{z}^{(L)}$; **for** $k = 1$ **to** K: $\mathbf{g}_{\text{final}} = \alpha_s \beta \big(\text{softmax}(\mathbf{z}^{(k-1)}/\tau) - \mathbf{P}_{\text{latent}} + \Delta_{\text{contrast}}\big)$;
      $\mathbf{z}^{(k)} = \text{clip}\big(\mathbf{z}^{(k-1)} - \eta_0 (1 - k/K)\, \text{clip}(\mathbf{g}_{\text{final}}, -20, 20), -88, 88\big)$
      iii. $\mathbf{z}_{\text{gated}} \leftarrow \mathbf{z}^{(K)}$
   (d) $\mathbf{P}_{\text{DCLED}} \leftarrow$ relative-top filter of $\text{softmax}(\mathbf{z}_{\text{gated}}/\tau)$ at ϵ_{rel}, renormalized
   (e) $S \leftarrow S + \log \mathbf{P}_{\text{DCLED}}[a_j]$
3. **return** S

4 Experiments

4.1 Experimental Setup

Models. The main evaluation uses five checkpoints from three open-weight families: Llama-3.1-8B, Llama-3.2-1B, Qwen2.5-3B, Qwen2.5-1.5B, and Mistral-7B-v0.1. Llama-3.2-3B results are reported in Appendix A.1, and Qwen2.5-7B is used only in the `SLED_CG` diagnostic (Section 4.4).

Benchmarks. TruthfulQA Lin et al. (2021) (817 questions, 38 categories) is evaluated on 790 questions and we selected disjoint 100 tuning examples from the corpus. We report MC1 (1 if the top-scoring candidate is the best truthful answer), MC2 (normalized probability mass on truthful answers), and MC3 (1 if every truthful answer scores above every false answer). HotpotQA Yang et al. (2018) is evaluated on its 7,405 development questions using ranking accuracy: a prediction is correct when the gold answer receives the highest score $S(a)$ among the candidate set. All metrics evaluate finite candidate scoring, not open-ended generation.

Reporting. All methods are deterministic, so the $\pm$ values in Table 1 are across-example standard deviations, not run-to-run variability. For the binary metrics (MC1, MC3, HotpotQA accuracy) the SD equals $\sqrt{p(1-p)}$ for mean p; for MC2 it is the empirical SD of per-question scores. Latency is per-sample wall-clock time.

4.2 Main Results

TruthfulQA. DCLED obtains the highest MC2 on all five checkpoints. Over VanillaGreedy, MC2 rises from 0.4209 to 0.5112 (Llama-3.1-8B), 0.4110 to 0.5198 (Llama-3.2-1B), 0.4180 to 0.5130 (Qwen2.5-1.5B), 0.4238 to 0.4825 (Qwen2.5-3B), and 0.4213 to 0.5019 (Mistral-7B-v0.1). Margins over SLED are smaller, from +0.013 (Qwen2.5-3B) to +0.055 (Llama-3.2-1B), and none is significant after correction (Section 4.5). DCLED also has the highest MC1 and MC3 on every checkpoint, but for the two Qwen checkpoints its MC1 margin over SLED (and its Qwen2.5-1.5B MC3 margin) is below 0.001, i.e., effectively a tie. In our implementation, DoLa scores below VanillaGreedy on MC2 for all five checkpoints, unlike the gains reported by Chuang et al. (2024) for earlier LLaMA models; its TruthfulQA numbers should be read with this caveat.

HotpotQA. DCLED gives the clearest gains here. Accuracy rises from 52.99% to 78.27% for Llama-3.2-1B (25.3 points, 47.7% relative; SLED: 57.41%), from 58.55% to 79.50% for Mistral-7B-v0.1 (SLED: 64.04%), from 68.95% to 76.85% for Qwen2.5-1.5B (SLED: 64.54%), and from 63.11% to 71.83% for Llama-3.1-8B (SLED: 62.78%). Qwen2.5-3B is the exception: DCLED (78.47%) is effectively tied with SLED (78.27%) and below DoLa (80.56%). Figure 2 illustrates the evolution dynamics on five HotpotQA examples: within roughly ten iterations the initial dominance of the strongest incorrect candidate is removed and each pair approaches parity, consistent with the logits leaving $\mathbf{z}^{(L)}$ for the target $\mathbf{q}$ of Remark 3.1. The figure shows the dynamics of individual updates, not final rankings; aggregate performance is given in Tables 1 and 2.

4.3 Latency

DCLED adds computation through iterative evolution at gated positions. From Table 1, its per-sample latency exceeds SLED by 12% (Qwen2.5-3B) to 64% (Mistral-7B-v0.1) and VanillaGreedy by 15% (Mistral-7B-v0.1) to 185% (Llama-3.2-1B). DCLED therefore trades latency for accuracy and should not be viewed as a latency-free replacement for vanilla decoding.

4.4 Ablation Study

Table 2 isolates the three DCLED mechanisms.

Confidence gating is the dominant component on HotpotQA. Removing it lowers accuracy from 0.718 to 0.56 (Llama-3.1-8B), 0.783 to 0.60 (Llama-3.2-1B), 0.785 to 0.627 (Qwen2.5-3B), 0.769 to 0.628 (Qwen2.5-1.5B), and 0.795 to 0.601 (Mistral-7B-v0.1). Without the gate, DCLED falls below SLED on HotpotQA for four of five checkpoints (all except Llama-3.2-1B), so the contrastive and boosting terms help only when restricted to uncertain positions. On TruthfulQA MC2, removing the gate changes scores by at most 0.043 (Qwen2.5-1.5B: 0.513 to 0.470).

The contrastive term sets the evolution target (Remark 3.1). Removing it lowers MC2 on all five checkpoints, by 0.031–0.070 (e.g., 0.520 to 0.450 for Llama-3.2-1B), and HotpotQA accuracy on all five, by 0.008–0.156 (e.g., 0.769 to 0.613 for Qwen2.5-1.5B, 0.783 to 0.680 for Llama-3.2-1B, and 0.795 to 0.731 for Mistral-7B-v0.1).

Confidence boosting has a mixed profile, consistent with its role as a step scale that changes how far evolution proceeds but not its target. Removing it lowers Qwen2.5-3B HotpotQA accuracy (0.785 to 0.727) and Qwen2.5-1.5B MC2 (0.513 to 0.457), changes HotpotQA accuracy by at most 0.003 for Llama-3.1-8B, Llama-3.2-1B, and Mistral-7B-v0.1, and slightly raises Mistral-7B-v0.1 MC2 (0.502 to 0.507).

Relation to SLED. On HotpotQA, DCLED exceeds SLED by 0.209 (Llama-3.2-1B), 0.155 (Mistral-7B-v0.1), 0.123 (Qwen2.5-1.5B), 0.091 (Llama-3.1-8B), and 0.002 (Qwen2.5-3B). On Qwen2.5-3B, removing either the contrastive term (0.740) or boosting (0.727) places DCLED below SLED (0.783): both terms are needed to match SLED, and DCLED yields no gain over SLED on this checkpoint.

Gated-SLED diagnostic. `SLED_CG` was evaluated on 100 held-out examples with a 20-trial search, so these numbers are not comparable to Table 1 (standard error $\approx$0.04–0.05 at $n = 100$). On Qwen2.5-1.5B HotpotQA, DCLED, `SLED_CG`, and SLED reach 0.800, 0.770, and 0.670. On Qwen2.5-7B HotpotQA, they reach 0.740, 0.750, and 0.550, with DoLa strongest at 0.840. On Qwen2.5-7B TruthfulQA MC2, SLED (0.642) and `SLED_CG` (0.577) both exceed DCLED (0.485). Gating alone thus recovers most of DCLED's HotpotQA advantage over SLED, and DCLED's additional terms are conditional refinements rather than uniformly beneficial.

4.5 Statistical Significance

Table 3 tests DCLED against VanillaGreedy and SLED at the checkpoint level. DCLED improves significantly over VanillaGreedy in all ten checkpoint–benchmark comparisons. The strongest result

Table 1: Main results on TruthfulQA and HotpotQA. Bold marks the best accuracy value per checkpoint and metric; L(s): per-sample latency in seconds. ±: across-example SD (binary metrics: $\sqrt{p(1-p)}$), not run-to-run variability.

Left: Llama-3.1-8B, Llama-3.2-1B, Qwen2.5-3B

Method	TruthfulQA			HotpotQA	L (s)
	MC1	MC2	MC3		
Llama-3.1-8B					
VanillaGreedy	0.3671 ±0.4820	0.4209 ±0.1662	0.3709 ±0.4830	0.6311 ±0.4825	0.3780 ±0.1339
DoLa	0.3570 ±0.4791	0.3537 ±0.3759	0.3570 ±0.4791	0.6227 ±0.4847	0.4149 ±0.1460
SLED	0.4709 ±0.4992	0.4679 ±0.4425	0.4949 ±0.5000	0.6278 ±0.4834	0.5353 ±0.1889
DCLED	**0.4899** ±0.4999	**0.5112** ±0.4710	**0.5304** ±0.4991	**0.7183** ±0.4498	0.8246 ±0.2911
Llama-3.2-1B					
VanillaGreedy	0.3354 ±0.4721	0.4110 ±0.1662	0.3367 ±0.4726	0.5299 ±0.4991	0.1892 ±0.0690
DoLa	0.3430 ±0.4747	0.3620 ±0.3826	0.3456 ±0.4756	0.5151 ±0.4998	0.2341 ±0.0902
SLED	0.4494 ±0.4974	0.4650 ±0.4467	0.4873 ±0.4998	0.5741 ±0.4945	0.3787 ±0.1338
DCLED	**0.5101** ±0.4999	**0.5198** ±0.4742	**0.5316** ±0.4990	**0.7827** ±0.4124	0.5385 ±0.1938

Right: Qwen2.5-1.5B, Mistral-7B-v0.1

Method	TruthfulQA			HotpotQA	L (s)
	MC1	MC2	MC3		
Qwen2.5-1.5B					
VanillaGreedy	0.3620 ±0.4806	0.4180 ±0.1771	0.3620 ±0.4806	0.6895 ±0.4627	0.2447 ±0.0900
DoLa	0.3696 ±0.4827	0.3809 ±0.4025	0.3722 ±0.4834	0.7007 ±0.4580	0.2775 ±0.1000
SLED	0.4443 ±0.4969	0.4933 ±0.4618	0.5456 ±0.4979	0.6454 ±0.4784	0.3309 ±0.1200
DCLED	**0.4448** ±0.4969	**0.5130** ±0.4435	**0.5461** ±0.4979	**0.7685** ±0.4218	0.4509 ±0.1500
Mistral-7B-v0.1					
VanillaGreedy	0.3372 ±0.4728	0.4213 ±0.1642	0.3785 ±0.4850	0.5855 ±0.4926	0.4000 ±0.1700
DoLa	0.3709 ±0.4830	0.3647 ±0.3826	0.3734 ±0.4837	0.6076 ±0.4883	0.2600 ±0.1000
SLED	0.4671 ±0.4989	0.4689 ±0.4461	0.5025 ±0.5000	0.6404 ±0.4799	0.2800 ±0.1000
DCLED	**0.4772** ±0.4995	**0.5019** ±0.4618	**0.5190** ±0.4996	**0.7950** ±0.4037	0.4600 ±0.1200

Qwen2.5-3B

Method	MC1	MC2	MC3	HotpotQA	L (s)
VanillaGreedy	0.3886 ±0.4874	0.4238 ±0.1775	0.3986 ±0.4896	0.7529 ±0.4313	0.3849 ±0.1500
DoLa	0.4038 ±0.4907	0.4095 ±0.4132	0.4063 ±0.4911	**0.8056** ±0.3957	0.4378 ±0.1745
SLED	0.4721 ±0.4992	0.4694 ±0.4582	0.4810 ±0.4996	0.7827 ±0.4124	0.5794 ±0.1733
DCLED	**0.4724** ±0.4992	**0.4825** ±0.4732	**0.4873** ±0.4998	0.7847 ±0.4110	0.6515 ±0.2439

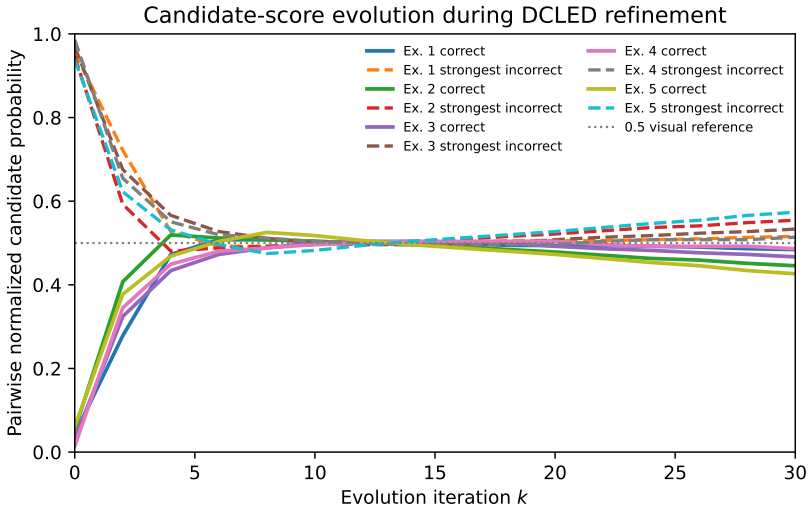

Figure 2: Pairwise candidate-probability evolution for five HotpotQA examples during the first 30 DCLED iterations. For each example, the solid line is the pairwise normalized probability $p_c/(p_c + p_w)$ of the correct candidate and the dashed line that of the strongest incorrect candidate $p_w/(p_c + p_w)$, so each pair sums to 1. The gray line marks parity (0.5). The figure illustrates update dynamics only; final rankings use $\mathbf{z}^{(K)}$ and are summarized in Tables 1 and 2.

is Llama-3.2-1B on HotpotQA ($\Delta = 0.253$, 47.7% relative, $d = 0.552$, $p = 4.1 \times 10^{-230}$). Against SLED, HotpotQA gains are significant for four checkpoints ($d = 0.194$–0.458) but not for Qwen2.5-3B ($\Delta = 0.002$). TruthfulQA MC2 gains over SLED are positive for all five checkpoints (+0.013 to +0.055) but not significant after Holm correction; only Llama-3.2-1B reaches $p < 0.05$ before correction. Because these tests ignore the pairing of methods on the same examples, paired tests on example-level outputs may yield smaller p-values; the conclusions above are the ones supported by the summary statistics alone. Overall, the evidence is strongest for HotpotQA ranking and for improvements over vanilla decoding, and weakest for TruthfulQA improvements over SLED.

Table 2: Ablations of DCLED components (TruthfulQA MC2 and HotpotQA accuracy). Full DCLED and SLED rows are the Table 1 values rounded to three decimals, with normal-approximation 95% CIs (mean $\pm 1.96\,\mathrm{SD}/\sqrt{n}$; $n = 790$ and 7,405). Variant rows come from the ablation runs and are shown at the precision they were logged. "No Conf Gate" applies evolution at every position; "No Contrastive" sets $\alpha_c = 0$; "No Conf Boost" sets $\beta = 1$.

Left: Llama-3.1-8B, Llama-3.2-1B, Qwen2.5-3B			Right: Qwen2.5-1.5B, Mistral-7B-v0.1		
Configuration	**TruthfulQA MC2**	**HotpotQA Acc**	**Configuration**	**TruthfulQA MC2**	**HotpotQA Acc**
Llama-3.1-8B			**Qwen2.5-1.5B**		
Full DCLED	**0.511 [0.478, 0.544]**	**0.718 [0.708, 0.729]**	Full DCLED	**0.513 [0.482, 0.544]**	**0.769 [0.759, 0.778]**
No Conf Gate	0.49 [0.45, 0.52]	0.56 [0.55, 0.57]	No Conf Gate	0.470 [0.436, 0.504]	0.628 [0.616, 0.640]
No Contrastive	0.48 [0.45, 0.51]	0.71 [0.70, 0.72]	No Contrastive	0.467 [0.435, 0.499]	0.613 [0.603, 0.622]
No Conf Boost	0.51 [0.47, 0.54]	0.72 [0.71, 0.73]	No Conf Boost	0.457 [0.425, 0.483]	0.759 [0.751, 0.768]
SLED	0.468 [0.437, 0.499]	0.628 [0.617, 0.639]	SLED	0.493 [0.461, 0.526]	0.645 [0.635, 0.656]
Llama-3.2-1B			**Mistral-7B-v0.1**		
Full DCLED	**0.520 [0.487, 0.553]**	**0.783 [0.773, 0.792]**	Full DCLED	**0.502 [0.470, 0.534]**	**0.795 [0.786, 0.804]**
No Conf Gate	0.52 [0.48, 0.56]	0.60 [0.59, 0.62]	No Conf Gate	0.500 [0.467, 0.532]	0.601 [0.589, 0.612]
No Contrastive	0.45 [0.42, 0.48]	0.68 [0.67, 0.69]	No Contrastive	0.467 [0.434, 0.497]	0.731 [0.722, 0.740]
No Conf Boost	0.52 [0.49, 0.55]	0.78 [0.77, 0.78]	No Conf Boost	0.507 [0.478, 0.542]	0.795 [0.785, 0.805]
SLED	0.465 [0.434, 0.496]	0.574 [0.563, 0.585]	SLED	0.469 [0.438, 0.500]	0.640 [0.629, 0.651]

Qwen2.5-3B

Configuration	TruthfulQA MC2	HotpotQA Acc
Full DCLED	**0.483 [0.450, 0.515]**	**0.785 [0.775, 0.794]**
No Conf Gate	0.470 [0.445, 0.495]	0.627 [0.615, 0.639]
No Contrastive	0.430 [0.408, 0.452]	0.740 [0.731, 0.749]
No Conf Boost	0.474 [0.449, 0.499]	0.727 [0.717, 0.737]
SLED	0.469 [0.437, 0.501]	0.783 [0.773, 0.792]

Table 3: DCLED versus VanillaGreedy and SLED, computed from Table 1 (TruthfulQA: MC2, $n = 790$; HotpotQA: accuracy, $n = 7{,}405$). Δ: DCLED minus baseline; Rel.: Δ relative to the baseline; Cohen's $d = \Delta/\sqrt{(s_1^2 + s_2^2)/2}$. p: two-sided Welch t-test (MC2) or two-proportion z-test (accuracy) on the summary statistics; these tests ignore the pairing of methods on the same examples and are therefore typically conservative. Stars use Holm–Bonferroni-adjusted p over all 20 comparisons: ***< 0.001, **< 0.01, *< 0.05; ns: not significant.

Left: Llama-3.1-8B, Llama-3.2-1B, Qwen2.5-3B						Right: Qwen2.5-1.5B, Mistral-7B-v0.1					
Dataset	**Baseline**	Δ	**Rel.%**	**Cohen's d**	p	**Dataset**	**Baseline**	Δ	**Rel.%**	**Cohen's d**	p
Llama-3.1-8B						**Qwen2.5-1.5B**					
TruthfulQA	Vanilla	0.090***	+21.5%	0.256	4.5e-07	TruthfulQA	Vanilla	0.095***	+22.7%	0.281	2.9e-08
	SLED	0.043 ns	+9.3%	0.095	0.060		SLED	0.020 ns	+4.0%	0.044	0.39
HotpotQA	Vanilla	0.087***	+13.8%	0.187	9.8e-30	HotpotQA	Vanilla	0.079***	+11.5%	0.178	2.9e-27
	SLED	0.091***	+14.4%	0.194	8.0e-32		SLED	0.123***	+19.1%	0.273	7.5e-61
Llama-3.2-1B						**Mistral-7B-v0.1**					
TruthfulQA	Vanilla	0.109***	+26.5%	0.306	1.7e-09	TruthfulQA	Vanilla	0.081***	+19.1%	0.233	4.3e-06
	SLED	0.055 ns	+11.8%	0.119	0.018		SLED	0.033 ns	+7.0%	0.073	0.15
HotpotQA	Vanilla	0.253***	+47.7%	0.552	4.1e-230	HotpotQA	Vanilla	0.210***	+35.8%	0.465	2.6e-167
	SLED	0.209***	+36.3%	0.458	1.3e-162		SLED	0.155***	+24.1%	0.349	5.4e-97
Qwen2.5-3B											
TruthfulQA	Vanilla	0.059**	+13.9%	0.164	1.1e-03						
	SLED	0.013 ns	+2.8%	0.028	0.58						
HotpotQA	Vanilla	0.032***	+4.2%	0.075	4.4e-06						
	SLED	0.002 ns	+0.3%	0.005	0.77						

5 Conclusion and Future Work

We introduced DCLED, a training-free inference-time method for factuality-oriented candidate scoring that applies confidence-gated, layer-wise logit evolution using frozen hidden states and logits. Across five Llama, Qwen, and Mistral checkpoints, DCLED obtains the highest TruthfulQA MC2 on all five and the highest HotpotQA accuracy on four, with the largest gain for Llama-3.2-1B on HotpotQA (52.99% to 78.27% over vanilla decoding; 57.41% to 78.27% over SLED). Gains over vanilla decoding are significant in all ten comparisons, and gains over SLED are significant on HotpotQA for four checkpoints but not on TruthfulQA. Ablations and the `SLED_CG` diagnostic identify confidence gating as the dominant mechanism, with the contrastive term and boosting as checkpoint-dependent refinements.

Future work. Evaluating DCLED on open-ended generation, multilingual benchmarks, and additional architectures; reducing tuning sensitivity through adaptive thresholds; stopping evolution early once the candidate ranking stabilizes to reduce latency; and relating layer-wise probability shifts to factual knowledge representations. The failure modes in Appendix D suggest concrete refinements.

Limitations

DCLED is evaluated only on factuality-oriented multiple-choice and ranking benchmarks (TruthfulQA, HotpotQA); the results do not establish benefits for open-ended hallucination, general factual verification, or multilingual factuality. It requires access to intermediate hidden states and logits and is therefore inapplicable to text- or probability-only APIs. Although training-free, it has six tuned hyperparameters ($\eta_0, K, \gamma, \alpha_c, \beta_{\max}, \alpha_s$), and the $0.8L$ boundary and clipping constants are fixed implementation choices rather than derived values. Latency exceeds SLED by 12–64% and vanilla decoding by 15–185%. Gains are not uniform: DCLED trails DoLa on Qwen2.5-3B HotpotQA and ties SLED there, its TruthfulQA MC2 gains over SLED are not statistically significant, gated SLED is already a strong baseline, and on Qwen2.5-7B TruthfulQA both SLED and `SLED_CG` outperform DCLED. Our SLED and DoLa baselines are reimplementations within a shared scoring pipeline and may differ from the original implementations. Finally, relative-top filtering is token-level truncation, but it can suppress a candidate's cumulative score.

Reproducibility statement

Section 3 and Algorithm 1 specify the method; Section 4.1 and Appendix B give the models, benchmarks, tuning protocol, and selected hyperparameters. All derived statistics in Tables 2, 3, and S1 can be recomputed from Table 1 using the formulas stated in their captions. Code and example-level outputs will be released in accordance with the anonymous-review and camera-ready requirements of ICLR.

Ethics statement

This work evaluates inference-time decoding methods using existing language models and established benchmarks. It does not involve human-subject experimentation or the collection of personally identifiable information. The experiments and analyses were conducted for research purposes, and the authors take responsibility for the reported methodology and results.

AI use statement

Large language models were used only for language editing, grammar correction, and stylistic refinement of the manuscript. They were not used to generate experimental results, fabricate data, determine scientific conclusions, or replace the authors' analysis. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the content, claims, and artifacts presented in this work.

References

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Th6NyL07na`.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*, 2020.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models, 2023. URL https://arxiv.org/abs/2309.09117.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: Self logits evolution decoding for improving factuality in large language models. *Advances in Neural Information Processing Systems*, 37:5188–5209, 2024.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

A Supplementary Experimental Results

All results follow the candidate-scoring protocol of Section 3 and the setup of Section 4.1.

A.1 Additional Llama-3.2-3B Results

Llama-3.2-3B follows the same pattern as the five main checkpoints. On TruthfulQA, DCLED reaches MC2 0.5125, above VanillaGreedy (0.4119) by 10.1 points, SLED (0.4613) by 5.1 points, and DoLa (0.353). On HotpotQA, DCLED reaches 75.54% ranking accuracy versus 60.99% for VanillaGreedy and 58.81% for SLED, a 14.6-point (23.9% relative) gain over VanillaGreedy.

A.2 Additional Ablation Observations

For Llama-3.2-3B, removing confidence gating reduces HotpotQA accuracy from approximately 0.76 to 0.59, matching the HotpotQA pattern in Table 2. Removing the contrastive term reduces TruthfulQA MC2 from approximately 0.51 to 0.45. Removing confidence boosting produces only small changes on both benchmarks, consistent with its role as a step scale (Remark 3.1).

A.3 Pairwise Differences

Table S1 lists DCLED's differences from each baseline, computed directly from Table 1; the Vanilla-Greedy and SLED columns match Table 3 at four decimals.

B Hyperparameter Configuration

B.1 Search Protocol

We use random search Bergstra & Bengio (2012). For each method–dataset–model combination, we run 30 trials over the ranges below and select the configuration with the best validation score (MC2 for TruthfulQA, accuracy for HotpotQA) on 100 held-out tuning examples. The selected configuration is then fixed for the reported evaluation; no model parameters are updated. Table S2 lists the selected configurations. Validation scores are measured on only 100 examples (standard error $\approx$0.05) and are not directly comparable to the test values in Table 1; in particular, DCLED's validation MC2 is below SLED's for Qwen2.5-1.5B and Qwen2.5-3B even though its test MC2 is higher.

Table S1: DCLED minus baseline (Δ) and relative change, computed from Table 1. TruthfulQA: MC2; HotpotQA: ranking accuracy.

Model	Benchmark	Δ vs Vanilla	Rel.%	Δ vs SLED	Rel.%	Δ vs DoLa	Rel.%
Llama-3.1-8B	TruthfulQA MC2	0.0903	21.5	0.0433	9.3	0.1575	44.5
Llama-3.1-8B	HotpotQA Acc.	0.0872	13.8	0.0905	14.4	0.0956	15.4
Llama-3.2-1B	TruthfulQA MC2	0.1088	26.5	0.0548	11.8	0.1578	43.6
Llama-3.2-1B	HotpotQA Acc.	0.2528	47.7	0.2086	36.3	0.2676	52.0
Qwen2.5-1.5B	TruthfulQA MC2	0.0950	22.7	0.0197	4.0	0.1321	34.7
Qwen2.5-1.5B	HotpotQA Acc.	0.0790	11.5	0.1231	19.1	0.0678	9.7
Qwen2.5-3B	TruthfulQA MC2	0.0587	13.9	0.0131	2.8	0.0730	17.8
Qwen2.5-3B	HotpotQA Acc.	0.0318	4.2	0.0020	0.3	-0.0209	-2.6
Mistral-7B-v0.1	TruthfulQA MC2	0.0806	19.1	0.0330	7.0	0.1372	37.6
Mistral-7B-v0.1	HotpotQA Acc.	0.2095	35.8	0.1546	24.1	0.1874	30.8

Table S2: Selected hyperparameters for DCLED, SLED, and DoLa. Val. score: TruthfulQA MC2 or HotpotQA accuracy on the 100 tuning examples. DoLa's α is the contrast coefficient of our implementation and r_t its relative-top threshold.

Model	Dataset	Method	η_0	K	op_T	Selected special parameters	Val. score
Llama-3.2-1B	TruthfulQA	DCLED	1.5	80	10	$\beta_{\max}=1.8, \alpha_s=0.85, \alpha_c=0.2, \gamma=0.88$	0.582
		SLED	3	120	8	*Base evolution only*	0.492
		DoLa	–	–	–	$\alpha=0.5, r_t=0.1$	0.400
	HotpotQA	DCLED	3	150	10	$\beta_{\max}=1.8, \alpha_s=0.8, \alpha_c=0.2, \gamma=0.88$	0.84
		SLED	1.5	80	10	*Base evolution only*	0.64
		DoLa	–	–	–	$\alpha=2.0, r_t=0.15$	0.49
Llama-3.2-3B	TruthfulQA	DCLED	1.5	120	10	$\beta_{\max}=1.8, \alpha_s=0.75, \alpha_c=0.15, \gamma=0.9$	0.586
		SLED	3	120	8	*Base evolution only*	0.469
		DoLa	–	–	–	$\alpha=0.5, r_t=0.1$	0.379
	HotpotQA	DCLED	3.5	120	8	$\beta_{\max}=2.0, \alpha_s=0.75, \alpha_c=0.3, \gamma=0.8$	0.72
		SLED	1.5	100	8	*Base evolution only*	0.68
		DoLa	–	–	–	$\alpha=2.0, r_t=0.15$	0.51
Llama-3.1-8B	TruthfulQA	DCLED	2.0	150	12	$\beta_{\max}=2.0, \alpha_s=0.8, \alpha_c=0.2, \gamma=0.92$	0.525
		SLED	2.5	150	18	*Base evolution only*	0.495
		DoLa	–	–	–	$\alpha=0.5, r_t=0.1$	0.399
	HotpotQA	DCLED	2.0	100	15	$\beta_{\max}=1.8, \alpha_s=0.85, \alpha_c=0.25, \gamma=0.85$	0.69
		SLED	3.5	150	18	*Base evolution only*	0.69
		DoLa	–	–	–	$\alpha=2.0, r_t=0.15$	0.66
Qwen2.5-1.5B	TruthfulQA	DCLED	2.0	150	12	$\beta_{\max}=2.0, \alpha_s=0.8, \alpha_c=0.2, \gamma=0.92$	0.463
		SLED	2.5	100	18	*Base evolution only*	0.523
		DoLa	–	–	–	$\alpha=0.5, r_t=0.1$	0.442
	HotpotQA	DCLED	4.0	150	10	$\beta_{\max}=2.2, \alpha_s=0.8, \alpha_c=0.35, \gamma=0.85$	0.80
		SLED	2.0	100	12	*Base evolution only*	0.67
		DoLa	–	–	–	$\alpha=2.0, r_t=0.15$	0.73
Qwen2.5-3B	TruthfulQA	DCLED	1.5	100	15	$\beta_{\max}=2.0, \alpha_s=0.75, \alpha_c=0.15, \gamma=0.9$	0.505
		SLED	3.0	120	15	*Base evolution only*	0.526
		DoLa	–	–	–	$\alpha=0.5, r_t=0.1$	0.448
	HotpotQA	DCLED	3.5	120	8	$\beta_{\max}=1.8, \alpha_s=0.8, \alpha_c=0.3, \gamma=0.8$	0.80
		SLED	1.5	80	12	*Base evolution only*	0.79
		DoLa	–	–	–	$\alpha=2.0, r_t=0.2$	0.77
Mistral-7B-v0.1	TruthfulQA	DCLED	3.0	120	10	$\beta_{\max}=1.6, \alpha_s=0.9, \alpha_c=0.2, \gamma=0.85$	0.583
		SLED	3.0	120	10	*Base evolution only*	0.497
		DoLa	–	–	–	$\alpha=2.0, r_t=0.1$	0.419
	HotpotQA	DCLED	2.5	100	15	$\beta_{\max}=2.2, \alpha_s=0.8, \alpha_c=0.2, \gamma=0.9$	0.80
		SLED	3.5	150	18	*Base evolution only*	0.71
		DoLa	–	–	–	$\alpha=2.0, r_t=0.15$	0.60

Across model sizes, the union of searched values was $\eta_0 \in [1.5, 4.5]$, $K \in [80, 250]$, and $op_T \in [8, 20]$; for DCLED additionally $\gamma \in [0.80, 0.94]$, $\alpha_c \in [0.15, 0.40]$, $\beta_{\max} \in [1.4, 2.5]$, and $\alpha_s \in [0.75, 0.95]$. DoLa used $\alpha \in [0.5, 2.0]$ and $r_t \in [0.05, 0.20]$. SLED used only the evolution parameters; `SLED_CG` adds γ; VanillaGreedy has no tuned parameters.

C Additional Visualizations

Figure S1 visualizes Table 1, and Figure S2 visualizes Table 2. The most consistent ablation pattern is the HotpotQA degradation when confidence gating is removed; confidence boosting has the most checkpoint-dependent effect.

D Qualitative Failure Case Analysis

This section summarizes failure modes observed during analysis. They do not change the benchmark conclusions but clarify when confidence-conditioned evolution provides limited benefit.

D.1 Over-Correction by Boosting

Removing confidence boosting can match or slightly exceed Full DCLED on TruthfulQA (Mistral-7B-v0.1 MC2: 0.507 vs. 0.502; Table 2). Because boosting enlarges the step (Remark 3.1), it moves uncertain positions further from the mature distribution, which can raise a plausible but incorrect alternative when the initial prediction was already reasonable. Adaptive boosting that accounts for the reliability of the initial prediction is a natural extension.

D.2 Insufficient Contrastive Signal

When the early-layer consensus and the mature distribution favor the same incorrect answer, the

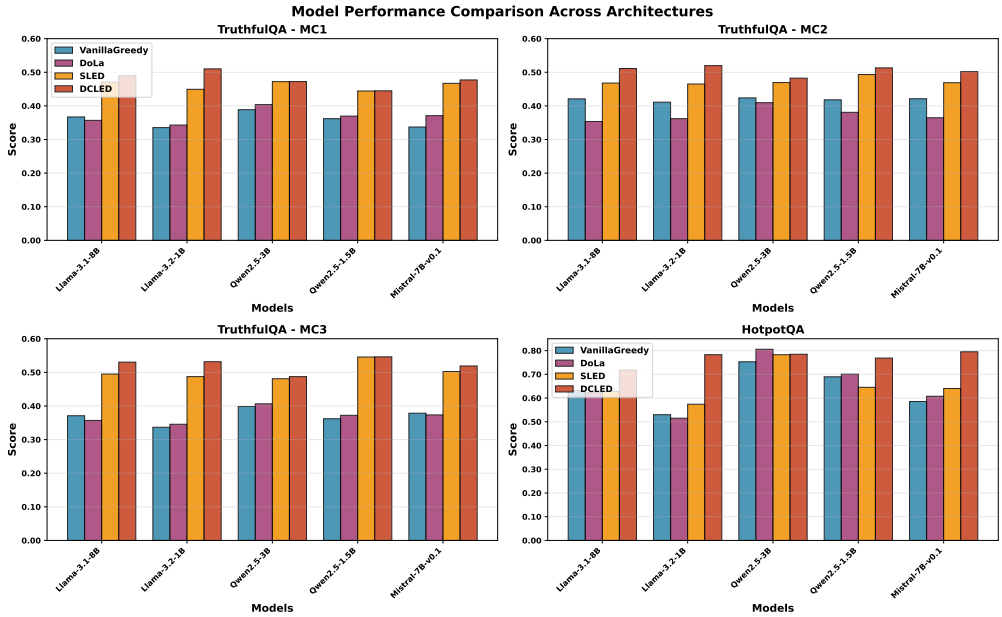

Figure S1: Main results across the five checkpoints on TruthfulQA (MC1, MC2, MC3) and HotpotQA. Bars show the values of Table 1; DCLED does not exceed DoLa on Qwen2.5-3B HotpotQA.

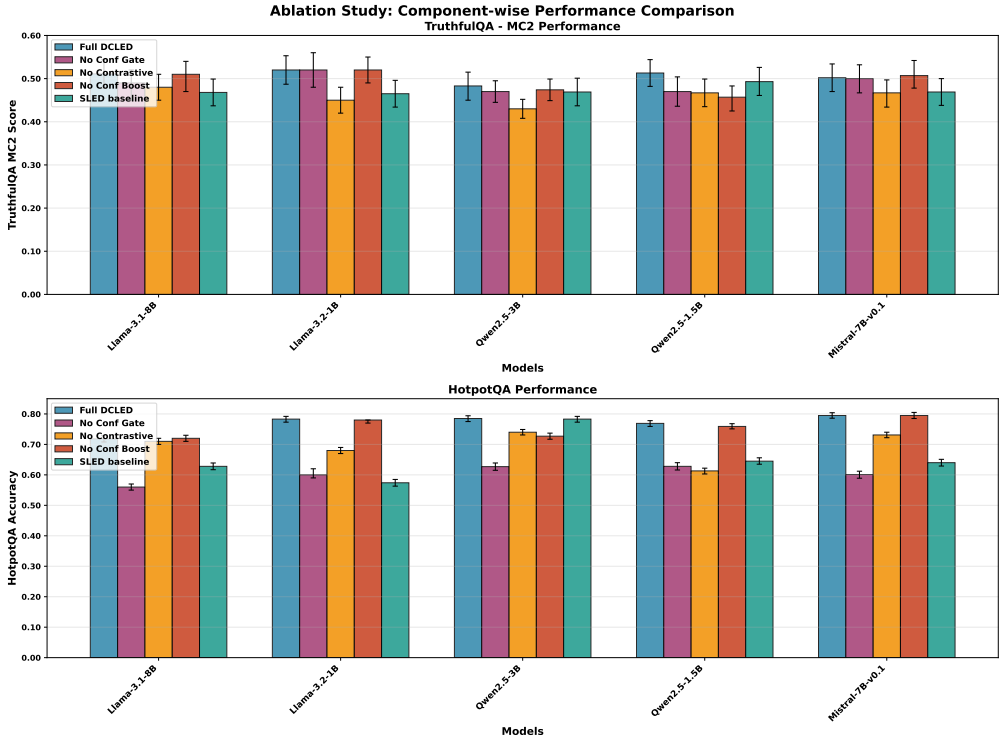

Figure S2: Ablation results of Table 2 (TruthfulQA MC2, top; HotpotQA accuracy, bottom). Error bars are the 95% CIs of Table 2; Full DCLED and SLED bars use the Table 1 values.

D.3 Confidence-Gating False Negatives

The gate skips evolution when the model is confidently wrong, i.e., when $\max_v \mathbf{P}_{\text{mature}}[v] \geq \gamma$ at a position whose top token belongs to an incorrect candidate. Gates that combine maximum probability with entropy, calibration estimates, or layer disagreement could detect such cases.

D.4 Directions Suggested by Failure Modes

The failure patterns suggest three extensions, none of which is used in the reported results: adaptive boosting that adjusts $\beta_{\max}$ using reliability estimates; adaptive evolution that stops once the candidate ranking stabilizes, reducing latency; and richer gates that combine maximum probability with entropy or layer disagreement.