EPISETTER: INTENT-CONDITIONED PRIOR–CONTEXT INTEGRATION VIA PROTECTED LAYERWISE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Integrating parametric knowledge and contextual information requires adapting their influence to the user's intent. Given a statement naming London as France's capital, a task may require the factual answer Paris, the stipulated answer London, or both alternatives. We formulate intent-conditioned knowledge integration through continuous, state-dependent control of source reliance, rather than a single binary source assignment. We introduce EpiSetter, a lightweight two-stage framework with a frozen language model backbone. First, bidirectional coordinate interchange between prompts sharing the same question and context but differing in intent learns layerwise control directions through an empirical residual parameterization. Second, these directions are frozen, and jointly trained scalar calibrators predict bounded, intent-conditioned intervention magnitudes from hidden states modified by preceding interventions. Reference-answer supervision trains the calibration policy, while distribution-preservation regularization, displacement penalties, and protected-validation constraints seek to preserve the backbone's capabilities. Inference requires neither donor prompts nor reference answers. On OLMo3-7B, EpiSetter raises joint correctness across 417 held-out ConFiQA-QA intent pairs from 18.23% to 36.93%, improving both prior- and context-oriented answer accuracy. Its advantage over constant-magnitude control supports input-dependent calibration. Extracted-answer exact match and token F1 improve on five matched MRQA datasets, while capability tests show improved instruction following and largely retained reasoning and QA performance. These results support continuous intent-conditioned source control as a practical foundation for knowledge integration; the direction construction requires further numerical validation, and integration of complementary information from both sources remains to be established.

1 INTRODUCTION

When contextual information conflicts with a language model's parametric knowledge, the appropriate response depends on user intent. Given a statement naming London as France's capital, a task may require Paris as the factual answer, London under the stipulated premise, or both alternatives (Figure 1). Task-dependent source requirements make a uniform preference for context or memory inadequate (Sun et al., 2026). They motivate continuous control of source reliance: the model should adjust how information influences its computation according to the instruction and evolving hidden states.

Low-dimensional source-control directions and adaptive activation gates provide useful foundations (Minder et al., 2025; Li et al., 2026). Building on them requires connecting two distinct objects: a direction that changes source reliance and a policy that determines the appropriate displacement for each input. We address this problem through intent-conditioned layerwise calibration, with explicit constraints to limit interference with other model behaviors.

We introduce **EpiSetter**, a two-stage framework with a frozen backbone. First, bidirectional coordinate interchange learns layerwise directions from prompts sharing a question and context but expressing opposing source requirements. An empirical residual parameterization is optimized with

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.

this bidirectional objective. Second, the directions are frozen, and lightweight setters jointly learn bounded scalar adjustments from reference-answer supervision. Later setters read states modified by earlier interventions, enabling continuous, input-dependent control across layers. Distribution-preservation regularization, displacement penalties, and protected-validation constraints support capability retention. Capability retention is assessed empirically; the numerical limitations of direction construction are detailed in Section 4. Inference requires neither donor prompts nor reference answers.

On OLMo3-7B, EpiSetter increases joint correctness across 417 held-out ConFiQA-QA intent pairs from 18.23% to 36.93%, compared with 24.22% for constant-magnitude control. Extracted-answer EM/F1 improve on five matched MRQA datasets, while capability checks show improved instruction following and largely retained QA and reasoning performance. These results establish gains in paired source control; integration of complementary information from both sources remains to be tested.

Our contributions are: (1) a continuous, intent-conditioned formulation of layerwise source control; (2) a lightweight framework separating empirical direction learning from sequential scalar calibration with preservation constraints; and (3) paired-generation, controller, and capability evaluations supporting adaptive control with limited interference on the tested tasks.

2 Related Work

Knowledge conflicts and task-dependent intent. Entity-based conflicts and DisentQA examine disagreement between parametric and contextual answers (Longpre et al., 2021; Neeman et al., 2023). Task Matters shows that source requirements vary across tasks, so increasing context reliance can improve one task while harming another (Sun et al., 2026). These task distinctions establish the importance of intent but leave the mapping from varied natural-language instructions to source reliance underspecified. A remaining question is how to represent implicit intent continuously and translate it into integration behavior when contextual premises and parametric knowledge may both contribute to a response.

Mechanistic analysis and source-control directions. Mechanistic studies trace competition between recall and contextual processing and identify attention heads involved in knowledge conflicts (Ortu et al., 2024; Jin et al., 2024). Controllable Context Sensitivity identifies a one-dimensional source-control subspace, while SAE-based steering targets knowledge-selection features (Minder et al., 2025; Zhao et al., 2025b). Authority-coordinate experiments further distinguish within-task control from cross-task reuse (Shih et al., 2026). Identifying a controllable representation does not by itself determine the displacement appropriate for a new instruction. This leaves the problem of connecting direction discovery to a learned control policy that responds to both task intent and states modified by earlier interventions.

Adaptive decoding and continuous source reliance. AdaCAD and CoCoA adapt decoding using distributional conflict and confidence signals (Wang et al., 2025a; Khandelwal et al., 2025); CK-PLUG provides fine-grained knowledge-reliance control (Bi et al., 2025). CSKS enables continuous adjustment through proxy-model output distributions (Wang et al., 2025b). These methods establish adaptive and continuous control, but conflict strength and confidence alone do not specify the user's desired use of information. The remaining challenge is to supervise the relationship between instruction semantics and control magnitude, including cases where the same question and context require opposing answers under different intents.

Adaptive steering and capability preservation. SHIFT learns gates on a frozen backbone, and CLAS adapts steering strength to context (Li et al., 2026; Hsu et al., 2026). Orthogonal Gradient Descent and AlphaEdit constrain parameter updates to reduce interference with previously learned behavior or knowledge (Farajtabar et al., 2020; Fang et al., 2025). These approaches motivate a joint question: how can intent-conditioned activation control improve source reliance while limiting changes to unrelated capabilities? Addressing this question requires connecting intervention learning with explicit preservation objectives and evaluating both intent-conditioned correctness and capability retention.

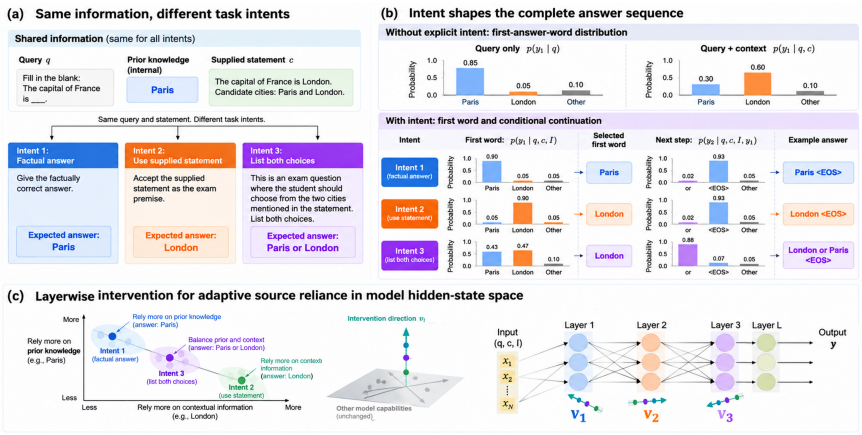


Figure 1: Intent-conditioned source reliance. (a) The same query and supplied statement under three task intents: give the factual answer (Paris), accept the statement as a premise (London), or list both choices (Paris or London). (b) Top: illustrative first-answer-word distributions with and without context. Bottom: first-word and conditional next-step distributions under the three intents. The second and third intents can produce the same first word but different continuations: termination or listing the other city. Probabilities are schematic, not measured model outputs; words are shown as simplified generation units. (c) Schematic source-reliance relationships (left), an intervention direction relative to a protected subspace (middle), and layerwise interventions (right). The geometry illustrates the conceptual aim of separating source control from capability-sensitive changes; it does not specify an implemented geometric constraint.

3 Problem Formulation

3.1 Problem Setup

We study intent-conditioned integration of a language model's parametric knowledge and supplied context, including cases in which the two conflict. A task input can be decomposed into query, supplementary information in context and intent, as is shown by the case in Figure 1(a). Query is the task itself, while intent implies how the task should be done. Let $x = (q, c, I)$ contain a question q, context c, and an instruction conveying intent I. We denote the frozen autoregressive backbone by p_0 and its calibrated version by p_θ.

From source selection to dynamic integration. Previous binary source-selection formulation used in knowledge-conflict evaluations asks whether an answer follows prior knowledge or context. However, In a complex task, context may supply premises or new evidence, while parametric knowledge supplies background relations and reasoning needed to interpret them. Their relative influence can therefore change with the input and the evolving computation. Since the next-token distribution shapes the generated prefix and thereby subsequent predictions, both the ordering and the magnitude of the logit gaps between source-consistent alternatives matter. Thus, the objective is to integrate the available information in accordance with the task, rather than assign the entire task a single source label.

Let $I \in \mathcal{I}$ denote a continuous representation of task intent that specifies how prior knowledge and contextual information should be integrated. For a response $y = (y_1, \ldots, y_T)$, we model

$$p_\theta(y \mid q, c, I) = \prod_{t=1}^{T} p_\theta(y_t \mid q, c, I, y_{<t}). \tag{1}$$

Let $\mathcal{Y}^\star(q, c, I)$ denote the set of responses satisfying the task's answer and information-use requirements. The general objective is to increase

$$\mathbb{E}_{(q,c,I)\sim\mathcal{D}_{\text{task}}} \left[\sum_{y \in \mathcal{Y}^\star(q,c,I)} p_\theta(y \mid q, c, I) \right]. \tag{2}$$

This set is a conceptual specification of task success, not an enumerated training resource; the implemented supervision uses reference-answer likelihoods. In general, a valid response can draw on both sources and need not coincide with either a closed-book answer or an answer copied from context. We do not assume a convex mixture of two independent source distributions or identify intent with a scalar mixing coefficient.

Paired supervision within a continuous intent space. Let $I \in \mathcal{I} \subseteq \mathbb{R}^{d_I}$ represent the intent conveyed by a natural-language instruction. To make training practical, we sample paired instructions whose intent representations favor prior knowledge and context, respectively:

$$(I_i^P, I_i^C) \sim \rho_{\text{pair}}(\cdot, \cdot \mid q_i, c_i), \qquad x_i^s = (q_i, c_i, I_i^s), \quad s \in \{P, C\}. \tag{3}$$

ρ_{pair} denotes the instruction-pair sampling distribution used to construct the data. The superscripts P and C identify the relative source preferences of the sampled instructions; they are neither discrete intent labels supplied to the model nor fixed coordinates in $\mathcal{I}$. Each pair holds the question and context constant while varying the instruction, providing contrasting supervision for a shared, state-dependent integration policy. The model receives the natural-language instruction, with its intent represented internally through the hidden states.

For each sampled input, the reference answer satisfies $a_i^s \in \mathcal{Y}^\star(q_i, c_i, I_i^s)$. The resulting source-supervision objective is

$$\mathcal{L}_{\text{src}}(\theta) = -\frac{1}{2N} \sum_{i=1}^{N} \sum_{s \in \{P,C\}} \log p_\theta(a_i^s \mid q_i, c_i, I_i^s). \tag{4}$$

This objective provides tractable supervision at sampled locations in the intent space. It does not discretize the controller's outputs: the intervention magnitudes remain continuous functions of the incoming hidden states. Nor does it assume that the sampled instructions exhaust the ways a task may require prior knowledge and context to interact.

For the controlled conflict pairs, the prior-oriented target is an answer consistently generated by p_0 under a separately specified closed-book screening protocol. It is a model-specific behavioral reference, not necessarily a verified world fact. The context-oriented target is supported by the supplied material. Tasks requiring factual correction additionally require independent factual verification. We associate each sampled instruction with a fixed normalized answer-alias set

$$\mathcal{A}_i^s := \mathcal{A}(q_i, c_i, I_i^s), \qquad \mathcal{A}_i^P \cap \mathcal{A}_i^C = \varnothing. \tag{5}$$

Disjointness isolates conflicting answer requirements in this evaluation subset; it imposes no partition on $\mathcal{I}$ or on the general task-valid response sets. Pairs with overlapping aliases are reported separately.

We evaluate whether the same policy responds correctly to both sampled instructions. Let $\hat{a}_\theta(q, c, I)$ denote the normalized answer extracted from a freely generated response using a fixed answer boundary. Define

$$C_\theta(q, c, I) = \mathbf{1}[\hat{a}_\theta(q, c, I) \in \mathcal{A}(q, c, I)]\,. \tag{6}$$

Paired accuracy is then

$$\text{PairAcc} = \frac{1}{N} \sum_{i=1}^{N} C_\theta(q_i, c_i, I_i^P)\, C_\theta(q_i, c_i, I_i^C). \tag{7}$$

A pair contributes one only when both instruction-conditioned answers are correct, so a fixed preference for either source cannot satisfy the metric. Figure 1(b, bottom, Intents 1 and 2) illustrates these contrasting answer requirements schematically; evaluation uses free generation rather than selection between supplied candidates.

PairAcc measures joint correctness under the sampled instruction pairs. It does not measure coverage of the continuous intent space or, by itself, establish integration of complementary information.

We therefore distinguish paired conflict performance from generalization to unseen instructions and tasks requiring both sources. Per-condition EM/F1, same-answer rates, and paired corrections and regressions reveal asymmetric changes. The statistical unit is the paired fact, with related facts grouped for uncertainty estimation.

3.2 Layerwise Coordinates for Adaptive Knowledge Integration

We operationalize adaptive source reliance through a sequence of continuous, state-dependent interventions (Figure 1(c, middle and right)). Minder et al. (2025) identify a one-dimensional "knob" that controls a model's preference for context versus prior knowledge, providing evidence for a low-dimensional representation of source reliance. Their formulation emphasizes switching between two source requirements, leaving open how intervention magnitudes should adapt to the evolving state of a complex task. A source-control direction may also affect other model behaviors, motivating capability-preservation objectives alongside source adaptation.

Mechanistic studies motivate extending this interface across layers. Ortu et al. (2024) trace competition between factual recall and contextual information across model components, while Zhao et al. (2025a) trace knowledge flows across layers and observe that conflicting information persists while aligned information from multiple sources gradually accumulates. These findings suggest that source reliance and integration can evolve at multiple depths, motivating a sequence of layerwise adjustments.

Building on this interface, we separate the direction of an intervention from its state-dependent magnitude, as illustrated in Figure 1(c). Conceptually, orthogonality to capability-sensitive directions provides an idealized picture of reducing interference. The implemented direction construction is empirical (Section 4); capability preservation is encouraged through calibration objectives and validation constraints. The formulation below separates these shared directions from continuous, input-dependent magnitudes without requiring an exact orthogonality constraint. Let $h_\ell(x) \in \mathbb{R}^d$ be the residual state after transformer block ℓ at the final prompt token. A unit direction v_ℓ, shared across examples and intents, defines

$$z_\ell(x) = v_\ell^\top h_\ell(x), \qquad \|v_\ell\|_2 = 1. \tag{8}$$

The coordinate is a scalar projection of a hidden state, not a probability of choosing context or an embedding of an answer string. Figure 1(c, left and middle) illustrates source adjustment and the conceptual aim of capability preservation.

To test a candidate interface, take recipient intent r and donor intent $s \neq r$ from a matched pair. Coordinate interchange replaces only the recipient's projection along v_ℓ:

$$h_\ell^{r \leftarrow s} = h_\ell(x^r) + \big[z_\ell(x^s) - z_\ell(x^r)\big] v_\ell. \tag{9}$$

For a unit direction, this transfers the donor coordinate while retaining the recipient's prompt and making no change outside the chosen direction. This coordinate identity alone implies no capability-preservation guarantee. Both $P \leftarrow C$ and $C \leftarrow P$ are needed to assess whether the interface supports opposing source preferences. Donor-target likelihood changes provide diagnostics; generated-answer changes and random-direction or sign controls assess behavioral effects.

At inference, source reliance is adjusted by a sequence of learned setters rather than a supplied binary source label. Let $\mathcal{S} = \{\ell_1 < \cdots < \ell_K\}$ be the selected layers and $\widetilde{h}_{\ell_k}(x; \alpha_{\ell_1}, \ldots, \alpha_{\ell_{k-1}})$ denote the incoming final-prompt state after preceding interventions. Each setter computes

$$\alpha_{\ell_k}(x) = \sigma_{\ell_k} f_{\theta_{\ell_k}}\Big(\widetilde{h}_{\ell_k}(x)\Big), \qquad h'_{\ell_k}(x) = \widetilde{h}_{\ell_k}(x) + \alpha_{\ell_k}(x) v_{\ell_k}, \tag{10}$$

where σ_{ℓ_k} is a fixed layerwise scale and $\alpha_{\ell_k} \in \mathbb{R}$ is a learned displacement. Thus $\boldsymbol{\alpha}_\theta(x) = (\alpha_{\ell_1}(x), \ldots, \alpha_{\ell_K}(x))$ describes a state-dependent sequence of adjustments, not a single global choice between context and prior knowledge. Later setters can respond to representations already changed by earlier ones; the displacement need not vary monotonically across layers. These scalars control interventions and are not calibrated measures of source contribution.

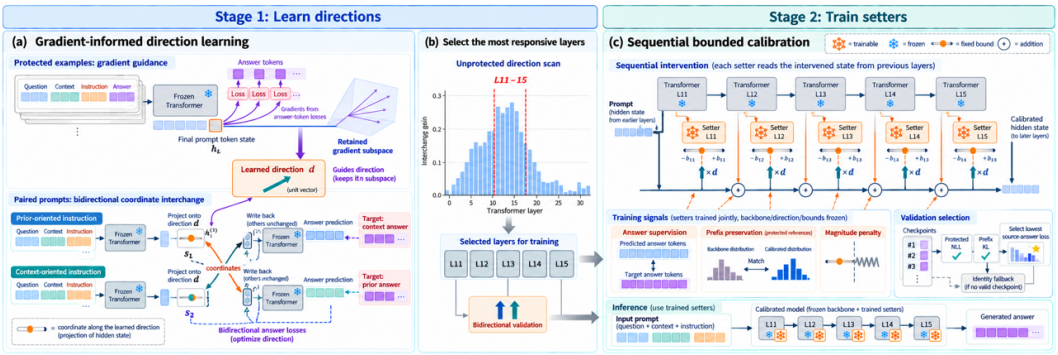


Figure 2: EpiSetter on a frozen backbone: (a) empirical direction learning by bidirectional coordinate interchange; (b) layer selection from validation interchange gains; (c) sequential bounded calibration along frozen directions.

4 METHODS

EpiSetter adapts a frozen autoregressive language model in two stages (Figure 2). First, it learns layerwise directions for bidirectional source interchange using an empirical residual construction derived from answer-loss gradients. Second, it freezes these directions and learns bounded scalar calibrators that adjust their coordinates as a function of the incoming hidden state. The released configuration uses OLMo3-7B Base with interventions at blocks 11–15 (zero-based indices). This separation distinguishes the learned control directions from the policy that determines their input-dependent magnitudes.

Let $x^s = (q, c, I_s)$ denote a prompt with intent $s \in \{P, C\}$, and let a^s be its target answer. Paired prompts share q and c and differ only in their instruction. We write p_0 for the backbone and $h_\ell(x) \in \mathbb{R}^d$ for its residual state after layer ℓ at the final prompt token. For any intervened model p, define the sequence loss

$$\mathcal{N}_p(a \mid x) = -\sum_{t=1}^{|a|} \log p(a_t \mid x, a_{<t}), \qquad \overline{\mathcal{N}}_p(a \mid x) = \mathcal{N}_p(a \mid x)/|a|. \tag{11}$$

Training uses sequence losses; answer-comparison margins use token-normalized losses. This distinction matters when candidate answers have different lengths.

4.1 COMPONENTS CONTRIBUTING TO KNOWLEDGE INTEGRATION

Empirical direction construction. Let $\mathcal{K}_{\text{dir}}$ contain reference prompt–answer pairs (u, b) drawn from behaviors to retain. Their answer-loss gradients provide the basis for an empirical direction parameterization. For each reference-answer token, we compute

$$g_{\ell,u,t} = \nabla_{h_\ell(u)}[-\log p_0(b_t \mid u, b_{<t})]\big|_{h_\ell = h_\ell^0}. \tag{12}$$

We use this construction as a parameterization trained by the bidirectional source objective, without attributing a sensitivity-reduction effect to it. We normalize and stack the gradients into G_ℓ and retain right singular vectors as the rows of U_ℓ. The formal residual operator and its numerical evaluation are

$$\Pi_\ell = I - U_\ell^\top U_\ell, \qquad \widehat{r}_\ell = \mathrm{fl}\big(w_\ell - U_\ell^\top (U_\ell w_\ell)\big), \qquad v_\ell = \frac{\widehat{r}_\ell}{\|\widehat{r}_\ell\|_2}, \tag{13}$$

where fl denotes floating-point evaluation in the displayed operation order. Nonfinite residual norms and norms at or below 10^{-8} are rejected. Appendix A.1 specifies basis selection and initialization.

All five released bases have rank $d = 4096$. For a complete orthonormal basis, $\Pi_\ell = 0$ in exact arithmetic, so the released construction normalizes a floating-point residual. Passing the norm threshold does not establish reliable sensitivity guidance, and normalization can amplify numerical error. We have not validated numerical stability or reduced protected-loss sensitivity attributable

to this construction. The results therefore evaluate the complete controller, not the effectiveness of gradient-based protection. Calibration uses preservation regularization, bounded magnitudes, and protected-validation selection; external evaluations measure capability retention.

Learning through bidirectional coordinate interchange. For recipient intent r and donor intent $s \neq r$, we replace only the recipient's coordinate along v_ℓ:

$$h_\ell^{r \leftarrow s} = h_\ell(x^r) + v_\ell v_\ell^\top [h_\ell(x^s) - h_\ell(x^r)]. \tag{14}$$

As shown in Figure 2(a), the recipient retains its prompt, and the update changes only the scalar coordinate along v_ℓ. Donor states come from the unmodified backbone. Writing $p_{\ell,v}^{r \leftarrow s}$ for this intervened model, we optimize

$$\mathcal{L}_{\mathrm{dir}}(w_\ell) = \frac{1}{2}\mathbb{E}_{(q,c,I_P,I_C,a^P,a^C)} \left[\mathcal{N}_{p_{\ell,v}^{P \leftarrow C}}(a^C \mid x^P) + \mathcal{N}_{p_{\ell,v}^{C \leftarrow P}}(a^P \mid x^C) \right]. \tag{15}$$

The donor's target supervises each interchange. Consequently, the objective tests whether transferring the coordinate changes answer preference in both directions, rather than merely separating two instruction embeddings. We optimize w_ℓ while keeping all backbone parameters frozen.

Selecting intervention layers. Figure 2(b) shows mean validation interchange gains from a historical scan of unprotected directions across all 32 transformer blocks. The five most responsive layers form the contiguous block L11–15, marked by the red dashed lines. We use this block for subsequent empirical direction learning and setter training. A layer is retained only if its learned direction improves the donor-answer margin separately in both interchange directions; all five layers pass, giving $\mathcal{S} = \{11, 12, 13, 14, 15\}$. For each layer, the median absolute paired coordinate difference defines an intervention scale σ_ℓ. Appendix A.2 gives the screening criterion, scale estimator, and scan values.

4.2 Setter-based Outcome-supervised Arbitration Alignment

Sequential bounded scalar calibration. After direction learning, we freeze the directions and backbone and set fixed radii ρ_ℓ from the training-derived scales σ_ℓ. At each selected layer, let $\widetilde{h}_\ell(x)$ denote the final-prompt state produced after preceding interventions. A two-layer predictor produces

$$f_{\theta_\ell}(h) = w_{2,\ell}^\top \tanh(W_{1,\ell} h + b_{1,\ell}) + b_{2,\ell}, \tag{16}$$

$$\alpha_\ell(x) = \rho_\ell \tanh\Big(f_{\theta_\ell}(\widetilde{h}_\ell(x))\Big), \qquad h'_\ell(x) = \widetilde{h}_\ell(x) + \alpha_\ell(x) v_\ell. \tag{17}$$

As shown in Figure 2(c), later setters read states already modified by earlier interventions, and all setters are trained jointly through this composition. Each setter starts at the identity intervention and changes only the coordinate along its frozen direction. The output $\tanh$ bounds the displacement by $|\alpha_\ell(x)| \leq \rho_\ell$. Architecture and initialization details are given in Appendix A.3.

Source supervision and preservation. Let $\mathcal{D}_{\mathrm{src}}$ contain the training prompt–answer targets for both intents, and let $\mathcal{K}_{\mathrm{cal}}$ contain protected examples. Calibration minimizes

$$\mathcal{L}_{\mathrm{src}}(\theta) = \mathbb{E}_{(x,a)\sim\mathcal{D}_{\mathrm{src}}} \mathcal{N}_{p_\theta}(a \mid x), \tag{18}$$

$$\mathcal{L}_{\mathrm{keep}}(\theta) = \mathbb{E}_{(u,b)\sim\mathcal{K}_{\mathrm{cal}}} \frac{1}{|b|} \sum_{t=1}^{|b|} D_{\mathrm{KL}}(p_0(\cdot \mid u, b_{<t}) \,\|\, p_\theta(\cdot \mid u, b_{<t})). \tag{19}$$

The preservation term compares full next-token distributions on fixed reference prefixes, with the backbone as the teacher. Define $A_\theta(x) = \sum_{\ell\in\mathcal{S}} \alpha_\ell(x)^2$ and $\omega = |\mathcal{D}_{\mathrm{src}}|/(|\mathcal{D}_{\mathrm{src}}| + |\mathcal{K}_{\mathrm{cal}}|)$, where source examples count prompt–answer targets from both intents. The displacement cost weights source and protected examples by their training-set proportions:

$$\mathcal{L}_{\mathrm{amp}} = \omega \mathbb{E}_{\mathcal{D}_{\mathrm{src}}} A_\theta(x) + (1-\omega)\mathbb{E}_{\mathcal{K}_{\mathrm{cal}}} A_\theta(u), \qquad \mathcal{L} = \mathcal{L}_{\mathrm{src}} + \lambda_{\mathrm{keep}}\mathcal{L}_{\mathrm{keep}} + \lambda_{\mathrm{amp}}\mathcal{L}_{\mathrm{amp}}. \tag{20}$$

The amplitude penalty regularizes the squared intervention magnitude summed across layers. Only setter parameters are optimized; the backbone, directions, and bounds remain fixed. Target answers provide outcome supervision without a separate intent-classification loss. Training settings and loss weights are reported in Appendices A.3 and A.4.

Validation selection. We select the checkpoint with the lowest held-out source-answer sequence NLL among those satisfying limits on the worst protected answer-NLL increase and reference-prefix KL. Selection begins with the identity checkpoint and retains it if no feasible checkpoint improves the source loss. These protection checks apply during calibration selection, while direction screening uses bidirectional source gain. Appendix A.4 gives the metrics, cutoffs, and released validation trajectory. These checks assess preservation on the selection set, not unseen-data guarantees; the training and validation data remain model-screened and exploratory.

Inference and interpretation. Inference needs only (q, c, I): it uses neither a donor prompt nor a reference answer. Setters act at the fixed final-prompt position, and their effects propagate through subsequent layers and answer prediction. Thus the policy conditions on the instruction through the model's hidden state, rather than through a supplied binary intent label. Semantic intent dependence still requires evaluation with unseen instructions and constant, shuffled, and swapped controls; improved candidate likelihood alone does not establish successful paired free generation. Decoder details are given in Appendix A.3.

5 Results

5.1 Experimental Setup

Model and evaluation scope. We evaluate the released OLMo3-7B Base with EpiSetter at blocks 11–15. To measure intent-conditioned integration of prior and contextual knowledge, we evaluate paired prior/context selection on 417 held-out examples from the ConFiQA-QA MR and MC test splits. Layer controls isolate the intervention components responsible for paired performance. To assess answer quality and capability retention beyond this source-integration task, we evaluate six MRQA development sets and the independent ARC-Challenge, IFEval, GSM8K, and SQuAD-dev checks.

Comparisons and uncertainty. EpiSetter is benchmarked against other methods that enable models to choose from parametric and contextual knowledge and the base model. The comparisons include CK-PLUG, which adaptively controls reliance on parametric and contextual knowledge (Bi et al., 2025), SHIFT, which uses input-dependent gates for activation steering under knowledge conflict (Li et al., 2026), and Knowledgeable-R1, which reinforces parametric knowledge to resist misleading context (Lin et al., 2026).

5.2 EpiSetter Enables Epistemic Integration of Prior and Contextual Knowledge

Joint correctness improves under the frozen answer boundary. Appendix Table 2 and Figure 3 compare Base, CK-PLUG, SHIFT, Knowledgeable-R1, and L11–15 on the ConFiQA-QA test set. EpiSetter increases PairAcc from 18.23% to 36.93%, a gain of 18.71 percentage points. Prior EM increases from 28.54% to 49.40%, while context EM increases from 65.47% to 68.59%. This demonstrates EpiSetter's ability to integrate prior and contextual knowledge sources.

Adaptive direction calibration enables intent-conditioned source integration As shown in Figure 3, the three controller conditions separate source integration from an undirected shift. Base reaches 18.23% PairAcc, with prior/context EM of 28.54/65.47%, whereas the learned L11–15 controller reaches 36.93% and 49.40/68.59%, respectively. Reversing the learned direction reduces PairAcc to 2.88% and prior/context EM to 8.39/50.60%. Thus, the gain is associated with directing the intervention toward the source required by each prompt: an incorrectly oriented shift actively disrupts both source choices, while the learned controller improves prior selection without sacrificing context selection on this test. A mean-constant controller reaches 24.22% PairAcc, below L11–15, indicating that source integration benefits from input-dependent control rather than a fixed displacement alone.

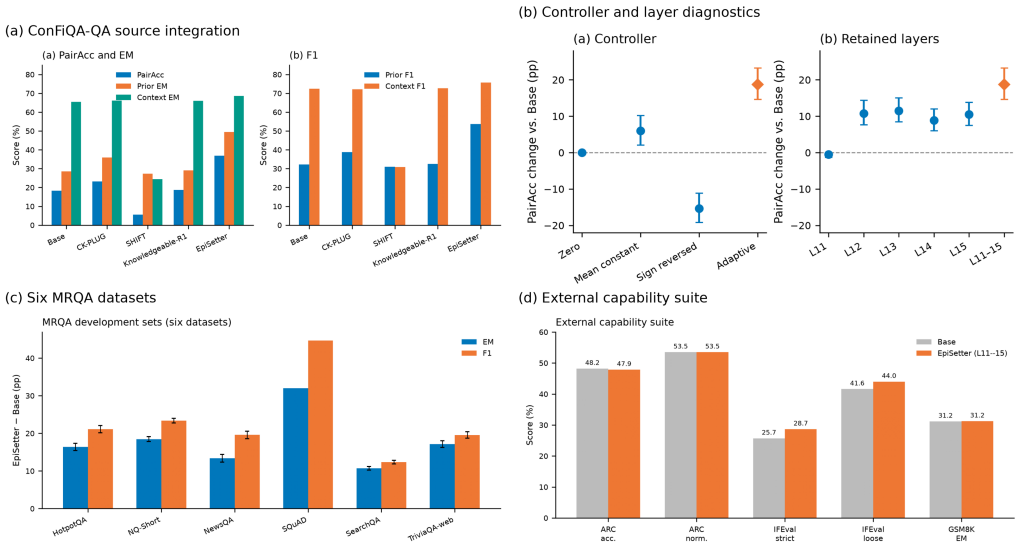


Figure 3: Overview of all figures used in the Results section. (a) ConFiQA-QA source integration across Base, external adaptations, and L11–15; (b) controller and retained-layer diagnostics; (c) six MRQA development datasets; (d) Matched ARC-Challenge, IFEval, and GSM8K capability checks for Base and L11–15.

5.3 EpiSetter Preserves Model Capability on General Tasks

Capability is retained across unrelated QA benchmarks. Appendix Table 3 reports all six MRQA development sets. Figure 3 visualizes their EM/F1 and answer-containment changes. On the five datasets with validated shared question IDs, extracted- answer EM improves by 10.70–18.49 pp and F1 by 12.35–23.38 pp.

Figure 3 also gives the independent capability suite. The first-100-question SQuAD-dev slice is unchanged (90.0% EM for Base and L11–15). On the matched capability suite, L11–15 keeps ARC-Challenge normalized accuracy at 53.50% (accuracy 47.86% versus 48.21% for Base), improves IFEval strict/loose instruction following from 25.69/41.59% to 28.65/43.99%, and keeps GSM8K EM nearly unchanged at 31.24% versus 31.16% (Appendix Table 4). Together with the MRQA and SQuAD checks, these results show preservation of the evaluated general-task capabilities.

CK-PLUG uses the same unified answer-tag prompt and scoring protocol on the six named MRQA development sets. Its scores are included in Appendix Table 3; the earlier zero-shot CK-PLUG run is excluded because it used a different extraction contract. These MRQA measurements are auxiliary answer-quality evidence; they neither replace the paired-intent endpoint nor demonstrate acquisition of new factual knowledge. We also exclude previously recorded CK-PLUG MRQA scores obtained with a different prompt/extraction contract from this matched comparison.

6 Conclusion

We introduced EpiSetter, a lightweight framework for intent-conditioned source reliance with a frozen language model backbone. By separating bidirectional direction learning from sequential scalar calibration, EpiSetter extends prior/context selection to continuous, input-dependent adjustments along layerwise control directions. Experiments show improved joint correctness under paired intents, with control comparisons supporting the value of adaptive intervention magnitudes. These results demonstrate the promise of learned source control with empirical capability-preservation constraints. The residual-based direction construction requires further numerical stability analysis. Establishing broader knowledge integration also requires experiments on tasks that combine complementary information from both sources.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, finding related works, generating execution codes, generating synthetic datasets and draft sections of the paper. We have not used generative AI tools for mathematical proof. Additionally, we used generative AI tools for concept graph generation. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness manually. We checked the test cases manually after a rough screening by LLM. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

ETHICS STATEMENT

We foresee no particular ethical concerns.

REPRODUCIBILITY STATEMENT

Reproducible code is provided at https://github.com/zw-huang/EpiSetter.git

REFERENCES

Bi et al. Parameters vs. context: Fine-grained control of knowledge reliance in language models, 2025. URL https://arxiv.org/abs/2503.15888.

Fang et al. Alphaedit: Null-space constrained knowledge editing for language models, 2025. URL https://arxiv.org/abs/2410.02355.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Hsu et al. Contextual linear activation steering of language models, 2026. URL https://arxiv.org/abs/2604.24693.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.findings-acl.70/.

Khandelwal et al. Cocoa: Confidence- and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of EMNLP*, 2025. URL https://aclanthology.org/2025.emnlp-main.348/.

Ruochang Li, Pengcheng Huang, Zhenghao Liu, Yukun Yan, Huiyuan Xie, Yu Gu, Ge Yu, and Maosong Sun. SHIFT: Gate-modulated activation steering for knowledge conflict mitigation in retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2606.27786.

Lin et al. Resisting contextual interference in RAG via parametric-knowledge reinforcement. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4dbcd4ba82866fdefc7ea0d15e9a72f2-Abstract-Conference.html.

Shayne Longpre, Yi Lu, Joachim Daiber, et al. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://aclanthology.org/2021.emnlp-main.565/.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. Controllable context sensitivity and the knob behind it. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.07404.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.559/.

Massimo Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Sch"olkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.458/.

Andrew Shih, Emily Winnicki, and Yuxin Cao. How do language models choose between context and memory?, 2026. URL https://arxiv.org/abs/2609.00753.

Wei Sun, Yushi Bai, and Mark Dredze. Task matters: Knowledge requirements shape LLM responses to context–memory conflict. In *Findings of the Association for Computational Linguistics*, 2026. URL https://aclanthology.org/2026.findings-acl.202/.

Wang et al. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of NAACL*, 2025a. URL https://aclanthology.org/2025.naacl-long.581/.

Wang et al. Continuously steering LLMs sensitivity to contextual knowledge with proxy models. In *Proceedings of EMNLP*, 2025b. URL https://aclanthology.org/2025.emnlp-main.233/.

Zhao et al. Understanding parametric and contextual knowledge reconciliation within large language models. In *Advances in Neural Information Processing Systems*, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/94f7c80260eea8d52a5c0ec7b1637c04-Abstract-Conference.html.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025b. URL https://aclanthology.org/2025.naacl-long.264/.

A Appendix

A.1 Direction construction and numerical limitations

The released implementation casts the gradient matrix to float32 and divides each row by its norm clamped below at 10^{-8}. It computes a reduced SVD and retains right singular vectors whose singular values exceed 10^{-5} times the largest singular value (itself clamped below at 10^{-8}), subject to an optional rank cap. These vectors form the rows of U_ℓ. The residual is evaluated as $w_\ell - U_\ell^\top (U_\ell w_\ell)$, without explicitly forming Π_ℓ, and then normalized as in Equation 13. A nonfinite norm or a norm at or below 10^{-8} raises an error.

Initialization applies this residual-and-normalization operation to the mean paired hidden-state difference and uses the resulting vector as the initial trainable w_ℓ. Adam updates w_ℓ through the same operation at every training step; donor states are cached from the unmodified backbone. The final vector is processed once more and its sign is oriented toward the mean context-minus-prior difference.

All five retained bases in the released checkpoint have rank 4096, equal to the hidden dimension. Consequently, the exact residual operator for a complete orthonormal basis is zero; the computed nonzero directions depend on floating-point residuals. Their recorded $\|U_\ell v_\ell\|_\infty$ values range from approximately 0.068 to 0.086, which does not establish protection. In particular, small individual

basis coordinates would not imply small total overlap with a full-dimensional span. The absolute norm threshold only prevents normalization of very small or invalid residuals; it does not test conditioning or numerical reproducibility.

The reported experiments do not establish stability across precision settings or SVD perturbations, nor isolate a sensitivity-reduction benefit of this direction construction. Capability-retention results apply to the complete calibrated model and cannot be attributed to residual projection. A nondegenerate alternative would require an explicitly truncated basis with rank below d, or an explicit soft sensitivity penalty, followed by retraining and matched controls. Neither alternative is represented by the reported checkpoint or results.

A.2 Layer screening and intervention scales

Define the donor-oriented margin $m_p^{s,r}(x) = \overline{\mathcal{N}}_p(a^r \mid x) - \overline{\mathcal{N}}_p(a^s \mid x)$. A candidate layer is source-feasible when its mean validation interchange gain $m^{s,r}_{p^{r\leftarrow s}_{\ell,v}}(x^r) - m^{s,r}_{p_0}(x^r)$ is positive separately for both swap directions. The layers passing this test under this empirical construction form $\mathcal{S}$. All five candidate layers, $\mathcal{S} = \{11, 12, 13, 14, 15\}$, are enabled in the released checkpoint. For each, we estimate

$$\sigma_\ell = \mathrm{median}_{i \in \mathcal{D}_{\mathrm{dir}}} \left| v_\ell^\top [h_\ell(x_i^C) - h_\ell(x_i^P)] \right|. \tag{21}$$

Scales below 10^{-6} are rejected. The sign is oriented toward the mean context minus prior difference for consistent reporting. We also measure protected loss and first-token KL changes under shifts $\pm\sigma_\ell v_\ell$. In the current implementation these are diagnostics: direction feasibility is determined by bidirectional source gain, not a hard protection cutoff.

The historical unprotected-direction scan uses the bidirectional coordinate-interchange objective over all 32 blocks. Its five largest mean validation signed margin gains occur at L14, L11, L13, L15, and L12, forming the contiguous block L11–15. Their gains range from 0.2455 to 0.2636, compared with 0.1793 and 0.1863 at the adjacent layers L10 and L16. This scan motivates the candidate block; all five layers subsequently pass bidirectional source-gain validation after empirical direction learning.

A.3 Setter architecture, training, and decoding

Each setter uses a two-layer predictor with hidden width 32, giving 655,685 trainable parameters across the five setters. Zero-initializing the final linear layer yields the identity intervention at initialization. The intervention radius is $\rho_\ell = m\sigma_\ell$ with $m = 1$ and remains fixed throughout calibration.

Direction learning runs for 1,200 updates per candidate. Calibration runs for 2,000 updates with Adam, a learning rate of 10^{-3}, and validation every 200 updates. Each calibration update takes one source and one protected example from separately shuffled training lists, optimizes only the predictor parameters, and clips their gradient norm to 1. The loss weights are reported in Appendix A.4.

At inference, setters act at the fixed final-prompt position. The reference decoder recomputes the prefix with that position fixed; an equivalent cached decoder must preserve the intervened prompt's downstream keys and values.

A.4 Calibration weights and validation constraints

Let $\mathcal{D}^{\mathrm{val}}_{\mathrm{src}}$ and $\mathcal{K}^{\mathrm{val}}$ be held-out source and protected sets. Define the mean source sequence NLL $J(\theta)$, the worst protected token-normalized NLL increase $B(\theta)$, and the worst protected reference-prefix KL $K(\theta)$:

$$J(\theta) = \mathbb{E}_{(x,a)\sim\mathcal{D}^{\mathrm{val}}_{\mathrm{src}}} \mathcal{N}_{p_\theta}(a \mid x), \tag{22}$$

$$B(\theta) = \max_{(u,b)\in\mathcal{K}^{\mathrm{val}}} \left[\overline{\mathcal{N}}_{p_\theta}(b \mid u) - \overline{\mathcal{N}}_{p_0}(b \mid u) \right], \tag{23}$$

$$K(\theta) = \max_{(u,b)\in\mathcal{K}^{\mathrm{val}}} \frac{1}{|b|} \sum_{t=1}^{|b|} D_{\mathrm{KL}}(p_0(\cdot \mid u, b_{<t}) \,\|\, p_\theta(\cdot \mid u, b_{<t})). \tag{24}$$

Among evaluated checkpoints satisfying $B(\theta) \le 0.05$ and $K(\theta) \le 0.05$, we retain the one with the lowest $J(\theta)$. Selection starts from the zero-output identity checkpoint and replaces it only on a

Table 1: Validation trajectory for the released seed-42 calibration run.

Step	J	B	K	Feasible ($B, K \leq 0.05$)
0	2.60115	0.00000	0.00000	Yes (identity)
200	1.03163	0.14166	0.00662	No
400	1.06269	0.03732	0.00248	Yes
600	1.04042	0.01794	0.00368	Yes
800	1.04023	0.01889	0.00235	Yes
1000	1.05654	0.02249	0.00104	Yes
1200	1.03560	0.01096	0.00198	Yes
1400	1.03169	0.04301	0.00075	Yes
1600	1.04249	0.02437	0.00196	Yes
1800	**1.02994**	**0.01457**	**0.00069**	**Yes; selected**
2000	1.05992	0.01217	0.00160	Yes

Table 2: Paired generation on 417 held-out examples from the ConFiQA-QA MR and MC test splits. EM and F1 use the frozen leading-answer parser. ΔPairAcc is relative to Base in percentage points with paired anchor-cluster 95% intervals.

Method	PairAcc	Prior EM	Prior F1	Context EM	Context F1	ΔPairAcc	95% CI
Base	18.23	28.54	32.19	65.47	72.40	–	–
SHIFT (local)	5.52	27.34	31.04	24.46	30.90	-12.71	[-16.31,-8.95]
Knowledgeable-R1 (local)	18.71	29.02	32.51	65.95	72.72	+0.48	[-0.45,+1.47]
EpiSetter (seed 42)	36.93	49.40	53.63	68.59	75.69	+18.71	[+14.59,+23.23]
CK-PLUG	23.26	35.97	38.73	66.19	72.20	+5.04	[+1.57,+8.28]

strict improvement, so ties retain the earlier checkpoint and no feasible improvement yields identity fallback. Margin gains are diagnostics, not the calibration selection objective; selection does not use Pareto ranking.

For the released seed42 run, this rule selects step 1800: validation source sequence NLL decreases from 2.60115 to 1.02994, with $B = 0.0145733$ and $K = 0.000688647$. These are selection-set measurements, not guarantees on unseen data or substitutes for external capability evaluation. Training and validation data remain marked as model-screened and exploratory.

The released OLMo3-7B run uses $\lambda_{\text{keep}} = 1$, $\lambda_{\text{amp}} = 0.01$, and validation cutoffs $B, K \leq 0.05$.

The repository also contains OLMo sweep configurations proposing $\lambda_{\text{amp}} \in \{0.001, 0.01, 0.1, 1\}$ in the separate `soft_constraint_calibration` implementation. The associated output directories contain no calibration results, so this planned sweep cannot be used as empirical evidence for the released bounded calibrator. We therefore report the observed release validation trajectory and a replay of its checkpoint selector under alternative cutoffs below. The cutoff replay is a post hoc sensitivity analysis of the recorded checkpoints, not a new run or an independent hyperparameter search.

Table 1 gives the validation trajectory for the released seed-42 run. The B limit rejects step 200 despite its low source NLL. Step 1800 is the lowest-J feasible checkpoint under the reported cutoffs. All trained checkpoints have $K < 0.0067$, so the $K = 0.05$ condition does not bind in this run.

A.5 Benchmark Results

Table 5 replays the same checkpoint rule, including the identity checkpoint, while varying one cutoff at a time and holding the other at 0.05. It shows that the selected checkpoint remains step 1800 over a range of cutoffs, while a tighter B limit of 0.01 or a tighter K limit of 0.0005 falls back to identity. In particular, these results show the practical effect of the chosen cutoff on this validation trajectory, but do not identify 0.05 as an optimal tolerance.

Table 3: MRQA dev benchmark results. Each row identifies a dataset and method; metric columns are reported separately in percent. Counts are common question IDs for the five usable Base comparisons.

Dataset	Method	n	EM	F1	ACC
HotpotQA	Base	5,892	10.42	22.19	57.42
HotpotQA	EpiSetter (42)	5,892	26.80	43.33	57.93
HotpotQA	CK-PLUG	5,892	25.54	42.78	57.96
NQ-Short	Base	12,464	15.56	31.59	64.07
NQ-Short	EpiSetter (42)	12,464	34.04	54.97	63.70
NQ-Short	CK-PLUG	12,464	32.08	53.44	62.15
NewsQA	Base	4,126	3.73	18.97	54.10
NewsQA	EpiSetter (42)	4,126	17.09	38.60	54.58
NewsQA	CK-PLUG	4,126	19.70	41.07	53.08
SQuAD†	Base	10,507	15.06	31.36	78.01
SQuAD†	EpiSetter (42)	10,507	34.17	54.58	78.05
SQuAD†	CK-PLUG	10,507	34.92	54.94	77.15
SearchQA	Base	16,966	8.94	16.97	60.96
SearchQA	EpiSetter (42)	16,966	19.65	29.32	63.13
SearchQA	CK-PLUG	16,966	19.37	29.01	62.18
TriviaQA-web	Base	7,754	9.35	18.95	78.62
TriviaQA-web	EpiSetter (42)	7,754	26.49	38.49	79.13
TriviaQA-web	CK-PLUG	7,754	28.22	40.22	78.98

Table 4: General-task capability checks for the frozen Base checkpoint and the matched seed-42 L11–15 EpiSetter checkpoint.

Dataset	Method	n	Metric / score (%)
SQuAD-dev slice	Base	100	EM / 90.00
SQuAD-dev slice	EpiSetter (42)	100	EM / 90.00
ARC-Challenge	Base	1,170	accuracy / 48.21
ARC-Challenge	Base	1,170	normalized accuracy / 53.50
ARC-Challenge	EpiSetter (42)	1,170	accuracy / 47.86
ARC-Challenge	EpiSetter (42)	1,170	normalized accuracy / 53.50
IFEval	Base	541	strict prompt / 25.69
IFEval	Base	541	loose prompt / 41.59
IFEval	EpiSetter (42)	541	strict prompt / 28.65
IFEval	EpiSetter (42)	541	loose prompt / 43.99
GSM8K	Base	1,319	EM / 31.16
GSM8K	EpiSetter (42)	1,319	EM / 31.24

Table 5: Post hoc sensitivity of checkpoint selection to validation cutoffs.

Varied cutoff	Value	Selected step	Selected J
B (with $K = 0.05$)	0.01	Identity (0)	2.60115
B (with $K = 0.05$)	0.02	1800	1.02994
B (with $K = 0.05$)	0.05	1800	1.02994
K (with $B = 0.05$)	0.0005	Identity (0)	2.60115
K (with $B = 0.05$)	0.001	1800	1.02994
K (with $B = 0.05$)	0.05	1800	1.02994