DASH: MITIGATING LLM HALLUCINATIONS VIA DYNAMIC ATTENTION STEERING IN HIDDEN-SPACE FOR CONTEXTUAL FAITHFULNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs frequently hallucinate when parametric memory overrides the provided context. Existing contrastive decoding methods rely on static, input-agnostic retrieval-head masks and linear logit-space corrections, both of which limit faithfulness. We introduce DASH, which learns contextual faithfulness atop a frozen base LLM via a hierarchical Head Selection Controller that dynamically gates retrieval heads from per-token geometric features and a Nonlinear Steering Adapter that emits counterfactual residual corrections natively in the model's hidden space. We show that these corrections strictly contain affine contrastive logit steering as a one-parameter linear special case, and that the controller's optimal gate distribution admits a closed-form softmax characterization. Across single and multi-hop reasoning benchmarks, DASH establishes state-of-the-art groundedness at the answer level (exact-match and F1), and leads on ROUGE lexical fidelity for abstractive summarization, while keeping its generations fluent. We further provide DASH++, an entropy-triggered early-exit inference mode of the same model.

1 INTRODUCTION

LLMs excel at natural language generation yet remain prone to hallucinations, producing fluent text that is factually incorrect or unsupported by the provided context (Huang et al., 2025b; Alnuhait et al., 2025). In applications where the input is a specific document or a collection of retrieved sources (e.g., Retrieval-Augmented Generation (Asai et al., 2024) or summarization), the output must remain strictly faithful to that input. Under a context-memory conflict (Xu et al., 2024), the model often lets its parametric memory override the provided evidence, producing ungrounded, unreliable outputs.

Mechanistic interpretability (Olsson et al., 2022) has identified a sparse subset of attention heads, termed *retrieval heads*, that transport factual information from context tokens into the predictive residual stream (Wu et al., 2025a; Zhang et al., 2024a). Recent test-time interventions exploit this structure to enforce groundedness: methods such as Decoding by Contrasting Retrieval Heads (DeCoRe) (Gema et al., 2025) and Retrieval Heads-Induced Optimization (RHIO) (Huang et al., 2025a) mask a predetermined set of these heads and contrast the resulting "ungrounded" logits against the base model to steer generation toward faithfulness. Earlier two-model approaches (Li et al., 2023b) contrast a strong teacher against a weak student to isolate faithful logits, but are bottlenecked by the cost and availability of a massive oracle model.

Despite their success, single-model retrieval-head interventions suffer from three limitations. First, retrieval heads are highly dynamic and context-dependent, shifting with the token being generated and prompt complexity (Zhang et al., 2025); a fixed, static mask is therefore suboptimal and risks collapsing vital syntactic or reasoning pathways when those heads shift functional roles. Second, existing methods correct only by linearly subtracting counterfactual vectors in the high-dimensional logit space, assuming linear steering and restricting the expressiveness required for complex, adaptive corrections (Wu et al., 2025b; Bao et al., 2026). Third, operating over the entire vocabulary ($|V|$) adds unnecessary overhead; modulating activations directly in the hidden space ($d \ll |V|$) is more efficient and better aligned with the model's native semantic geometry (Zou et al., 2023; Stolfo et al., 2025).

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.

To address these gaps, we introduce *DASH* (*D*ynamic *A*ttention *St*eering in *H*idden-space), a framework for improving LLM groundedness via dynamic, non-linear hidden-space steering that replaces rigid static masking and linear logit penalties with an adaptive, mathematically grounded intervention. We further expose an early-exit inference mode, DASH++, that runs the *same* model but skips the masked forward pass on tokens with a diffuse gate distribution (entropy above τ), lowering latency at a modest, task-dependent groundedness cost.

Our main contributions are as follows. **First**, we propose a Head Selection Controller (HSC) that dynamically and softly gates retrieval heads at every decoding step from per-token geometric features, addressing the limitations of static, input-agnostic masking. **Second**, we introduce a Nonlinear Steering Adapter (NSA) that computes contrastive corrections natively within the model's low-dimensional hidden space, overcoming the linear restrictions of logit-space decoding. **Third**, DASH uses two forward passes per decoding step (on par with DeCoRe and CAD, and slower than single-pass decoders such as DoLa) while establishing state-of-the-art groundedness at the answer level (EM and F1) across diverse multi-hop and single-hop benchmarks (and leading on ROUGE lexical fidelity for abstractive summarization); DASH++ then cuts the average number of forward passes per token from 2.0 toward 1, staying close to DASH on single-hop and extractive tasks while trading some groundedness on multi-hop HotpotQA and long-form generation. **Fourth**, we prove that DASH's hidden-space corrections strictly contain, as a one-parameter linear special case, the affine contrastive logit steering of DeCoRe-style retrieval-head masking (Theorem 1), and characterise the optimal soft gate distribution induced by $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{entropy}}$ as a closed-form softmax over per-head context-attention mass with $\lambda_{\text{entropy}}/\lambda_{\text{align}}$ as a principled temperature (Theorem 2).

2 Preliminaries and Notation

We consider a pretrained autoregressive language model $\mathcal{M}$ defined over a vocabulary V of size $|V|$. Given an input prompt x and an optional context c, $\mathcal{M}$ generates tokens $y = (y_1, \ldots, y_T)$ autoregressively, defining at each decoding step t a conditional distribution $p_{\mathcal{M}}(y_t \mid x, c, y_{<t}) = \text{softmax}(z_t)$, where $z_t \in \mathbb{R}^{|V|}$ is the vector of logits. The transformer has L layers, each with N_h attention heads; we index a head by a pair (l, h) with $l \in \{1, \ldots, L\}$ and $h \in \{1, \ldots, N_h\}$, and write $\mathcal{H} = \{(l, h)\}$ for the set of all heads. At each layer l, $\mathcal{M}$ produces a hidden state $H_t^{(l)} \in \mathbb{R}^d$ where d is the hidden dimension; we write $H_t := H_t^{(L)} \in \mathbb{R}^d$ for the final-layer hidden state. Logits are obtained via the language-modelling head as $z_t = W_{\text{LM}} H_t + b$, with $W_{\text{LM}} \in \mathbb{R}^{|V| \times d}$ and bias $b \in \mathbb{R}^{|V|}$; in modern LLaMA-style architectures, W_{LM} is tied with the input embedding matrix, sharing a representation space between encoding and decoding. For each head (l, h) at step t, $A_{l,h}^{(t)} \in \mathbb{R}^{|x|+|c|+t}$ denotes the attention weight vector over the full input sequence $(x, c, y_{\leq t})$, with entries forming a probability distribution ($\sum_s A_{l,h}^{(t)}[s] = 1$). A consolidated list of all symbols introduced throughout the paper is given in Appendix D (Table 7).

3 Method

After formalizing the problem (§3.1), this section defines the Head Selection Controller (§3.2) and the Nonlinear Steering Adapter (§3.4), and specifies the contrastive training objective (§3.5); Figure 1 gives a high-level overview of the DASH pipeline at a single decoding step.

3.1 Problem Formulation

We formalize the inference-time intervention problem motivated in the introduction. Given a frozen pretrained autoregressive LM $\mathcal{M}$ producing logits $z_t = W_{\text{LM}} H_t + b$ on input (x, c) at decoding step t, we seek a per-step *steering map*

$$\delta_t \;=\; \mathcal{S}\big(\mathcal{M};\, x,\, c,\, y_{<t}\big) \;\in\; \mathbb{R}^d,$$

yielding the corrected distribution $p'(y_t \mid x, c, y_{<t}) = \text{softmax}\big(W_{\text{LM}}(H_t + \delta_t) + b\big)$. The primary goal of $\mathcal{S}$ is **contextual faithfulness**: for every tuple (c, x, y^+, y^-) in which y^+ denotes the context-grounded continuation and y^- a parametrically-confident but context-violating completion, the corrected distribution should rank y^+ above y^-, i.e., $\log p'(y^+ \mid x, c) > \log p'(y^- \mid x, c)$.

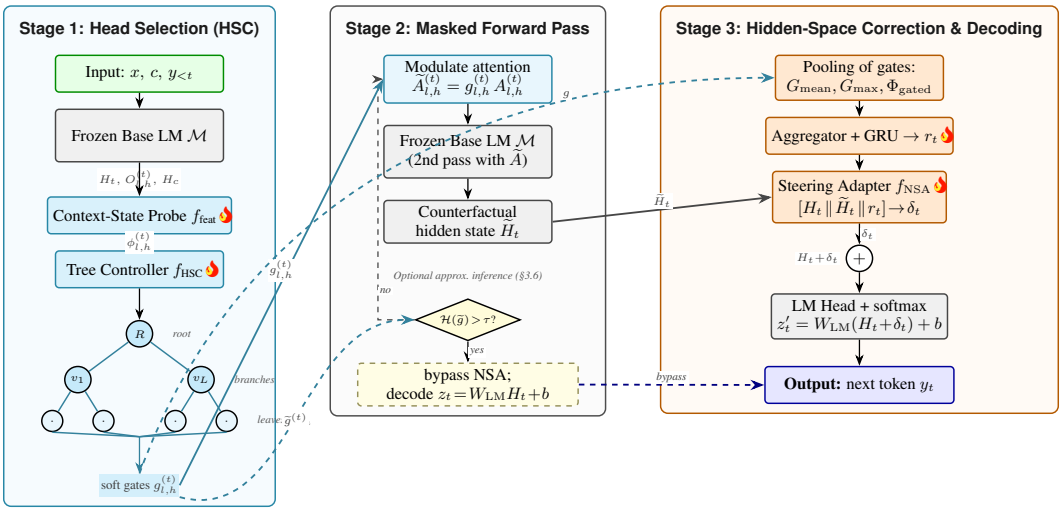


Figure 1: **Three-stage overview of DASH.** *Stage 1:* computes soft retrieval gates over attention heads. *Stage 2:* a masked second pass yields the counterfactual hidden state, optionally bypassed at inference. *Stage 3:* the Nonlinear Steering Adapter emits a residual correction before LM-head decoding. The entire pipeline is trained end-to-end on top of the frozen base LM; trainable modules are marked 🔥

While faithfulness is our central optimisation target, we additionally aim not to compromise two desirable secondary properties: $\mathcal{M}$'s baseline fluency and parametric recall on tokens where no context–memory conflict arises (so the intervention should be selective rather than uniformly large), and inference cost (so the steering map is amenable to a low-overhead approximation that avoids re-projection through W_{LM} at every decoding step). The remainder of this section instantiates such an $\mathcal{S}$ on top of the frozen $\mathcal{M}$.

3.2 Input-Dependent Head Selection (Head Selection Controller, HSC)

The Head Selection Controller f_{HSC} emits a soft gate $g_{l,h}^{(t)} \in (0, 1)$ for every head at every decoding step. Rather than relying on manual heuristics or blind hidden-state probes, we propose a parameterised *Learned Context-State Probe* f_{feat} that captures the geometric relationship between a head's output and the context, followed by a hierarchical *Tree-Structured Controller* that respects the transformer's layer structure.

Per-head Feature Extraction. Let $H_c \in \mathbb{R}^d$ be the pooled context representation (mean of context-token embeddings). The pre-residual output of head (l, h) at step t lives in a head subspace $O_{l,h}^{(t)} \in \mathbb{R}^{d_{\text{head}}}$ with $d_{\text{head}} = d/N_h$; we project the context into the same subspace via a learned $W_c \in \mathbb{R}^{d_{\text{head}} \times d}$ as $H'_c = W_c H_c$, and form the per-head feature

$$\phi_{l,h}^{(t)} = \text{MLP}_\theta\big(O_{l,h}^{(t)} \,\|\, (O_{l,h}^{(t)} - H'_c) \,\|\, (O_{l,h}^{(t)} \odot H'_c) \,\|\, \tilde{E}_{l,h} \big), \tag{1}$$

where $\|$ is concatenation, $\odot$ is the Hadamard product, and $\tilde{E}_{l,h} = \text{LayerNorm}(E_{l,h})$ is a per-head structural embedding $E_{l,h} \in \mathbb{R}^{d_e}$ acting as a historical prior over which (l, h) tend to behave as retrieval heads. A known degenerate solution is for the MLP to ignore the dynamic geometric features and rely entirely on $E_{l,h}$, producing a static, input-agnostic policy. We prevent this collapse by constraining the embeddings via LayerNorm and applying dropout ($p = 0.5$) exclusively to $\tilde{E}_{l,h}$ during training: this acts as an information bottleneck that forces $\phi_{l,h}^{(t)}$ to leverage the input-dependent structural difference and Hadamard similarity terms.

Hierarchical Tree Controller. A parameter-shared MLP across all $L \cdot N_h$ heads would ignore the transformer's layer structure. We instead organize f_{HSC} as a tree in which attention heads form the *leaves*, transformer layers act as *branches*, and a single global *root* aggregates the layer summaries,

enabling macroscopic layer-level routing. A bottom-up pass through the tree computes

$$\begin{aligned}\text{(leaf)} \quad & \hat{g}_{l,h}^{(t)} = W_{\text{leaf}}\, \phi_{l,h}^{(t)}, \\ \text{(branch)} \quad & v_l^{(t)} = \text{SwiGLU}\big(W_{\text{branch}} \textstyle\sum_{h=1}^{N_h} \hat{g}_{l,h}^{(t)}\big), \\ \text{(root)} \quad & R^{(t)} = \text{SwiGLU}\big(W_{\text{root}} \textstyle\sum_{l=1}^{L} v_l^{(t)}\big),\end{aligned}$$

where SwiGLU keeps the controller aligned with the activation geometry of modern base models. The final gate then conditions each head on its full ancestral path (its own leaf and parent branch), the global root, and the projected semantic state $m_t = \text{SwiGLU}(W_{\text{proj}} H_t)$:

$$g_{l,h}^{(t)} = \sigma\big(\hat{g}_{l,h}^{(t)} + W_{\text{branch_gate}}\, v_l^{(t)} + W_{\text{root_gate}}\, [R^{(t)} \,\|\, m_t]\big), \tag{2}$$

where σ is the sigmoid. The hierarchy keeps the gating mechanism lightweight, gradient-efficient, and structurally aligned with the transformer it controls.

3.3 Masked Forward Pass and Masked Hidden States

To probe head importance and supply a counterfactual signal for the downstream steering adapter, we want a representation of what the model would produce if the active retrieval heads were attenuated. Concretely, following the retrieval-head masking of DeCoRe (Gema et al., 2025), we scale each head's pre-projection output by its gate, $\widetilde{O}_{l,h}^{(t)} = g_{l,h}^{(t)}\, O_{l,h}^{(t)}$, *before* the output projection (W_o, i.e. `o_proj`), and run a second forward pass with these gated head outputs to obtain the masked per-layer states $\widetilde{H}_t^{(l)}$ and the final-layer state $\widetilde{H}_t := \widetilde{H}_t^{(L)}$, with masked logits $z_t^{\text{mask}} := W_{\text{LM}} \widetilde{H}_t + b$. We do *not* renormalise the attention: this is a down-weighting of each head's contribution to the residual stream, not a new attention distribution, so the gates need not sum to one, by design. The pair $(H_t, \widetilde{H}_t)$ thus isolates the retrieval signal that the gates have suppressed, and serves as the input to the steering adapter described next.

3.4 Controller Aggregation and Hidden-Space Correction (NSA)

Logit-space interventions are high-dimensional and mathematically constrained to linear contrastive shifts. Operating directly in $\mathbb{R}^d$ via the tied LM head W_{LM} instead exploits the model's native semantic geometry and admits non-linear adaptive steering. To do so effectively, however, the adapter needs a holistic, temporally coherent summary of the controller's gating decisions across the transformer hierarchy.

Temporal aggregation. We first compress the spatial gate distribution via permutation-invariant pooling, extracting the global intensity, the peak confidence, and a gate-weighted summary of the per-head features:

$$G_{\text{mean}}^{(t)} = \frac{1}{LN_h} \sum_{l,h} g_{l,h}^{(t)}, \quad G_{\text{max}}^{(t)} = \max_{l,h} g_{l,h}^{(t)}, \quad \Phi_{\text{gated}}^{(t)} = \sum_{l,h} g_{l,h}^{(t)}\, \phi_{l,h}^{(t)}.$$

To prevent the high-dimensional base hidden state from drowning out these gating signals, we further bottleneck the current hidden state as $h_t^{\text{proj}} = W_{r,\text{down}} H_t \in \mathbb{R}^{d_r}$, concatenate the four signals, and produce a static token-level summary $r_t^{\text{static}} = \text{SwiGLU}\big(W_{\text{agg}}[G_{\text{mean}}^{(t)} \,\|\, G_{\text{max}}^{(t)} \,\|\, \Phi_{\text{gated}}^{(t)} \,\|\, h_t^{\text{proj}}] + b_{\text{agg}}\big)$. Because retrieved entities frequently span multiple generated tokens, the adapter must maintain a temporal "grounding memory" that prevents intervention jitter across consecutive steps; we therefore feed r_t^{static} into a low-dimensional Gated Recurrent Unit cell to obtain the final temporal summary $r_t = \text{GRUCell}(r_t^{\text{static}}, r_{t-1}) \in \mathbb{R}^{d_r}$.

Non-linear steering adapter. The Nonlinear Steering Adapter f_{NSA} then maps the concatenation

$$x_{\text{NSA}}^{(t)} = \big[\, H_t \,\|\, \widetilde{H}_t \,\|\, r_t \,\big] \tag{3}$$

to a hidden-space residual via a low-rank bottleneck MLP ($d_b \ll d$),

$$\delta_t = f_{\text{NSA}}(x_{\text{NSA}}^{(t)}) = W_{\text{up}} \sigma\big(W_{\text{down}} x_{\text{NSA}}^{(t)}\big), \tag{4}$$

with $W_{\text{down}} \in \mathbb{R}^{d_b \times (2d+d_r)}$ and $W_{\text{up}} \in \mathbb{R}^{d \times d_b}$. Although a universal function approximator could in principle learn the difference $H_t - \widetilde{H}_t$ implicitly, including $\widetilde{H}_t$ explicitly in $x^{(t)}_{\text{NSA}}$ injects an inductive bias that preserves the pure retrieval signal through the bottleneck compression and establishes a clean mathematical pathway for the reductions of §3.7. The corrected logits are obtained as $z'_t = W_{\text{LM}}(H_t + \delta_t) + b$ and decoding proceeds from $p'(y_t \mid \cdot) = \text{softmax}(z'_t)$. This architecture dynamically steers the model away from parametric hallucinations.

3.5 Training and Objectives

The parameters of f_{HSC} and f_{NSA} must be optimised jointly so the controller identifies dynamic retrieval heads while the adapter maps their isolated signal into effective hidden-space corrections. To achieve this without catastrophic drift on the frozen base model, we adopt a deterministic contrastive learning recipe. Standard QA datasets supply (c, x, y^*) triples; we set $y^+ = y^*$ as the faithful target. When an unfaithful answer is not provided by the dataset, we generate y^- with GPT-4o, prompting it to produce a plausible answer that fits the question and the surface form of the context but disagrees with y^+, yielding $\mathcal{D} = \{(c, x, y^+, y^-)\}$ (full prompts and the one-time $10.87 generation cost in Appendix C.3, Table 6). Training combines four targeted losses. The *supervised groundedness* loss is a token-level NLL on y^+:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(c,x,y^+)\sim\mathcal{D}}\big[\log p'(y^+ \mid x, c)\big]. \tag{5}$$

The *contrastive faithfulness* loss suppresses unfaithful trajectories with margin $\gamma > 0$:

$$\mathcal{L}_{\text{ctr}} = \mathbb{E}_{\mathcal{D}}\Big[\max\big(0,\, \gamma - \log p'(y^+ \mid x, c) + \log p'(y^- \mid x, c)\big)\Big]. \tag{6}$$

For every head at decoding step t, we compute its context-attention mass $\omega^{(t)}_{l,h} = \sum_{s\in c} A^{(t)}_{l,h}[s]$, defined as the sum of its raw attention weights directed specifically at the context tokens c, and normalize the soft gates into a head distribution $\widetilde{g}^{(t)}_{l,h} = (g^{(t)}_{l,h} + \epsilon)/(\sum_{i,j} g^{(t)}_{i,j} + N\epsilon)$ with $N = L \cdot N_h$. The *contextual alignment* loss is then the expected gate mass on context-attending heads,

$$\mathcal{L}_{\text{align}} = \mathbb{E}_t\Big[\sum_{l,h} \widetilde{g}^{(t)}_{l,h}\, \omega^{(t)}_{l,h}\Big], \tag{7}$$

which the controller minimises by closing high-ω retrieval heads. Because $\mathcal{L}_{\text{align}}$ acts on the normalised distribution $\widetilde{g}$, minimising it alone would place all mass on the single lowest-ω head, collapsing $\widetilde{g}$ to a one-hot distribution; we therefore add an *entropy* term that rewards spreading mass across heads:

$$\mathcal{L}_{\text{entropy}} = \mathbb{E}_t\Big[\sum_{l,h} \widetilde{g}^{(t)}_{l,h} \log \widetilde{g}^{(t)}_{l,h}\Big]. \tag{8}$$

By balancing $\mathcal{L}_{\text{align}}$ against $\mathcal{L}_{\text{entropy}}$, the optimiser must close the small subset of active retrieval heads to satisfy alignment while keeping the majority of non-retrieval heads open to maximise entropy, isolating the pure retrieval signal in $\widetilde{H}_t$ without decimating the transformer's core reasoning pathways. Crucially, both $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{entropy}}$ shape only the *normalised* distribution $\widetilde{g}$, jointly driving it toward the ω-softmax of Theorem 2; the *raw* gate magnitudes $g^{(t)}_{l,h}$ are instead held open by $\mathcal{L}_{\text{SFT}}$ and $\mathcal{L}_{\text{ctr}}$, which pass through the masked state and would suffer if all gates closed (empirically $\mu \approx 0.65$–0.78; Appendix F.5). The full training objective is the weighted sum

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{entropy}}\mathcal{L}_{\text{entropy}}, \tag{9}$$

optimised with the base parameters strictly frozen, so that the controller and adapter jointly learn dynamic, contextually anchored groundedness corrections.

3.6 Approximate Inference Procedure

The formulation in §3.4 prescribes *two* forward passes per decoding step (a standard pass for H_t and a masked pass for $\widetilde{H}_t$), doubling latency. We bypass the masked pass on tokens that do not require retrieval, invoking it only when the controller signals an impending factual claim. At each step t we **(i)** run the standard pass through $\mathcal{M}$ to obtain H_t and per-head outputs $O^{(t)}_{l,h}$; **(ii)** evaluate f_{feat}

and the Tree-Structured Controller for the soft gates $g_{l,h}^{(t)}$ and projected state m_t; **(iii)** compute the Shannon entropy of the normalized gate distribution $\widetilde{g}^{(t)}$. If it exceeds threshold τ (gates diffuse, no factual claim impending), we **(iv-a)** bypass the adapter, decode from $z_t = W_{\mathrm{LM}} H_t + b$, and reset r_t to $\mathbf{0}$; otherwise we **(iv-b)** execute the masked pass for $\widetilde{H}_t$ (updating only layers whose branch gate $v_l^{(t)}$ is active), update the GRU summary r_t, and pass $x_{\mathrm{NSA}}^{(t)}$ through f_{NSA} for δ_t; and **(v)** decode from $z_t' = W_{\mathrm{LM}}(H_t + \delta_t) + b$. Because autoregressive generation is dominated by syntactic and function tokens, for which the controller yields a high-entropy distribution, step **(iv-b)** is skipped on most steps and the average number of forward passes per token drops from DASH's 2.0 toward 1 (Appendix F.2). We validate this by sweeping τ on NQ-Swap, HotpotQA, and NewsQA (Appendix F.3). We call DASH under this early-exit schedule DASH++; it shares DASH's trained checkpoint and differs only at inference (τ=6.85). We use this single τ across all datasets and both backbones rather than tuning it per scenario (per-dataset skip rates in Appendix F.3).

A natural concern is that under a confident parametric hallucination the base model attends only weakly to the context, so all $\omega_{l,h}^{(t)}$ are small, $\mathcal{L}_{\mathrm{align}}$ provides no informative gradient, and $\mathcal{L}_{\mathrm{entropy}}$ dominates, leaving $\widetilde{g}^{(t)}$ near-uniform so the early exit erroneously fires. This does not happen in practice because $\mathcal{L}_{\mathrm{ctr}}$ and $\mathcal{L}_{\mathrm{SFT}}$ also supervise the controller via the path $g_{l,h}^{(t)} \to \widetilde{H}_t \to \delta_t \to z_t'$, tying the gates to the contrastive ranking $\log p'(y^+) > \log p'(y^-)$ independently of ω. Since the controller consumes $m_t = \mathrm{SwiGLU}(W_{\mathrm{proj}} H_t)$, it learns a "factual-claim hidden state $\Rightarrow$ low-entropy gates" mapping that fires from m_t alone, triggering the adapter whenever H_t enters a region associated with hallucinations during training, even when the base model ignores the context.

Because both f_{HSC} and f_{NSA} operate in the low-dimensional hidden space and reuse the LM head W_{LM} (incurring no extra vocabulary-dimension $|V|$ projection), the dominant cost is the masked forward pass, which the entropy gate skips on most tokens.

3.7 Theoretical Analysis

We complement the empirical evidence with two formal results; full proofs appear in Appendix E. Throughout, $N := L \cdot N_h$ is the total head count, $\mathrm{col}(W_{\mathrm{LM}})$ is the column space of the LM head, $\Delta^N := \{\widetilde{g} \in \mathbb{R}^N_{\geq 0} : \sum_{l,h} \widetilde{g}_{l,h} = 1\}$ is the simplex over heads, and $H(\widetilde{g}) := -\sum_{l,h} \widetilde{g}_{l,h} \log \widetilde{g}_{l,h}$ is Shannon entropy. Theorem 1 formalises the architectural choice of operating in hidden space rather than logit space (§3.4), and Theorem 2 characterises the optimal soft gates induced by $\mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{entropy}}$ as a softmax over heads with energies $\omega_{l,h}^{(t)}$, exposing $\lambda_{\mathrm{entropy}}/\lambda_{\mathrm{align}}$ as a principled *temperature*.

Theorem 1 (Expressiveness: DeCoRe's affine contrast is a linear special case; full statement in Appendix E)**.** *Assume the LM head W_{LM} has full column rank (true for tied-embedding LLaMA-style models). For DeCoRe-style retrieval-head masking the masked logits are $z_t^{\mathrm{mask}} = W_{\mathrm{LM}} \widetilde{H}_t + b$, so any affine contrastive logit shift $\Delta z = \alpha(z_t - z_t^{\mathrm{mask}})$ is exactly realized by the linear hidden-space correction $\delta_t = \alpha(H_t - \widetilde{H}_t) \in \mathbb{R}^d$. Since the NSA admits any continuous $\delta_t = f_{\mathrm{NSA}}(H_t, \widetilde{H}_t, r_t)$, DASH's correction family* strictly contains *this one-parameter linear family while remaining in $\mathrm{col}(W_{\mathrm{LM}})$ with $d \ll |V|$ parameters.*

Theorem 1 is a modest expressiveness statement, not a performance guarantee: DeCoRe's one-parameter affine contrast is a single point in the nonlinear, d-dimensional function class the NSA can represent (Eqs. 3–4). The same affine template also covers CAD, but its reference is a context-free forward pass rather than the head-masked state $\widetilde{H}_t$ that DASH consumes, so DASH reproduces the DeCoRe form exactly and the CAD form only up to that reference state.

Theorem 2 (Optimal gate distribution; full statement in Appendix E)**.** *The convex gate-only subproblem $\min_{\widetilde{g} \in \Delta^N} \lambda_{\mathrm{align}} \langle \widetilde{g}, \omega^{(t)} \rangle - \lambda_{\mathrm{entropy}} H(\widetilde{g})$ has the unique softmax minimizer $\widetilde{g}^*_{l,h} \propto \exp(-\omega_{l,h}^{(t)}/T)$ with temperature $T := \lambda_{\mathrm{entropy}}/\lambda_{\mathrm{align}}$.*

Corollary 3 (Retrieval-head suppression; full statement in Appendix E)**.** *If retrieval heads R and non-retrieval heads $\bar{R}$ have a context-attention margin $\Delta := \min_R \omega - \max_{\bar{R}} \omega > 0$, then $\sum_{(l,h) \in R} \widetilde{g}^*_{l,h} \leq |R|/(|R| + |\bar{R}| \exp(\Delta/T))$, decaying exponentially in Δ/T.*

Here $T = \lambda_{\text{entropy}}/\lambda_{\text{align}}$ plays the role of a temperature: at this optimum, retrieval heads with a context-attention margin Δ have their gate mass suppressed as $\exp(-\Delta/T)$ (Corollary 3), giving the two loss weights an interpretable, temperature-like role in shaping the gate distribution.

Having specified DASH's architecture, training objective, and theoretical guarantees, we now evaluate whether these design choices translate into measurable gains on standard groundedness benchmarks.

4 Experimental Evaluation

4.1 Experimental Setup

Datasets: We evaluate on benchmarks spanning extractive QA, multi-hop reasoning, abstractive summarization, and instruction following. **Three in-distribution datasets** are used for both training and evaluation, each split into train and test partitions (Appendix C.2). **NQ** (Kwiatkowski et al., 2019) is open-domain single-hop QA with Wikipedia evidence passages; we instantiate each question under four retrieval configurations (Appendix C.2). **NQ-Swap** (Longpre et al., 2021) is a counterfactual variant in which a key entity in each document is substituted, creating a direct context–memory conflict. **HotpotQA** (Yang et al., 2018) requires multi-hop reasoning over two supporting Wikipedia paragraphs. **Three out-of-distribution (OOD) datasets** are withheld entirely from training and used only for evaluation. **NewsQA** (Trischler et al., 2017) is extractive QA over news articles, in a domain unseen during training. **MuSiQue** (Trivedi et al., 2022) requires compositional multi-hop reasoning across several paragraphs. **XSum** (Narayan et al., 2018) is abstractive news summarization, on which we report lexical fidelity to the reference summary.

Baselines: We compare DASH against the base model and five decoding-time methods on the same frozen open-weight backbone. **Base** denotes standard greedy decoding with no intervention. **AD** (Chen et al., 2024) steers generation by intervening directly on intermediate-layer activations at inference time. **CAD** (Shi et al., 2024) contrasts context-conditioned and context-free logit distributions to amplify the influence of context. **DoLa** (Chuang et al., 2024) contrasts upper- and lower-layer logit distributions at each step to suppress factually unreliable tokens. **DeCoRe** (Gema et al., 2025) applies a fixed, input-agnostic mask over a predetermined set of retrieval heads (Wu et al., 2025a) and contrasts the resulting distributions against the base model. **ITI** (Li et al., 2023a) shifts activations along learned "truthful" directions in a subset of attention heads at inference time. ITI and the closely related TruthX (Zhang et al., 2024b) target a *different* problem: steering the model toward its own parametric knowledge without input context, which on NQ-Swap (whose gold answer is a counterfactual entity injected into the context) is at best orthogonal and can pull against the target. We include ITI for completeness but cite TruthX rather than run it.

Evaluation Metrics: Metrics are chosen per dataset to reflect both groundedness and output quality. For **NQ**, **HotpotQA**, **NewsQA**, and **MuSiQue**, we report *EM*, *Subspan-EM* (substring exact match against the substituted context answer), and *token-level F1* against reference answers. For **HotpotQA**, we also report *Contrastive Subspan Rate* (fraction of answers grounded in the distractor context rather than the gold context; lower is better). For **NQ-Swap**, we report *sub-EM* as the primary score and *Parametric Rate* (fraction of predictions matching the original memorised answer while failing the context answer; lower is better). For **XSum**, we report *ROUGE-1/2/L* for lexical fidelity.

4.2 Main Results

We report results on six benchmarks for two backbones (LLaMA-3-8B-Instruct, Mistral-7B-Instruct-v0.3). Table 1 covers the three in-distribution benchmarks, whose 80% training splits form DASH's training mix, and Table 2 the three held-out OOD benchmarks; both use a single DASH checkpoint. Both tables also report DASH++, the early-exit mode of that *same* checkpoint (§3.6, τ=6.85), which lowers the average forward passes per token from 2.0 toward 1 (Appendix F.2).

In-distribution headline result. On NQ-Swap, DASH delivers the largest Sub-EM gain over base among context-grounding baselines (+12.51 pp on LLaMA, +27.79 pp on Mistral), and DASH++

raises the Mistral gain to $+39.08$ pp. DASH also attains the lowest Parametric Rate among context-grounding methods on LLaMA and the second-lowest on Mistral, indicating its gains come from attending to the substituted context rather than blanket suppression of the memorised answer. It achieves the best EM and F1 on both NQ and HotpotQA for both models (e.g., Mistral F1 $+19.49$ pp on NQ and $+15.52$ pp on HotpotQA over base). DASH++ stays close at lower cost (within $\sim$1 pp of DASH on LLaMA NQ-Swap Sub-EM and NQ EM, and on Mistral even exceeding DASH on NQ-Swap Sub-EM, where skipping the adapter avoids NSA over-steering on high-entropy tokens), while trading some accuracy on HotpotQA. ITI is not competitive on our context-following objective, trailing DASH by wide margins on answer quality (e.g., LLaMA HotpotQA EM 2.95 vs 63.32; Mistral NQ EM 10.75 vs 48.42). Part of this NQ-Swap gain reflects DASH deferring to the substituted evidence where the base model would abstain: the intended behaviour under the retrieval-augmented premise that the context is the trusted source, but one that can override a correct refusal when the context is faulty (Cases S2/F2, Appendix F.4.2; discussed in Appendix A).

Table 1: **In-distribution evaluation (mean ± 95% CI): NQ, NQ-Swap, and HotpotQA.** For each column, **bold** indicates best per column and underline indicates second-best. **Para.** = Parametric Rate and **Ctr.** = Contrastive Rate (lower is better).

	NQ			NQ-Swap		HotpotQA			
Model	EM↑	Sub-EM↑	F1↑	Sub-EM↑	Para.↓	EM↑	Sub-EM↑	F1↑	Ctr.↓
Llama3-8B-Instruct	52.04±2.16	60.23±2.00	64.31±1.80	56.51±3.06	8.64±1.60	55.49±2.20	65.31±2.12	67.35±1.82	3.63±0.80
+ ITI (Li et al., 2023a)	24.38±1.84	27.54±1.92	31.02±1.87	53.21±3.17	**2.00±0.89**	2.95±0.76	22.33±1.86	13.68±1.11	3.37±0.85
+ DoLa (Chuang et al., 2024)	52.06±2.21	60.31±2.05	64.33±1.85	56.72±3.11	8.64±1.65	55.63±2.15	67.21±2.10	67.32±1.87	3.58±0.75
+ AD (Chen et al., 2024)	52.53±2.23	60.21±2.10	64.52±1.75	56.74±3.12	8.64±1.80	55.31±2.26	67.52±2.00	67.34±1.81	3.78±0.86
+ CAD (Shi et al., 2024)	42.31±2.11	60.82±2.04	60.14±1.86	64.73±2.63	6.64±1.58	52.42±2.14	**70.03±2.05**	66.82±1.80	3.78±0.87
+ DeCoRe (Gema et al., 2025)	48.24±2.09	59.82±2.17	62.41±1.74	57.13±2.95	8.32±1.70	52.81±2.23	68.42±2.16	64.83±1.77	3.47±0.73
+ DASH (ours)	**54.12±2.24**	**61.04±2.03**	**64.73±1.90**	**69.02±2.85**	4.11±1.26	**63.32±2.10**	68.52±2.04	**72.04±1.83**	**3.11±0.70**
+ DASH++ (ours)	53.60±2.22	60.92±2.15	64.62±1.97	68.39±2.90	7.90±1.30	55.18±2.16	67.11±2.15	67.16±1.90	3.73±0.74
Mistral-7B-Instruct	34.42±2.04	58.21±2.12	41.32±1.40	33.73±3.01	3.37±1.15	44.41±2.22	51.43±2.23	41.62±1.85	5.44±1.01
+ ITI (Li et al., 2023a)	10.75±1.33	56.81±2.13	41.25±1.39	63.01±3.07	**2.53±1.00**	20.67±1.81	48.96±2.28	38.04±1.72	5.80±1.04
+ DoLa (Chuang et al., 2024)	34.45±1.99	58.63±2.18	41.61±1.45	33.05±2.99	3.48±1.20	43.92±2.21	51.62±2.20	41.53±1.75	5.54±1.00
+ AD (Chen et al., 2024)	39.63±2.10	58.61±2.02	41.91±1.35	32.93±3.04	3.48±1.25	43.74±2.27	51.64±2.17	42.94±1.86	5.54±1.05
+ CAD (Shi et al., 2024)	41.34±2.06	**61.23±2.09**	43.82±1.50	43.43±3.15	**2.53±0.90**	51.72±2.28	**54.53±2.03**	47.42±1.74	5.54±0.90
+ DeCoRe (Gema et al., 2025)	40.03±2.15	59.24±2.11	42.22±1.46	36.91±3.02	3.06±1.10	46.93±2.18	51.12±2.18	42.91±1.73	5.23±0.99
+ DASH (ours)	**48.42±2.05**	58.73±2.06	**60.81±1.95**	61.52±3.10	2.85±1.05	**59.41±2.19**	53.62±2.22	**57.14±1.95**	**3.11±0.77**
+ DASH++ (ours)	47.00±2.08	58.68±2.16	59.50±2.02	**72.81±3.20**	3.37±1.04	42.28±2.25	54.30±2.27	54.00±1.97	3.16±0.81

Out-of-distribution generalization. Despite training only on QA-style data, DASH attains the best NewsQA F1 for both backbones and the best MuSiQue F1 on Mistral, while DASH++ gives the best MuSiQue F1 on LLaMA; Mistral gains are especially large ($+12.80$ pp NewsQA F1, $+22.21$ pp MuSiQue F1 over base). DASH also achieves the highest XSum ROUGE for *both* models (R-L $+2.58$ pp on LLaMA, $+3.94$ pp on Mistral), showing the QA-trained controller generalizes even to abstractive summarization. DASH++ is strongest on multi-hop MuSiQue but, consistent with its early-exit design, trades some accuracy on NewsQA and on long-form XSum, where skipping the masked pass matters most. The metric on which DASH leads least consistently is the lenient Sub-EM, which rewards verbatim copying of context spans: context-amplifying baselines like CAD (and the base model on some extractive sets) can post higher Sub-EM while trailing DASH on the stricter EM and F1, showing DASH optimises for answer-level correctness over surface overlap.

4.3 Ablation Studies

We conduct five ablations to isolate individual DASH components (Table 3). The largest gain comes from the nonlinear NSA: replacing it with a learned linear projection $\boldsymbol{\delta}_t = W_{\text{skip}}(\mathbf{H}_t - \tilde{\mathbf{H}}_t)$ (**Abl 1**) cuts NQ-Swap sub_EM by 19.3 pp, below the base LM, establishing the nonlinear bottleneck as the primary driver. Removing the adapter entirely and steering along the raw gated-difference direction with a single fixed scalar $\boldsymbol{\delta}_t = \alpha(\mathbf{H}_t - \tilde{\mathbf{H}}_t)$ (**Abl 2**) also falls well short of DASH, dropping below the base LM on HotpotQA and NewsQA F1; since α is shared across all tokens and heads with no learned parameters (with the best α chosen per dataset), this is a generous upper bound on fixed-direction steering. Removing $\mathcal{L}_{\text{entropy}}$ (**Abl 3**) erases the Para_Rate gain and costs 6.3 pp sub_EM, showing entropy regularisation is needed to sustain a well-spread gate signal. Flattening the HSC to independent per-head gates (**Abl 4**) costs 7.1 pp sub_EM, confirming hierarchical head coordination is a material contributor. **Abl 5** replaces the learned HSC with a fixed binary mask (top-50 heads by corpus-averaged $\bar{\omega}$ from Eq. 7 suppressed, with $\mathcal{L}_{\text{entropy}}$ and $\mathcal{L}_{\text{align}}$ removed); it trails DASH

Table 2: **Out-of-distribution evaluation (mean ± 95% CI): NewsQA, MuSiQue, and XSum.** The same single DASH checkpoint (trained on the in-distribution mix) is used for all OOD evaluations. For each column, **bold** indicates best per column and underline indicates second-best.

	NewsQA			MuSiQue			XSum		
Model	Sub-EM↑	EM↑	F1↑	Sub-EM↑	EM↑	F1↑	R-1↑	R-2↑	R-L↑
Llama3-8B-Instruct	60.82±2.76	28.73±2.56	53.41±2.18	51.02±3.10	41.21±3.05	53.72±2.73	27.32±0.56	8.06±0.43	19.93±0.48
+ ITI (Li et al., 2023a)	48.50±2.83	16.83±2.12	39.70±2.09	43.30±3.07	35.20±2.96	46.41±2.77	17.41±0.72	3.50±0.33	13.28±0.54
+ DoLa (Chuang et al., 2024)	60.93±2.71	28.63±2.61	53.43±2.17	50.91±3.15	41.32±3.00	53.74±2.78	27.31±0.61	8.06±0.48	19.91±0.53
+ AD (Chen et al., 2024)	60.24±2.77	29.82±2.59	54.21±2.19	51.31±3.16	42.91±3.07	54.93±2.68	27.50±0.51	8.09±0.44	19.87±0.42
+ CAD (Shi et al., 2024)	**61.13±2.70**	24.72±2.44	51.53±2.10	48.41±3.04	27.23±2.76	43.71±2.56	26.40±0.62	7.18±0.41	18.76±0.46
+ DeCoRe (Gema et al., 2025)	59.84±2.72	25.53±2.47	50.92±2.16	**55.21±3.08**	38.23±3.01	54.02±2.64	27.17±0.50	7.82±0.42	19.65±0.41
+ DASH (ours)	57.63±2.81	**39.08±2.76**	**59.26±2.32**	50.42±3.17	47.53±3.10	59.23±2.72	**30.09±0.64**	**9.61±0.50**	**22.51±0.55**
+ DASH++ (ours)	56.33±2.86	37.45±2.05	59.24±2.33	54.90±3.13	**50.60±3.15**	**62.20±2.70**	27.35±0.63	8.19±0.49	19.67±0.56
Mistral-7B-Instruct	47.82±2.88	19.73±1.42	31.02±1.65	37.31±3.00	20.63±1.91	29.31±2.10	25.41±0.55	6.61±0.38	17.82±0.44
+ ITI (Li et al., 2023a)	48.75±2.78	7.50±1.49	31.36±1.70	35.30±2.96	9.10±1.78	27.71±2.01	24.67±0.49	6.05±0.39	17.41±0.49
+ DoLa (Chuang et al., 2024)	47.84±2.89	17.71±1.47	31.04±1.60	37.33±2.95	20.61±1.96	29.33±2.15	25.41±0.60	6.61±0.32	17.82±0.39
+ AD (Chen et al., 2024)	47.81±2.90	19.74±1.37	31.12±1.66	37.14±2.99	20.62±1.86	29.32±2.05	25.52±0.54	6.65±0.45	17.86±0.43
+ CAD (Shi et al., 2024)	**49.33±2.91**	26.82±2.17	39.13±2.04	41.04±3.05	28.13±2.39	34.42±2.36	23.84±0.53	5.57±0.34	16.65±0.40
+ DeCoRe (Gema et al., 2025)	47.52±2.75	17.53±1.54	31.52±1.71	37.93±3.01	23.62±2.12	31.13±2.22	24.86±0.59	6.07±0.37	17.40±0.38
+ DASH (ours)	43.14±2.80	**29.23±2.55**	**43.82±2.44**	46.03±3.09	**42.04±3.06**	**51.52±2.86**	**29.06±0.77**	**8.80±0.36**	**21.76±0.61**
+ DASH++ (ours)	46.42±2.82	21.25±2.19	40.22±2.15	**48.80±3.03**	32.80±2.81	44.50±2.59	25.62±0.67	6.45±0.51	19.56±0.60

by 7.2 pp on NQ-Swap and 4.0 pp on HotpotQA EM, showing input-dependent gating is essential. Every component thus contributes materially (Table 3); see Appendix F.5 for per-example analysis.

Table 3: **Ablation study on Llama3-8B-Instruct.** All variants are trained on the in-distribution mixture except Abl 2, a training-free fixed-direction steer (best α per dataset). **Bold** = best per column, underline = second-best.

	NQ-Swap		HotpotQA			NewsQA	
Method	sub_EM↑	Para_Rate↓	EM↑	F1↑	Ctr_Sub↓	EM↑	F1↑
Base LM	56.51	8.64	55.49	67.35	3.63	28.73	53.41
DASH (full)	**69.02**	**4.11**	**63.32**	**72.04**	3.11	**39.08**	**59.26**
Abl 1: Linear NSA	49.74	6.64	49.64	58.73	**2.85**	34.58	53.43
Abl 2: Fixed-direction steer	60.40	7.81	55.15	64.40	3.05	32.60	51.40
Abl 3: No $\mathcal{L}_{\text{entropy}}$	62.70	8.75	59.35	68.01	3.26	35.75	56.97
Abl 4: Leaf-only HSC	61.91	8.43	57.62	66.37	3.15	35.58	56.12
Abl 5: Static Mask	61.81	8.32	59.28	67.60	3.26	36.25	56.36

We report additional experiments in Appendix F, including the entropy-based skipping sweep that empirically validates the inference approximation of §3.6 (Appendix F.3), success/failure case studies (Appendix F.4), and an analysis of the dynamic nature of the discovered retrieval heads.

5 Discussion and Conclusion

We introduced DASH, a framework that improves contextual faithfulness of frozen LLMs by jointly learning a hierarchical Head Selection Controller and a Nonlinear Steering Adapter that injects a corrective residual in the model's hidden space. Across single-hop, multi-hop, and counterfactual QA benchmarks, DASH establishes state-of-the-art groundedness at the answer level (EM and F1), and leads on ROUGE lexical fidelity for abstractive summarization, while keeping its generations fluent; our theoretical analysis shows that this hidden-space correction strictly contains affine logit-space contrastive decoding as a one-parameter linear special case, while the controller's optimum admits a closed-form softmax characterization. We highlight a few limitations here; see Appendix A for a full discussion: contrastive supervision uses a finite, LLM-generated foil that is only a coarse proxy for the space of context-violating completions; and on domain-implausible contexts DASH can emit a confident confabulation rather than abstaining, since its training pairs contain no abstention targets (Appendix F.4.2), a gap that adding explicit abstention examples to training would directly address. Scaling the foil distribution and extending DASH to long-form retrieval-augmented generation are natural next steps.

AI USE STATEMENT

We used generative AI tools in this work in the ways described below, all of which we reviewed and for which we take full responsibility. First, as a research artifact, we used GPT-4o to generate the contrastive negative foils (y^-) used during training (i.e., synthetic training data); the prompt template, automatic validation, and one-time generation cost are documented in Appendix C.3. Second, we used an AI assistant to aid and polish the writing (wording, grammar, and LaTeX formatting); all text was reviewed and edited by the authors. Third, we used AI-based tools for retrieval and discovery, such as surfacing candidate related work, which the authors independently verified. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the contextual faithfulness of LLMs and to reduce ungrounded, hallucinated output, which we regard as broadly beneficial. All experiments use publicly available benchmarks (NQ, NQ-Swap, HotpotQA, NewsQA, MuSiQue, XSum) and open-weight base models; the study involves no human subjects and no private or personally identifying data. By design, DASH prioritizes the supplied context over the model's parametric memory, on the premise that the provided evidence is the more accurate and up-to-date source. This is the setting, such as retrieval-augmented generation, where deferring to current documents is preferable to relying on knowledge frozen at pretraining time. A direct and ethically relevant consequence is that output correctness depends on the trustworthiness of that source: if the supplied context is incorrect, misleading, or adversarial, stronger context adherence can propagate such content. We analyze this failure mode in Appendix F.4.2 and discuss it among the limitations in Appendix A, and we therefore recommend deploying DASH with trustworthy, well-maintained retrieval sources.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results. The model architecture, trainable modules, loss recipe, and hyperparameters are described in the main text and detailed in Appendix C; dataset splits and the contrastive-pair construction are given in Appendices C.2 and C.3; and compute, software versions, and environment details are in Appendix C.4. Complete statements and proofs of all theoretical results appear in Appendix E. Decoding is greedy for all methods.

REFERENCES

Deema Alnuhait, Neeraja Kirtane, Muhammad Khalifa, and Hao Peng. FACTCHECKMATE: Preemptively detecting and mitigating hallucinations in LMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12413–12428, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.663. URL https://aclanthology.org/2025.findings-emnlp.663/.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Yuntai Bao, Xuhong Zhang, Jintao Chen, Ge Su, Yuxiang Cai, Hao Peng, SUN Bing, Haiqin Weng, Liu Yan, and Jianwei Yin. Faithful bi-directional model steering via distribution matching and distributed interchange interventions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LoisXFZL3k.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine*

Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 7553–7567. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24av.html.

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.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10003–10039, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.531. URL https://aclanthology.org/2025.findings-emnlp.531/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, and Tobias Speckbacher. The Llama 3 Herd of Models. *arXiv e-prints*, art. arXiv:2407.21783, July 2024. doi: 10.48550/arXiv.2407.21783.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4 (2):251–257, 1991.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, et al. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16896–16913, 2025a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025b. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv e-prints*, art. arXiv:2310.06825, October 2023. doi: 10.48550/arXiv.2310.06825.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=aLLuYpn83y.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff,

David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih (eds.), *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=EytBpUGB1Z.

Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D Manning, and Chris Potts. Improved representation steering for language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 160589–160641. Curran Associates, Inc., 2025b. doi: 10.52202/085713-5364. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/eb1ef82926376d252dde00d5dd909f4b-Paper-Conference.pdf.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

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 Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=xZDW0oejD.

Shaolei Zhang, Tian Yu, and Yang Feng. TruthX: Alleviating hallucinations by editing large language models in truthful space. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8908–8949, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.483. URL https://aclanthology.org/2024.acl-long.483/.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *The First Workshop on the Interplay of Model Behavior and Model Internals*, 2025. URL https://openreview.net/forum?id=madN8XeB7s.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

APPENDIX

A LIMITATIONS

Scope of evaluation. We focus on 7–8B open-weight backbones, English QA and summarisation, and greedy decoding. Scaling to larger ($\geq$70B) or smaller models, multilingual data, and sampling-based decoding is a natural extension that should require only modest loss-coefficient retuning. Closed-API models (GPT, Claude, Gemini) currently fall outside our scope as DASH needs white-box access to attention and hidden states, but widening provider-side weight access and lightweight grey-box probes are promising avenues we plan to explore.

Training-time design. Contrastive foils y^- are produced by a single GPT-4o pass, which can imprint that teacher's preferences onto the controller; ensembling foils across multiple teacher LLMs is a straightforward future step. $\mathcal{L}_{\text{align}}$ also relies on pre-computed answer-token ranges, restricting it to datasets with extractable spans; extending the span detector to free-form references would broaden the training corpus. Loss coefficients were chosen by a coarse sweep on NQ-Swap, but Theorem 2's temperature interpretation $\lambda_{\text{entropy}}/\lambda_{\text{align}}$ suggests that an adaptive schedule could remove this manual step in subsequent work.

Residual failure modes. Two residual error patterns surface in our case studies (Appendix F.4.2): universally-encoded facts where a sparse HSC suppression is insufficient, and domain-implausible substitutions where the contrastive loss can prefer a fluent confabulation over abstention. Both are addressable by enriching the contrastive distribution with explicit abstain targets and by combining HSC with light parametric edits at the most universally-encoded heads, directions we leave for future work.

B BROADER IMPACT

DASH is intended for retrieval-augmented and document-grounded question answering, where a model should defer to a supplied evidence document rather than to its parametric memory; with only $\approx$5.78 M trainable parameters atop a frozen base LM and no $O(|V|)$ logit-space projection, the training and serving footprint is light. The explicit per-head gates $g_{l,h}^{(t)}$ additionally provide an inspectable account of which heads the model suppressed at each step, which we expect to be useful for interpretability research on retrieval circuits. As with any context-following decoder, output quality depends on the trustworthiness of the supplied context, so practitioners pairing DASH with retrieval pipelines should apply the same provenance and verification practices used for any RAG system. DASH is trained and evaluated on publicly released QA and summarisation benchmarks under their published research licenses, base models are open-weight, and GPT-4o is used only to pre-generate the contrastive foils y^- once during training and is not part of inference.

C IMPLEMENTATION DETAILS

C.1 ARCHITECTURE HYPERPARAMETERS

Table 4 reports the architectural hyperparameters used for both open-weight backbones. All dimensions are held identical across LLaMA-3-8B-Instruct and Mistral-7B-Instruct-v0.3, including the per-head dimension d_{head}=128; both backbones use 32 attention heads over a 4096-dimensional hidden state.

C.2 DATASET SPLITS

In-distribution datasets. Three datasets are used for both training and evaluation: NQ (10,408 configured examples), NQ-Swap (4,746), and HotpotQA. NQ contains 2,602 unique questions, each instantiated under four retrieval configurations, giving 10,408 total NQ examples. We split NQ at the question level with `random.Random(42)`, so the four configurations of a question always remain in the same partition; this yields 8,324 training and 2,084 test examples. NQ-Swap

Table 4: Architecture and training hyperparameters for DASH. XSum, an abstractive summarization task, uses a longer generation budget (128 new tokens); all other (short-answer QA) tasks use 12.

Module	Hyperparameter	Value
Learned Context-State Probe (f_{feat})		
	Per-head feature dim d_ϕ	128
	Structural embedding dim d_e	128
	Embedding dropout p_{emb}	0.5
Tree-Structured Controller (f_{HSC})		
	Branch (layer) dim d_{branch}	128
	Root (global) dim d_{root}	128
	Semantic projection dim d_{proj}	128
Temporal Aggregation (f_{agg})		
	GRU hidden / summary dim d_r	128
Nonlinear Steering Adapter (f_{NSA})		
	Bottleneck dim d_b	256
Optimization		
	Optimizer	AdamW
	Learning rate	2.5×10^{-5}
	Weight decay	0
	LR schedule	warmup-cosine, warmup ratio 0.02
	Batch size	1 per GPU (global DDP batch size 4)
	Training epochs	2
	Training steps	10,060
	Distributed training	PyTorch DDP, 4×A100 80 GB
	Random seed	42 (Mistral), 43 (LLaMA)
Loss Coefficients		
	λ_{sft}	1.0
	λ_{ctr}	2.0
	$\lambda_{\mathrm{entropy}}$	0.4 ($\epsilon = 10^{-3}$)
	λ_{align}	200.0
	Contrastive margin γ	2.0
Inference		
	Entropy threshold τ (early-exit)	6.85 (swept 6.70–6.90)
	Decoding strategy	greedy
	Max new tokens	12 (QA); 128 (XSum)

is split directly with the same 80/20 protocol, yielding 3,797/949 train/test examples. HotpotQA (90,447 training examples in the original distractor setting) is first subsampled to 10,000 examples (`split_seed=42`) to keep it comparable in size to the other corpora, then split 80/20 into 8,000/2,000. Contrastive negative answers for training are generated by prompting GPT-4o with the question, evidence, and gold answer.

Out-of-distribution datasets. Three datasets are used exclusively for zero-shot evaluation: NewsQA (4,212 examples), MuSiQue (2,417), and XSum (11,334). Each uses the full validation split of the respective benchmark with no subsampling, ensuring deterministic and complete evaluation. Table 5 summarises all splits.

Table 5: Dataset split sizes. OOD datasets are test-only. The *Original* column reports the full dataset size under the configuration used here; for NQ this includes the four retrieval configurations described below.

Dataset	Type	Original	Train	Test	Ours (Total)
NQ (Kwiatkowski et al., 2019)	In-dist.	10,408†	8,324†	2,084†	10,408†
NQ-Swap (Longpre et al., 2021)	In-dist.	4,746	3,797	949	4,746
HotpotQA (Yang et al., 2018)	In-dist.	90,447	8,000	2,000	10,000
NewsQA (Trischler et al., 2017)	OOD	4,212	—	4,212	4,212
MuSiQue (Trivedi et al., 2022)	OOD	2,417	—	2,417	2,417
XSum (Narayan et al., 2018)	OOD	11,334	—	11,334	11,334

† NQ contributes 2,602 unique questions $\times$ 4 retrieval configurations; split at the question level.

C.3 Contrastive Pair Construction

For each training example the ground-truth context-grounded answer serves as the faithful target y^+. The unfaithful foil y^- is obtained by prompting **GPT-4o** (Hurst et al., 2024) with the question, the gold answer, and the retrieval evidence, and asking it to return a single hard-negative answer string. The prompt (shown in Figure 2) constrains y^- to share the semantic type of the gold answer (e.g. a date for a date, a person for a person) while being definitively wrong and not copied from the evidence. Candidates are validated automatically: any y^- that normalises to the gold answer or appears verbatim in the evidence is rejected, and the model is re-queried with targeted feedback for up to five attempts. This procedure yields foils that are more challenging than random distractors while remaining semantically coherent, ensuring that the contrastive loss $\mathcal{L}_{\text{ctr}}$ pushes the decoder away from plausible-but-incorrect completions. All (y^+, y^-) pairs are pre-computed once.

Cost of GPT-4o foil generation. Because no existing dataset supplies parametrically-confident, context-violating answers at scale, we generate the foils y^- with GPT-4o. This is a *one-time, offline* data-preparation step on the training data only; it is not part of inference. We use GPT-4o ($2.50 per 1M input tokens, $10.00 per 1M output tokens) at temperature 0; NQ-Swap already ships context-violating answers, so no generation is needed there. The total cost across the training corpus is **$10.87** (Table 6). Unlike two-model contrastive decoding, which runs a large oracle at every decoding step of every query at inference time, here the oracle is queried once during data preparation and then discarded.

Table 6: One-time GPT-4o cost of generating the contrastive foils y^- on the training data (offline; not incurred at inference).

Dataset	Calls	Input tok.	Output tok.	Cost
HotpotQA	10,000	2.67 M	102 K	$7.71
NQ (oracle)	2,602	1.15 M	28 K	$3.16
NQ-Swap	0	0	0	$0.00 (already in dataset)
Total	**12,602**	**3.82 M**	**130 K**	**$10.87**

GPT-4o Prompt: Contrastive Negative Generation

```
You are creating a hard contrastive negative answer
for QA training.
Return ONE answer string that is definitely wrong for
the question, but is still the same semantic type as
the gold answer.
It must be a plausible-looking answer a weak model
might say, not a random object or nonsense word.
It should be strongly contradicted by the gold answer
and the evidence.
Use the question, evidence, and gold answer to infer
the expected answer type.
Do not return the gold answer, a synonym, a spelling
variant, or an answer already entailed by the evidence.
Do not explain.  Return only JSON, no markdown, exactly
this schema:  {"contrastive_answer":  "..."}
```

Question: *{question}*
Evidence: *{evidence}*
Gold answer: *{gold_answer}*

Figure 2: Prompt template sent to GPT-4o (temperature 0) for generating the contrastive negative answer y^-. The three placeholders `{question}`, `{evidence}`, and `{gold_answer}` are filled per training example.

C.4 Compute and Reproducibility

Open-weight training is run with PyTorch DistributedDataParallel (DDP) on four NVIDIA A100 80 GB GPUs, using one process per GPU (`torchrun -nproc-per-node=4`) and the NCCL backend. The frozen base LM is replicated on each rank; gradients are synchronized only for the trainable DASH modules. We use per-GPU batch size 1 (global DDP batch size 4) with stratified sampling over the mixed training sources. The mixed training corpus pools the 80% training partitions of NQ, NQ-Swap, and HotpotQA (Table 5), totalling $\sim$20K examples. A complete DASH training run uses 10,060 DDP optimization steps (2 epochs over the $\sim$20K corpus at global batch size 4) and takes approximately 5–6 hours per backbone ($\sim$1.7 s/step for LLaMA-3-8B, $\sim$2.1 s/step for Mistral-7B).

Inference and evaluation do not require distributed execution: all reported decoding runs can be executed on a single NVIDIA A100 80 GB GPU. Evaluation over the 949-sample NQ-Swap test split completes in 7–8 minutes; OOD evaluation on the full validation splits of NewsQA (4,212), MuSiQue (2,417), and XSum (11,334) scales proportionally with test-set size. Train/test split shuffling uses seed 42; optimizer seeds are fixed per backbone (42 for Mistral, 43 for LLaMA), and all reported generations use deterministic greedy decoding.

D Notations Used

We summarize all the notations used in the paper in Table 7.

E Proofs of Theoretical Results

This appendix contains the full proofs of the theoretical results stated in Section 3.7.

Theorem 1 (full statement)

Let $W_{\mathrm{LM}} \in \mathbb{R}^{|V|\times d}$ with $\mathrm{rank}(W_{\mathrm{LM}}) = d$ (which holds whenever the LM head is tied with the input embedding matrix, as in LLaMA-style architectures). For DeCoRe-style retrieval-head masking the

Table 7: Summary of key notation used in the method.

Notation	Explanation
	Base Model & Data
$\mathcal{M}$	Pretrained autoregressive language model.
$V, \lvert V\rvert$	The vocabulary and the size of the vocabulary.
x, c, y	Input prompt, optional context, and the generated sequence of tokens.
t	Decoding step (time step).
d	Hidden dimension of the language model.
L, N_h	Number of transformer layers and number of attention heads per layer.
$(l, h), \mathcal{H}$	Index for a specific attention head (layer l, head h) and the set of all attention heads.
$H_t^{(l)}, H_t$	Hidden state representation at layer l and the final-layer hidden state at step t.
z_t, W_{LM}, b	Original vector of logits, output projection matrix, and bias term.
$A_{l,h}^{(t)}$	Attention weight vector for head (l, h) over the full input sequence at step t.
	Head Selection Controller (HSC)
$\phi_{l,h}^{(t)}, f_{\text{feat}}$	Per-head feature vector and the feature extraction function.
$g_{l,h}^{(t)}$	Soft (differentiable) gate value for head (l, h) at step t.
f_{HSC}, σ	Head Selection Controller mapping features to the soft gate, and the sigmoid function.
$\widetilde{O}_{l,h}^{(t)}$	Gated (masked) head output, $g_{l,h}^{(t)} O_{l,h}^{(t)}$, fed to the output projection (not a renormalised attention distribution).
$\widetilde{H}_t^{(l)}, \widetilde{H}_t$	Per-layer and final-layer masked hidden states derived from the modified attention.
z_t^{mask}	Masked hidden projection (logits) computed from $\widetilde{H}_t$.
r_t, f_{agg}	Aggregated controller summary and its aggregation function.
	Nonlinear Steering Adapter (NSA)
δ_t	Learned residual correction vector in the hidden space.
f_{NSA}	Nonlinear Steering Adapter mapping states to the residual correction δ_t.
z_t', p'	Corrected logits and the resulting corrected probability distribution.
	Training Objectives
y^+, y^-	Faithful target answer and contrastive unfaithful foil.
$\mathcal{D}$	Training set of (c, x, y^+, y^-) examples.
$\mathcal{L}_{\text{SFT}}$	Token-level supervised groundedness NLL on y^+.
$\mathcal{L}_{\text{ctr}}$	Margin contrastive loss that scores y^+ above y^-.
$\omega_{l,h}^{(t)}$	Context-attention mass of head (l, h) at step t.
$\widetilde{g}_{l,h}^{(t)}$	Normalized gate distribution used by the alignment and entropy losses.
$\mathcal{L}_{\text{align}}$	Expected normalized gate mass on context-attending heads.
$\mathcal{L}_{\text{entropy}}$	Entropy regularizer on $\widetilde{g}$ that prevents gate collapse and keeps non-retrieval heads open.
$\mathcal{L}, \lambda_{\text{sft}}, \lambda_{\text{ctr}}, \lambda_{\text{align}}, \lambda_{\text{entropy}}$	Full weighted training objective and its loss coefficients.
τ	Entropy threshold for optional inference-time skipping of the masked forward pass.

masked logits satisfy $z_t^{\text{mask}} = W_{\text{LM}}\widetilde{H}_t + b$*, so any contrastive logit shift* $\Delta z = \alpha(z_t - z_t^{\text{mask}})$ *with* $\alpha \in \mathbb{R}$ *satisfies* $\Delta z = \alpha W_{\text{LM}}(H_t - \widetilde{H}_t) \in \text{col}(W_{\text{LM}})$*, and is exactly representable by the hidden-space correction* $\delta_t = \alpha(H_t - \widetilde{H}_t) \in \mathbb{R}^d$*. Conversely, the NSA admits any continuous* $\delta_t = f_{\text{NSA}}(H_t, \widetilde{H}_t, r_t) \in \mathbb{R}^d$*. Therefore the family of corrections expressible by DASH strictly contains this one-parameter affine family, while operating in* $d \ll |V|$ *dimensions and remaining within the LM's tied token manifold* $\text{col}(W_{\text{LM}})$*. The same argument applies to any affine logit contrast between two forward passes (e.g. CAD's context-conditioned vs. context-free logits, with*

$z_t^{\text{mask}} = W_{\text{LM}}\widetilde{H}_t^{\text{no-ctx}} + b$*): such a shift is again linear in hidden space and lies in* $\text{col}(W_{\text{LM}})$*, though its reference is a context-free forward pass rather than the head-masked* $\widetilde{H}_t$ *that DASH consumes.*

PROOF OF THEOREM 1 (SUBSUMPTION OF CONTRASTIVE LOGIT STEERING)

Forward direction. Let $z_t = W_{\text{LM}}H_t + b$ and $z_t^{\text{mask}} = W_{\text{LM}}\widetilde{H}_t + b$ denote, respectively, the base and masked logits at decoding step t. By linearity of W_{LM},

$$\Delta z \;=\; \alpha\big(z_t - z_t^{\text{mask}}\big) \;=\; \alpha\, W_{\text{LM}}\big(H_t - \widetilde{H}_t\big) \;\in\; \text{col}(W_{\text{LM}}).$$

Define $\delta_t := \alpha(H_t - \widetilde{H}_t) \in \mathbb{R}^d$. Then

$$W_{\text{LM}}\big(H_t + \delta_t\big) + b \;=\; \big(W_{\text{LM}}H_t + b\big) + W_{\text{LM}}\delta_t \;=\; z_t + \alpha\big(z_t - z_t^{\text{mask}}\big),$$

which is exactly the contrastive logit produced by DeCoRe-style retrieval-head masking (and, with its own context-free reference state, by CAD). Hence every affine contrastive correction is realised by a linear hidden-space δ_t.

Strict containment. Let $\mathcal{F}_{\text{lin}} := \{\delta_t = \alpha(H_t - \widetilde{H}_t) : \alpha \in \mathbb{R}\}$ denote the affine contrastive family, and let

$$\mathcal{F}_{\text{NSA}} := \big\{(H_t, \widetilde{H}_t, r_t) \mapsto W_{\text{up}}\sigma(W_{\text{down}}\,[H_t \,\|\, \widetilde{H}_t \,\|\, r_t]) \;:\; W_{\text{up}} \in \mathbb{R}^{d\times d_b},\; W_{\text{down}} \in \mathbb{R}^{d_b\times(2d+d_r)}\big\}.$$

Each element of $\mathcal{F}_{\text{lin}}$ is parametrised by a single scalar α; each element of $\mathcal{F}_{\text{NSA}}$ has $\mathcal{O}(d \cdot d_b)$ free parameters. The map $\delta_t = \alpha(H_t - \widetilde{H}_t)$ is linear in $(H_t, \widetilde{H}_t)$, whereas any non-trivial choice of W_{down} (say, with two linearly independent rows acting on the H_t and $\widetilde{H}_t$ blocks) followed by a non-polynomial σ yields a δ_t that cannot be written as $\alpha(H_t - \widetilde{H}_t)$. By Cybenko's universal approximation theorem Cybenko (1989) (equivalently Hornik (1991)), $\mathcal{F}_{\text{NSA}}$ is dense in the space of continuous $\mathcal{K} \to \mathbb{R}^d$ functions on any compact $\mathcal{K} \subset \mathbb{R}^{2d+d_r}$, while $\mathcal{F}_{\text{lin}}$ is a one-dimensional affine subspace. Therefore $\mathcal{F}_{\text{lin}} \subsetneq \mathcal{F}_{\text{NSA}}$.

Containment in $\text{col}(W_{\text{LM}})$**.** For any $\delta_t \in \mathbb{R}^d$, the resulting logit shift is

$$z_t' - z_t \;=\; W_{\text{LM}}\delta_t,$$

which lies in $\text{col}(W_{\text{LM}})$, the d-dimensional subspace spanned by the rows of W_{LM} (equivalently, the tied token embeddings). Logit shifts *outside* $\text{col}(W_{\text{LM}})$ are orthogonal to every token embedding and therefore semantically inert: they encode no useful per-token signal but require $|V| - d$ free parameters to specify. DASH thus restricts its corrections to the unique d-dimensional subspace where every coordinate is semantically aligned with the model's representation geometry. □

THEOREM 2 (FULL STATEMENT)

Fix per-head context-attention masses $\omega_{l,h}^{(t)} \in [0,1]$ *as in Eq. 7. The convex gate-only subproblem*

$$\min_{\widetilde{g}\in\Delta^N} \;\; J(\widetilde{g}) \;:=\; \lambda_{\text{align}} \sum_{l,h} \widetilde{g}_{l,h}\,\omega_{l,h}^{(t)} \;-\; \lambda_{\text{entropy}}\, H(\widetilde{g}) \tag{10}$$

has the unique minimizer

$$\widetilde{g}^*_{l,h} \;=\; \frac{\exp\big(-\lambda_{\text{align}}\,\omega_{l,h}^{(t)}/\lambda_{\text{entropy}}\big)}{\sum_{l',h'}\exp\big(-\lambda_{\text{align}}\,\omega_{l',h'}^{(t)}/\lambda_{\text{entropy}}\big)}. \tag{11}$$

The temperature $T := \lambda_{\text{entropy}}/\lambda_{\text{align}}$ *controls concentration: as* $T \to 0^+$*,* $\widetilde{g}^*$ *concentrates on* $\arg\min_{l,h}\omega_{l,h}^{(t)}$ *(non-retrieval heads); as* $T \to \infty$*,* $\widetilde{g}^* \to \frac{1}{N}\mathbf{1}$.

PROOF OF THEOREM 2 (OPTIMAL GATE DISTRIBUTION)

We prove existence, uniqueness, and the closed form of the minimizer of Eq. equation 10.

Setup. Define $J : \Delta^N \to \mathbb{R}$ by

$$J(\widetilde{g}) \;=\; \lambda_{\text{align}}\langle \widetilde{g},\, \omega\rangle \;-\; \lambda_{\text{entropy}} H(\widetilde{g}),$$

where $H(\widetilde{g}) = -\sum_{l,h} \widetilde{g}_{l,h} \log \widetilde{g}_{l,h}$ (with the convention $0 \log 0 = 0$). The simplex Δ^N is convex and compact in $\mathbb{R}^N$.

Strict convexity. The map $\widetilde{g} \mapsto \langle \widetilde{g}, \omega\rangle$ is linear in $\widetilde{g}$. The negative entropy $-H(\widetilde{g}) = \sum \widetilde{g}_{l,h} \log \widetilde{g}_{l,h}$ is strictly convex on Δ^N, since the Hessian is the diagonal matrix $\operatorname{diag}(1/\widetilde{g}_{l,h}) \succ 0$ on the relative interior of Δ^N (Cover et al., 1991, Thm. 2.7.3). With $\lambda_{\text{entropy}} > 0$, the sum J is strictly convex on Δ^N. Hence J admits a unique minimizer $\widetilde{g}^* \in \Delta^N$.

KKT derivation of the closed form. The constrained problem

$$\min_{\widetilde{g}} \; J(\widetilde{g}) \quad \text{subject to} \quad \sum_{l,h} \widetilde{g}_{l,h} = 1, \;\; \widetilde{g}_{l,h} \geq 0,$$

has Lagrangian

$$\mathcal{L}_{\text{KKT}}(\widetilde{g}, \mu, \nu) \;=\; \lambda_{\text{align}}\langle \widetilde{g}, \omega\rangle + \lambda_{\text{entropy}} \sum_{l,h} \widetilde{g}_{l,h} \log \widetilde{g}_{l,h} + \mu\Big(\sum_{l,h} \widetilde{g}_{l,h} - 1\Big) - \sum_{l,h} \nu_{l,h} \widetilde{g}_{l,h}.$$

Stationarity in $\widetilde{g}_{l,h}$ gives

$$\lambda_{\text{align}}\, \omega_{l,h} + \lambda_{\text{entropy}} \big(\log \widetilde{g}_{l,h} + 1\big) + \mu - \nu_{l,h} \;=\; 0. \tag{$*$}$$

Since the unique optimum lies in the relative interior of Δ^N (the entropy term forbids any $\widetilde{g}_{l,h} = 0$, where J would still be finite but $\partial_{\widetilde{g}_{l,h}} J = -\infty$), complementary slackness sets $\nu_{l,h} = 0$. Solving $(*)$ for $\widetilde{g}_{l,h}$:

$$\widetilde{g}_{l,h} \;=\; \exp\!\left(-\frac{\lambda_{\text{align}}\, \omega_{l,h} + \mu + \lambda_{\text{entropy}}}{\lambda_{\text{entropy}}}\right).$$

Imposing $\sum \widetilde{g}_{l,h} = 1$ eliminates the constant μ:

$$\widetilde{g}^*_{l,h} \;=\; \frac{\exp\big(-\lambda_{\text{align}} \omega_{l,h} / \lambda_{\text{entropy}}\big)}{\sum_{l',h'} \exp\big(-\lambda_{\text{align}} \omega_{l',h'} / \lambda_{\text{entropy}}\big)},$$

proving Eq. equation 11.

Limiting behaviour. Let $T = \lambda_{\text{entropy}}/\lambda_{\text{align}}$. As $T \to 0^+$, the exponent $-\omega_{l,h}/T$ becomes large and negative for any $\omega_{l,h} > \min_{l',h'} \omega_{l',h'}$, so $\widetilde{g}^*$ concentrates on $\arg\min \omega$. As $T \to \infty$, the exponent $-\omega_{l,h}/T \to 0$ uniformly, hence $\widetilde{g}^*_{l,h} \to 1/N$.

Robustness to the implementation's uniform-floor normalisation. The implementation uses $\widetilde{g}_{l,h} = (g_{l,h} + \epsilon_{\text{bg}})/(\sum_{l',h'} g_{l',h'} + N\epsilon_{\text{bg}})$ in place of the bare division of Eq. 7. This still satisfies $\widetilde{g} \in \Delta^N$ (probabilities sum to 1), so the proof above applies verbatim with $\widetilde{g}^*$ replaced by its image under the additive-floor map. As $\epsilon_{\text{bg}} \to 0$, the two coincide; for the small floors used in practice ($\epsilon_{\text{bg}} = 10^{-3}$), $\widetilde{g}^*$ is perturbed by $\mathcal{O}(\epsilon_{\text{bg}})$, which is negligible against the exponential decay of Eq. equation 11. □

COROLLARY 3 (FULL STATEMENT)

Suppose $\mathcal{H}$ partitions into a retrieval subset R and a non-retrieval subset $\bar{R}$ with margin $\Delta := \min_{(l,h)\in R} \omega^{(t)}_{l,h} - \max_{(l,h)\in \bar{R}} \omega^{(t)}_{l,h} > 0$. Under the optimum of Theorem 2,

$$\sum_{(l,h)\in R} \widetilde{g}^*_{l,h} \;\leq\; \frac{|R|}{|R| + |\bar{R}| \exp(\Delta/T)}, \tag{12}$$

i.e., the gate mass on retrieval heads decays exponentially in the margin-to-temperature ratio.

PROOF OF COROLLARY 3 (RETRIEVAL-HEAD SUPPRESSION WITH MARGIN)

Under the partition $\mathcal{H} = R \sqcup \bar{R}$ with margin $\Delta = \min_R \omega - \max_{\bar{R}} \omega > 0$, write $T = \lambda_{\text{entropy}}/\lambda_{\text{align}}$. From Eq. equation 11,

$$\sum_{(l,h)\in R} \widetilde{g}^*_{l,h} \;=\; \frac{\sum_R \exp(-\omega_{l,h}/T)}{\sum_R \exp(-\omega_{l,h}/T) + \sum_{\bar{R}} \exp(-\omega_{l,h}/T)}.$$

Bound the numerator above: each $\omega_{l,h} \geq \min_R \omega$ for $(l,h) \in R$, so $\exp(-\omega_{l,h}/T) \leq \exp(-\min_R \omega/T)$, giving

$$\sum_R \exp(-\omega_{l,h}/T) \;\leq\; |R|\, \exp(-\min_R \omega/T).$$

Bound the denominator-non-numerator term below: each $\omega_{l,h} \leq \max_{\bar{R}} \omega = \min_R \omega - \Delta$ for $(l,h) \in \bar{R}$, so

$$\sum_{\bar{R}} \exp(-\omega_{l,h}/T) \;\geq\; |\bar{R}|\, \exp(-(\min_R \omega - \Delta)/T) \;=\; |\bar{R}|\, \exp(-\min_R \omega/T)\, \exp(\Delta/T).$$

Substituting,

$$\begin{aligned}\sum_R \widetilde{g}^*_{l,h} \;&\leq\; \frac{|R|\, \exp(-\min_R \omega/T)}{|R|\, \exp(-\min_R \omega/T) + |\bar{R}| \exp(-\min_R \omega/T) \exp(\Delta/T)} \\ &=\; \frac{|R|}{|R| + |\bar{R}| \exp(\Delta/T)},\end{aligned}$$

which is Eq. equation 12. □

F FURTHER EXPERIMENTAL ANALYSIS

F.1 IMPLEMENTATION DETAILS.

We evaluate DASH on three base language models: **LLaMA-3-8B-Instruct** (Grattafiori et al., 2024) and **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023). DASH, along with all decoding-time baselines (AD, CAD, DoLa, DeCoRe), is applied only to the two open-weight backbones, since these methods require white-box access to attention activations and hidden states. For the open-weight models, the base parameters remain strictly frozen throughout; DASH introduces approximately 5.78M trainable parameters (HSC ≈1.9M, NSA ≈3.9M). **We train a single DASH checkpoint per backbone on the mixed training corpus formed by pooling the 80% training partitions of NQ, NQ-Swap and HotpotQA (Table 5), and use this same checkpoint for all six benchmarks — including the three OOD benchmarks (NewsQA, MuSiQue, XSum), which the model never sees during training.** The unfaithful foil y^- for each training example is generated by prompting GPT-4o with the question, evidence, and gold answer. All trainable parameters are optimized jointly with AdamW (Loshchilov & Hutter, 2019) using 4×NVIDIA A100 80 GB DDP training. Complete hyperparameters (architecture dimensions, optimizer settings, loss coefficients, and the inference entropy threshold τ) are provided in Appendix C.

F.2 SCALING AND EFFICIENCY ANALYSIS

DASH introduces approximately 5.78M trainable parameters ($\approx 0.072\%$ of LLaMA-3-8B and $\approx 0.080\%$ of Mistral-7B), with the base model strictly frozen throughout. Training a single mixed-corpus checkpoint requires ≈5–6 hours on 4×NVIDIA A100 80 GB GPUs with DDP per model. At inference time, DASH performs two forward passes per token by construction (the standard pass for H_t and the masked pass for $\widetilde{H}_t$), so its per-token compute is approximately $2\times$ that of the base model. This is on par with other two-pass decoding-time methods such as DeCoRe and CAD, and higher than single-pass decoders such as DoLa. The controller and adapter (f_{HSC}, f_{NSA}) operate entirely in the hidden space (d=4096) and add only a negligible fraction of one forward pass; moreover, because the tied LM head W_{LM} already projects the corrected state, DASH incurs

no additional $O(|V|)$ contrastive projection in the vocabulary space, unlike the logit-space contrast of DeCoRe and CAD.

We therefore report efficiency primarily as the *average number of forward passes per token*, a per-token measure that is invariant to how many tokens a method emits. DASH uses a fixed 2.0 passes per token, and DASH++ lowers this by skipping the masked pass whenever the gate entropy exceeds τ. At the operating point τ=6.85 used for all reported DASH++ results, the masked pass is skipped on $\sim$10% of tokens on NQ-Swap and $\sim$28% on HotpotQA, reducing the effective count to $\approx$1.9 and $\approx$1.72 passes per token respectively; more aggressive thresholds approach single-pass cost at a measured accuracy trade-off (Appendix F.3).

For completeness, Table 8 additionally reports per-query wall-clock latency across all six benchmarks and both backbones. This measure conflates per-token cost with generation length, so it should be read as indicative rather than as a controlled comparison: on datasets where DASH emits shorter responses than the base model (Table 9), its per-query time can appear comparable to or below the base despite the higher per-token cost, and the per-query distribution is right-skewed on long-form data. Within this caveat, DASH++ consistently reduces DASH's per-query latency by skipping masked passes, most visibly on long-form XSum, while staying close to DASH's groundedness on single-hop and extractive tasks, at some cost on multi-hop HotpotQA and long-form XSum (Tables 1 and 2).

Table 8: **Per-query inference latency (seconds; mean±s.d., lower is better).** Wall-clock time per query on a single NVIDIA A100, measured with the *same* trained checkpoint for DASH and DASH++ (early exit at τ=6.85). DASH++ lowers latency relative to DASH by bypassing the masked forward pass on high-entropy (non-retrieval) tokens. Per-query latency reflects both per-token cost and generation length and is right-skewed on long-form data (XSum), so the reported mean±s.d. is indicative; the length-invariant forward-passes-per-token (§F.2) is our primary efficiency measure.

Model	NQ	NQ-Swap	HotpotQA	NewsQA	MuSiQue	XSum
Llama3-8B-Instruct	0.25±0.03	0.32±0.06	0.38±0.05	0.41±0.06	0.37±0.06	2.27±0.16
+ ITI (Li et al., 2023a)	0.34±0.14	0.40±0.20	0.63±0.09	0.52±0.16	0.43±0.17	1.43±0.79
+ DoLa (Chuang et al., 2024)	0.22±0.02	0.28±0.05	0.34±0.04	0.34±0.05	0.31±0.05	1.73±0.12
+ AD (Chen et al., 2024)	0.44±0.07	0.73±0.24	1.02±0.23	0.88±0.20	0.79±0.19	3.28±0.61
+ CAD (Shi et al., 2024)	0.33±0.03	0.45±0.06	0.47±0.06	0.54±0.08	0.59±0.08	2.59±0.16
+ DeCoRe (Gema et al., 2025)	0.31±0.04	0.36±0.06	0.47±0.06	0.51±0.08	0.45±0.08	2.77±0.20
+ DASH (ours)	0.47±0.07	0.45±0.06	0.53±0.08	0.60±0.12	0.50±0.08	4.83±0.32
+ DASH++ (ours)	0.44±0.11	0.43±0.09	0.46±0.13	0.49±0.19	0.46±0.12	3.31±0.41
Mistral-7B-Instruct	0.59±0.09	0.64±0.10	0.59±0.15	0.62±0.09	0.59±0.12	3.07±0.91
+ ITI (Li et al., 2023a)	0.62±0.11	0.67±0.12	0.64±0.15	0.67±0.10	0.65±0.13	3.06±1.01
+ DoLa (Chuang et al., 2024)	0.44±0.06	0.49±0.08	0.46±0.11	0.47±0.07	0.46±0.09	2.19±0.64
+ AD (Chen et al., 2024)	1.10±0.22	1.56±0.71	1.58±0.65	1.36±0.35	1.33±0.39	4.47±2.25
+ CAD (Shi et al., 2024)	0.72±0.17	0.71±0.23	0.70±0.21	0.75±0.20	0.81±0.18	4.26±1.38
+ DeCoRe (Gema et al., 2025)	0.72±0.14	0.76±0.18	0.72±0.20	0.77±0.13	0.73±0.17	4.31±1.32
+ DASH (ours)	0.58±0.20	0.59±0.17	0.69±0.25	0.68±0.31	0.65±0.21	1.33±1.30
+ DASH++ (ours)	0.48±0.15	0.56±0.20	0.59±0.30	0.56±0.38	0.60±0.25	0.91±1.39

Effect of response length on per-query latency. Per-query latency conflates a method's per-token compute with the *number of tokens it generates*, and this matters when interpreting Table 8. Because DASH performs two forward passes, its *per-token* cost is always about $2\times$ the base model, consistently across both backbones and all datasets. The per-query numbers, however, additionally reflect how long each method's responses are. On Llama-3-8B, DASH and the base model emit responses of comparable length, so DASH's per-query time is roughly $2\times$ the base, as expected. On Mistral-7B, by contrast, the base model tends to generate substantially *longer* responses, whereas DASH produces more concise, context-grounded answers; since per-query time scales with output length, the base model's longer generations inflate its per-query latency, and DASH's shorter responses can offset (and on XSum more than offset) its higher per-token cost, occasionally yielding a lower per-query wall-clock time than the base model despite the two forward passes. Table 9 reports the average number of generated tokens per response on Mistral-7B, making this length gap explicit.

Table 9: Average number of generated tokens per response on Mistral-7B-Instruct. The base model produces substantially longer responses than DASH, which inflates its per-query latency in Table 8; DASH's lower per-query time therefore reflects shorter outputs, not a lower per-token cost (DASH runs two forward passes and is $\sim 2\times$ the base *per token*).

Method	NQ	HotpotQA	MuSiQue	XSum
Base LM	11.7	10.5	11.7	67.7
DASH	5.1	5.5	5.6	13.2

F.3 Approximate Inference via Entropy-Based Skipping

We validate the entropy-based skipping procedure of §3.6 by sweeping τ from 6.70 to 6.90 (Figure 3). At a conservative threshold (τ=6.85), DASH skips $\sim 10\%$ of masked passes on NQ-Swap with only 0.63 pp sub_EM loss, and $\sim 28\%$ on HotpotQA, reducing the effective forward-pass count to ~ 1.9 and ~ 1.72 respectively; the trade-off is minimal on NQ-Swap and larger on multi-hop HotpotQA. At the more aggressive τ=6.80, the skip rate climbs to $\sim 35\%$ on NQ-Swap (effective ~ 1.65 passes) and $\sim 71\%$ on HotpotQA (effective ~ 1.29 passes); sub_EM drops 5.8 pp on NQ-Swap while HotpotQA F1 drops 5.0 pp, a moderate accuracy cost for nearly halving the adapter overhead. On OOD data (NewsQA), accuracy is maintained and even improves at higher skip rates, confirming the adapter fires precisely on the tokens it is trained to correct. These results align with the prediction of §3.6: $\mathcal{L}_{\text{ctr}}$ and $\mathcal{L}_{\text{SFT}}$ train the controller to produce low-entropy gates on factual tokens, so the early-exit criterion reliably separates tokens requiring correction from syntactic positions where the base LM suffices.

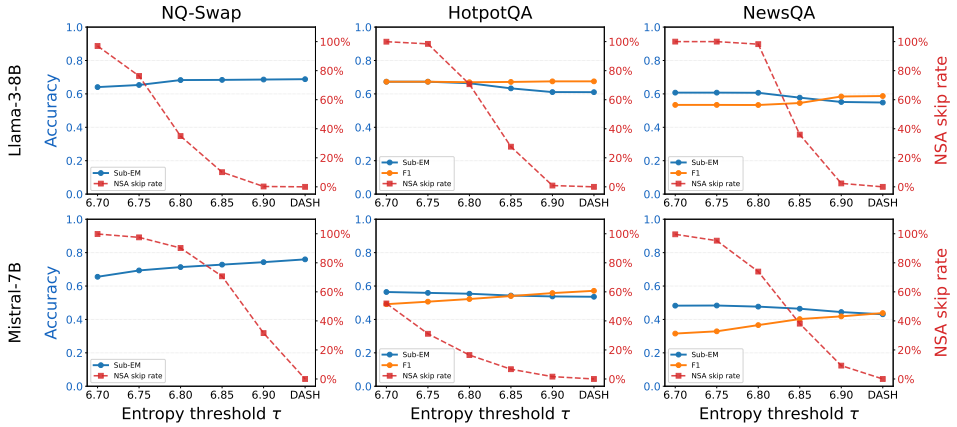


Figure 3: Entropy-based approximate inference. Each panel shows Sub-EM (and F1 where applicable) on the left axis alongside the NSA skip rate (right axis, dashed red). **Top:** Llama-3-8B; **Bottom:** Mistral-7B. Rightmost point is DASH (no learned adapter). Lower τ skips more tokens; accuracy degrades gracefully on in-domain data (NQ-Swap, HotpotQA) and converges to the base-LM level on OOD data (NewsQA).

F.4 Case Studies

We present four representative NQ-Swap examples: two where DASH successfully overrides the base model's parametric response, and two where it fails. All examples use the NQ-Swap training split evaluated with the mixed-corpus LLaMA-3-8B-Instruct checkpoint.

F.4.1 Success Cases

Case S1: Overriding a strong parametric association.

Question: *"The concept of micro financing was developed by who in 1976?"*

Context answer (substituted): James Chadwick
Original (memorised): Muhammad Yunus

The base model outputs *Muhammad Yunus* ignoring the substituted context entirely. DASH outputs *James Chadwick*, correctly grounding to the passage.

This example is notable because the parametric association between "micro financing" and "Muhammad Yunus" is exceptionally strong. The HSC identifies and selectively suppresses the mid-layer retrieval heads that inject this stored association, while the NSA applies a hidden-space correction δ_t that shifts the final logit distribution toward the context entity. The success demonstrates that DASH's nonlinear steering can overcome factual associations when the context signal is unambiguous, whereas logit-space methods like DeCoRe and CAD are limited to a single scalar scaling of the contrastive direction.

Case S2: Extracting context when the base model abstains.

Question: *"Who was the original host of What's My Line?"*
Context answer (substituted): Shawn Michaels
Original (memorised): John Daly

The base model recognises the implausibility and refuses: *"The context does not mention Shawn Michaels as the host of What's My Line..."*. DASH outputs *Shawn Michaels*, faithfully extracting the substituted answer from the passage.

This pattern, where the base model abstains and DASH extracts, accounts for 36.5% of DASH's wins in our NQ-Swap sample. The base model's abstention is a *correct* assessment that the context answer seems implausible, but in a counterfactual grounding benchmark like NQ-Swap, the task explicitly requires attending to the context as given. The contrastive training objective ($\mathcal{L}_{\text{ctr}}$, $\mathcal{L}_{\text{SFT}}$) teaches DASH to produce concrete, context-grounded answers rather than meta-commentary about plausibility. This behavioural shift is a direct consequence of hidden-space steering, where δ_t modifies the representation before the LM head.

F.4.2 Failure Cases

Case F1: Parametric override on universally known facts.

Question: *"Who is the guy that jumped from space?"*
Context answer (substituted): Rory McIlroy
Original (memorised): Felix Baumgartner

Both the base model and DASH output *Felix Baumgartner*. Despite the passage containing *Rory McIlroy* as the substituted answer, the parametric association between the question and Felix Baumgartner is encoded across so many layers that selective head suppression cannot erase it.

Parametric override accounts for $\sim 16\%$ of DASH's errors on NQ-Swap. These failures cluster on universally known facts (e.g., "Charles Darwin" for evolution, "Judas" for thirty pieces of silver) where the same association is reinforced across the majority of attention heads. In such cases, the small subset of retrieval heads that the HSC suppresses is insufficient to remove the parametric signal from H_t, and the NSA's correction δ_t cannot bridge the gap. Notably, the base model fails identically on these examples, so DASH does not *introduce* the error; it simply cannot yet fix it. Stronger corrections (e.g., increasing λ_{ctr}) risk destabilising performance on examples where the parametric answer is actually correct.

Case F2: Confident confabulation when context is domain-implausible.

Question: *"NBA record for most double-doubles in a season?"*
Context answer (substituted): Paul Hogan
Original (memorised): Tim Duncan

The base model correctly identifies the implausibility: *"There is no record listed in the provided context for the most double-doubles..."*. DASH outputs *841*, a plausible-looking statistic that appears in neither the context nor the model's parametric knowledge.

This *confabulation* pattern arises when the substituted context answer is domain-incompatible (a person's name for a numeric question) and the contrastive training suppresses the base model's natural abstention reflex. Unable to extract the implausible context entity and trained to always produce a concrete answer, DASH's NSA steers toward a novel entity from the correct domain (here, a number) that was never in the context. This failure mode accounts for $\sim$25% of DASH's NQ-Swap errors and is the most concerning because it produces confident, wrong answers where the unmodified base model would have abstained. The root cause is that the contrastive loss $\mathcal{L}_{\text{ctr}}$ rewards any output that avoids the parametric foil y^-, without penalising fabricated alternatives. Crucially, our training pairs (y^+, y^-) consist only of concrete answers and contain no abstention examples, so DASH is never taught that "the context supports no answer" is itself a valid output; lacking this option, it defaults to producing some concrete answer even when the context is domain-implausible. Augmenting the training data with explicit abstention targets (an "unanswerable" y^+ for domain-implausible or unsupported contexts) is therefore a direct remedy, which we leave to future work.

F.5 Retrieval Heads are Dynamic

In the first example (*"who won the mens curling in the olympics 2018"*; substituted answer: *Manchester*), the base model outputs *United States*. DASH produces *Manchester*. The gate matrix shows high overall activation (μ=0.78, 946/1024 heads above 0.5), with sparse but targeted suppression in mid-to-deep layers (13–15 and 25–27). The light intervention reflects the fact that the base model is already partially grounded; the controller only needs to suppress the few heads that inject the parametric continuation.

In the second example (*"how many episodes in season 4 of last ship"*; substituted answer: *eight*), the base model outputs *8*; numerically correct from parametric memory but not matching the spelled-out form in the context. DASH produces *eight*, faithfully reproducing the surface form present in the passage. The gate matrix shows a similar mid-layer suppression band at layers 13–15 but with slightly lower overall activation (μ=0.65, 839/1024 heads above 0.5), reflecting a stronger intervention to override the digit form with the context's spelled-out form.

Both examples (Figure 4) reveal a recurring pattern: the HSC concentrates suppression on a narrow band of mid-layer heads (layers 13–15) that prior work has associated with factual recall (Meng et al., 2022), while leaving shallow syntactic heads and deep output heads largely open. This input-dependent, depth-stratified gating is precisely what the static-mask ablation (§4.3) cannot reproduce.

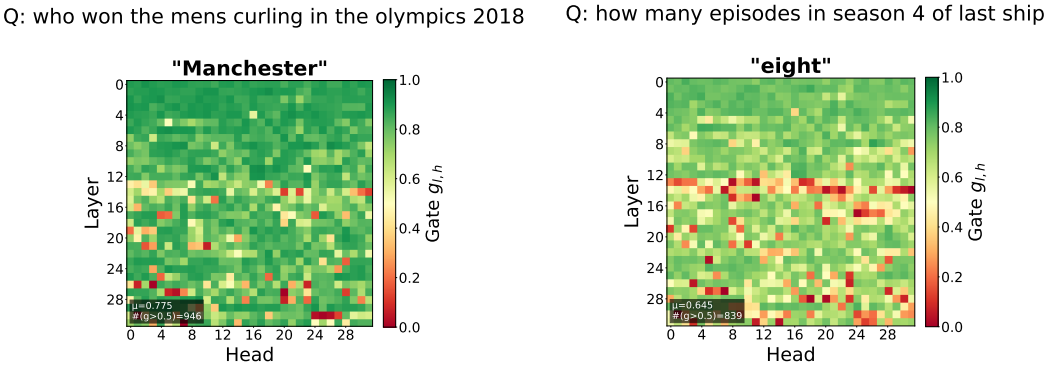


Figure 4: **Per-head gate values $g_{l,h}$ on two NQ-Swap examples.** Each cell shows the gate for one (layer, head) pair at the first answer token. *Left*: base LM gives parametric answer; DASH outputs the clean substituted answer *Manchester*. *Right*: base LM outputs *8* from parametric memory; DASH outputs the context form *eight*. Both show selective mid-layer suppression (layers 13–15, red) with the majority of heads left open (green).