SELFCUE: MAKING A 3D CT REPORT GENERATOR SAY WHAT IT ALREADY KNOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Progress in 3D CT report generation is usually sought in increasingly sophisticated architectures and larger pools of training data. We find instead that a 3D CT report generator already holds what its report leaves out, and loses it when the hidden state becomes tokens. Over the 18 CT-RATE abnormalities, this hidden-to-report surfacing gap is reflected by a drop in macro AUROC from 0.848 in the hidden states to 0.739 in the generated report. We propose **SelfCue** based on contrastive decoding. It promotes what the hidden state already supports and suppresses what it does not. It raises clinical efficacy F1 to 0.481 and the LLM-judged GREEN score to 0.510. Distilling that behaviour into the weights gives **SelfCue-KD**, a student that keeps most of the gain, needs nothing extra at inference, and drops into any pipeline already serving the baseline. Code is available at `https://anonymous.4open.science/r/SelfCue-anonymous-348F`.

1 INTRODUCTION

Computed tomography is a workhorse of diagnostic imaging, with about 93 million examinations performed in the United States in 2023 (Smith-Bindman et al., 2025). Reporting one takes a radiologist about 15 minutes (Kurmukov et al., 2024). Automating that reading has produced a steady line of 3D CT models that write reports (Hamamci et al., 2026; Liang et al., 2026a). They share one recipe. A 3D encoder turns the volume into visual tokens, a projector maps those tokens into a language model, and the model is fine-tuned to produce text. Their reports often get the abnormalities wrong, and fall short in overall quality. Most work has responded with stronger encoders (Wald et al., 2025; Cao et al., 2025) or an explicit abnormality classifier attached alongside (Di Piazza et al., 2025; Jin et al., 2024). We instead ask what the generator already encodes.

We probe every stage of the pipeline on one ruler (Liang & Xu, 2026), traced in Figure 1A. For AUROC over the 18 CT-RATE abnormalities, the raw CT volume is 1.0 by assumption, since the abnormality is physically present to be seen. It remains 0.848 before the language model, and stays there at every layer we probe. It drops to 0.739 in the generated report. A large part of the information therefore never reaches the text. The gap does not open gradually across the network but only when the hidden state becomes tokens, so what the report leaves out is missing from its text and not from the model that writes it. The same shape appears on two CT visual encoders and on two language models, so it is not a quirk of one pipeline. Our goal is to narrow this surfacing gap.

We propose **SelfCue** to correct the output distribution. We want the abnormalities the model already encodes to appear in the report, and the ones it does not to stay out. Following contrastive decoding (Sanchez et al., 2024; Shi et al., 2024), SelfCue runs the model twice for every token, once under a prompt carrying an abnormality condition and once without it. It decodes from a combination of the two that pushes away from the second. The abnormality condition is predicted for each volume, not read off the labels.

SelfCue has two costs at inference. It runs the model twice for every token, which doubles the work and does not fit standard inference engines. The abnormality condition also has to be computed for every volume before decoding. We distil the corrected decoder into a student model we call **SelfCue-KD**. Following sequence-level distillation (Kim & Rush, 2016), SelfCue writes a report for every training volume, and the student is fine-tuned on those reports under a prompt with no abnormality condition. At inference it decodes with a single forward pass, with no condition and no contrast, so the correction lives in the weights rather than in the decoding rule.

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.

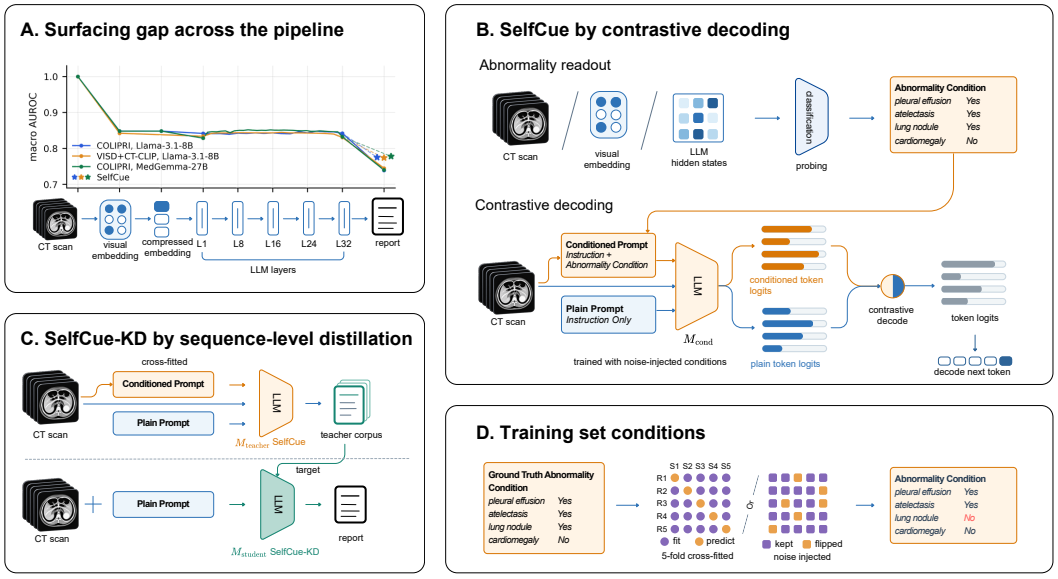

Figure 1: **A.** Macro AUROC over the 18 CT-RATE abnormalities at each stage of three report generators, the raw volume 1.0 by assumption. **B.** The abnormality condition is read by a probe, and each next token is chosen by contrasting the conditioned and plain branches. **C.** The teacher's reports train the student, which reads no condition. **D.** The two constructions of the training set conditions. Red marks a probe error.

Contributions.

- We measure abnormality information along the whole 3D CT reporting pipeline on one ruler and localise the information loss to the point where the hidden state becomes text, on two CT visual encoders and two language models.
- We propose **SelfCue**, which narrows the surfacing gap by contrasting the generator's output distribution with and without an abnormality condition. It reaches 0.481 macro F1@0.5 and 0.510 GREEN on COLIPRI with Llama-3.1-8B.
- We distil SelfCue into **SelfCue-KD**, which decodes with a single forward pass. It keeps most of the performance while no longer depending on the abnormality condition.

2 Related Work

2.1 3D CT Report Generation

Work on CT-RATE (Hamamci et al., 2026) has advanced largely through the visual pathway. Tokenizers built for volumes (Li et al., 2025) and stronger visual backbones (Shi et al., 2026) change what the language model is given, as does a cross-attention decoder built for 3D volumes (Hamamci et al., 2024). SliceWorld trains a world-state model that predicts the next slice and reads the report off the learned dynamics (Tian & Song, 2026). Others leave the generator alone and retrieve similar studies to supply what the volume does not (Liang et al., 2026a). Closest to us are the decoders that put an abnormality classifier's output into the prompt (Jin et al., 2024; Zhang et al., 2025), and CT-AGRG, which generates one sentence for each abnormality its own head predicts (Di Piazza et al., 2025). SelfCue starts from the same prediction and gets more out of it, changing only the decoding rule.

2.2 Contrastive Decoding and Representation Intervention

Burns et al. (2023) recover from a language model's activations the truth of statements its own text gets wrong. Orgad et al. (2025) and Goel et al. (2026) report the same gap for the model's own

errors. We find the same gap in CT report generation, and contrastive decoding is the family of rules built to close it. The original form contrasts a large model with a small one and keeps only the tokens the large model still finds plausible (Li et al., 2023b). Later work changes what the second branch is conditioned on: shallow layers against deep ones (Chuang et al., 2024), an image-conditioned branch against a distorted one (Leng et al., 2024), a prompt against its absence (Sanchez et al., 2024; Shi et al., 2024). In chest radiography the rule appears anchored on an external classifier (Zhang et al., 2026) or on category-wise visual prompts (Srivastava et al., 2026). SelfCue contrasts a prompt with its absence and closes part of the gap. Reshaping the residual stream is the other way to act on what a probe finds (Li et al., 2023a), and a recent report on medical failure regimes finds the information decodable there but not corrected by fixed linear steering (Liu, 2026). SelfCue leaves the states alone and changes only the output distribution.

2.3 Knowledge distillation

Removing a two-branch decoder by distilling it is a mature line in diffusion (Meng et al., 2023). The same has been done for autoregressive image tokens (Chen et al., 2025), and the medical analogue distils in-context vectors into a report generator (Wu et al., 2026). Those students keep the condition at inference. Ours does not receive one, so the teacher holds information the student can never be given, the setting of learning under privileged information (Vapnik & Izmailov, 2015). Distilling a prompt into the weights so that it need not be supplied at inference is context distillation (Snell et al., 2022). Ours distils a decoding rule rather than a prompt, and the condition the teacher reads cannot be given to the student at all. SelfCue-KD follows sequence-level distillation (Kim & Rush, 2016) rather than matching token distributions (Hinton et al., 2015).

3 Surfacing gap

3.1 Probing every stage

We first ask where abnormality information is retained and where it fails to reach the report. We score every stage of the pipeline on the same abnormality-level ruler. Each one is scored on the same 18 CT-RATE abnormalities, and every number is a macro average over them. The ruler is AUROC (Hanley & McNeil, 1982; Fawcett, 2006), which depends only on how the volumes are ranked and so is unchanged by any monotone transformation of the scores. We use the same CheapCT probe head (Liang & Xu, 2026) to score the visual tokens the encoder hands to the projector and the language model's hidden states. As a report is text, RadBERT (Yan et al., 2022) first assigns it a probability per abnormality, and the AUROC is taken over those. We score the baseline generator's hidden states at every second layer under the same text prompt for every volume. At each layer we average the hidden states over all visual and text prompt positions, a choice Appendix B.1 compares against two others. The probe is fit on the training split and every number reported here is on CT-RATE validation.

3.2 The gap along the pipeline

Figure 1A traces the macro AUROC along the pipeline. Every point is measured here except COLIPRI's embedding before ORCA's compression, which is quoted from ORCA (Liang et al., 2026b). The raw volume scores 1.0 by assumption, and the encoder carries part of that information into its embedding, where the score drops to 0.848 on COLIPRI. The language model holds it at every layer we probe, 0.839 to 0.848. The report then scores 0.739. That surfacing gap of 0.109 opens when the hidden state becomes tokens. Swapping the visual encoder does not change this: on VISD+CT-CLIP (Cao et al., 2025; Hamamci et al., 2026; Liang et al., 2026a) the gap is 0.100. Swapping the language model does not change it either: on MedGemma-27B (Sellergren et al., 2025) the gap is 0.112, with its 62 layers rescaled onto the 32-layer axis of Llama-3.1-8B.

4 Method

SelfCue surfaces abnormalities that a probe can recover from a CT report generator's internal representations but that are missing from its reports. It renders the probe's prediction as a textual

abnormality condition and uses that condition to drive contrastive decoding. Distilling the generated reports then gives SelfCue-KD, which needs only one forward pass per token.

4.1 SELF-CUED DECODING

We apply a contrastive decoding rule (Sanchez et al., 2024; Shi et al., 2024) to narrow the surfacing gap. Let x denote a chest CT volume and M_{base} a baseline 3D CT report generator. We threshold an abnormality readout and render the predicted abnormalities as a short sentence. That sentence is the abnormality condition, and it carries noisy information rather than ground truth. The readout can come from the hidden states or from the visual embedding. The baseline has never seen a conditional prompt, so feeding it one would be out of distribution. We therefore train M_{cond}, warm-started from M_{base} and fine-tuned on CT–report pairs with the condition prepended. Its training set comes in three kinds of prompt. The first carries no condition. The second carries the ground-truth abnormalities. The third carries the same labels degraded to the probe's measured accuracy. The first keeps the plain branch in distribution. The other two show the condition at its best and at its real accuracy (Figure 1D).

At inference the abnormality condition for this volume is prepended to its plain prompt, giving the conditional prompt (Figure 1B). Decoding then proceeds one token at a time. At each step both branches of M_{cond} see the same partial report, the conditional branch under the conditional prompt and the plain branch under the plain prompt. Writing $\boldsymbol{\ell}^{\text{cond}}$ and $\boldsymbol{\ell}^{\text{plain}}$ for their next-token logits, SelfCue decodes from

$$\widetilde{\boldsymbol{\ell}} = \boldsymbol{\ell}^{\text{cond}} + \alpha \left(\boldsymbol{\ell}^{\text{cond}} - \boldsymbol{\ell}^{\text{plain}} \right), \tag{1}$$

where α controls the contrast strength and is set to 0 on volumes whose condition mentions no abnormality. Here that means the readout found none, not that the volume is normal. The combination promotes tokens the conditional branch favours more strongly and demotes those the plain branch favours more strongly. We greedily select the next token from $\widetilde{\boldsymbol{\ell}}$ and append it to both branches, so the two always continue from the same text.

4.2 SELFCUE-KD

SelfCue runs two forward passes of M_{cond} at every step. We remove this inference-time overhead through sequence-level knowledge distillation (Kim & Rush, 2016), with SelfCue as the teacher M_{teacher} (Figure 1C). Because the student receives no condition at inference, sequence-level distillation here has to meet two conditions it ordinarily does not. First, the conditions used to write the teacher corpus must be honest out of sample (Dao et al., 2021). The probe is fit on the training split, so applying it back to that split returns conditions close to the ground truth and the teacher would carry information the student can never obtain. We split the training volumes into five disjoint folds $S_1, \ldots, S_5$ and fit a separate probe on each set of four, so every training volume is paired with a report whose condition came from a probe that never saw it. Writing $M_{\text{teacher}}^{(-k)}$ for SelfCue driven by the probe fit on the four folds other than k, the teacher corpus is

$$\mathcal{D}^T = \bigcup_{k=1}^{5} \left\{ \left(x_i,\ M_{\text{teacher}}^{(-k)}(x_i) \right) :\ x_i \in S_k \right\}. \tag{2}$$

Second, the teacher is told which abnormalities to mention and the student is not, so the student has to infer them from the image. M_{student} is initialized from M_{cond} but receives only the plain prompt. Rather than matching the teacher's token-level distributions, which would run both branches of the teacher at every training step, the student uses each complete teacher report as a sequence-level target and is optimized with standard autoregressive cross-entropy:

$$\mathcal{L}_{\text{SeqKD}} = - \sum_{(x_i, y_i) \in \mathcal{D}^T} \log p_{\text{student}} \left(y_i \mid x_i \right). \tag{3}$$

At inference M_{student} receives only the CT volume and the plain prompt, and generates in a single pass.

System	F1@0.5	$F1_{max}$	AUROC	GREEN	BLEU-1	BLEU-4	ROUGE-L	METEOR
Eligible works								
CT-CHAT (Hamamci et al., 2026)	0.184	–	–	–	0.395	–	0.334	–
RadAgent (Roschewitz et al., 2026)	0.220	–	–	–	–	–	–	–
BTB3D (Hamamci et al., 2025)	0.258	–	–	–	0.439	0.213	–	0.223
OmniCT (Lin et al., 2026)	0.363	–	–	–	–	–	–	–
U-VLM (Shi et al., 2026)	0.414	–	–	–	0.474	0.256	–	–
AdaRAG-CT (Liang et al., 2026a)	0.480	–	–	–	0.496	0.242	0.354	0.246
CLarGen (Maye-Lasserre et al., 2026)	0.486	–	–	–	–	0.208	0.298	0.335
CT-AGRG (Di Piazza et al., 2025)	–	0.501	–	–	–	0.172	0.280	0.196
μ^2Tokenizer (Li et al., 2025)	–	–	–	0.429	–	–	–	0.359
SliceWorld (Tian & Song, 2026)	–	–	–	0.507	0.394	0.226	0.382	0.230
COLIPRI + Llama-3.1-8B								
M_{base}	0.395	0.482	0.739	0.466	0.456	0.226	0.345	0.214
M_{cond}	0.430	0.514	0.762	0.510	0.425	0.220	0.356	0.209
SelfCue	0.481	0.542	0.775	0.510	0.424	0.221	0.356	0.210
SelfCue-KD	0.462	0.534	0.768	0.510	0.416	0.217	0.353	0.208
VISD+CT-CLIP + Llama-3.1-8B								
M_{base}	0.408	0.474	0.745	0.463	0.461	0.223	0.337	0.214
M_{cond}	0.449	0.515	0.761	0.512	0.428	0.224	0.358	0.211
SelfCue	0.487	0.539	0.774	0.504	0.428	0.222	0.355	0.211
SelfCue-KD	0.460	0.526	0.754	0.512	0.412	0.218	0.358	0.209
COLIPRI + MedGemma-27B								
M_{base}	0.348	0.471	0.740	0.459	0.451	0.213	0.326	0.206
M_{cond}	0.438	0.518	0.763	0.472	0.386	0.162	0.274	0.201
SelfCue	0.481	0.548	0.778	0.456	0.392	0.165	0.276	0.204
SelfCue-KD	0.472	0.535	0.771	0.467	0.386	0.160	0.270	0.200

Table 1: **Comparison on CT-RATE.** SelfCue decodes at $\alpha = 1.0$ and $\tau = 0.75$ on all three encoder–backbone pairs. M_{cond} is decoded with the condition in its prompt and no contrast.

5 Experiments

5.1 Setup

Dataset. All experiments use CT-RATE (Hamamci et al., 2026) and its 18 abnormality labels. COLIPRI is scored on $n = 1564$ volumes and VISD+CT-CLIP on $n = 2987$. COLIPRI keeps one reconstruction of each scan, and reconstructions of a scan share a report. Its 1564 volumes therefore cover every report in the validation split, so the two remain comparable.

Encoders. We use two visual embeddings. The first comes from ORCA (Liang et al., 2026b), which compresses the COLIPRI encoder (Wald et al., 2025) to 216 tokens and provides them. The second is the pairing of ViSD organ embeddings (Cao et al., 2025) with a mean-pooled CT-CLIP embedding (Hamamci et al., 2026) introduced by Liang et al. (2026a), the encoder's own 5 tokens.

Backbones. The generator is a LLaVA-style pipeline (Liu et al., 2023): the visual tokens pass through a projector into a language model fine-tuned with LoRA (Hu et al., 2022). We use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) which has 32 transformer layers, and MedGemma-27B (Sellergren et al., 2025), which has 62 layers.

Metrics. Abnormalities are read out of the generated text with RadBERT (Yan et al., 2022), as in Section 3.1. We report macro AUROC, which measures how well the volumes are ranked. Alongside the widely used clinical efficacy F1, macro over the 18 abnormalities at the fixed threshold of 0.5, we report macro $F1_{\max}$, the best F1 over all thresholds and therefore an upper bound. Because it maximises over thresholds, $F1_{\max}$ is unchanged by a systematic shift in the scores, as macro AUROC is. GREEN (Ostmeier et al., 2024) gives a language-model judgement of the report. We also report BLEU-1 and BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Banerjee & Lavie, 2005).

Comparison eligibility. We apply several eligibility conditions so that the systems can be compared fairly. A system must evaluate on the official CT-RATE validation split and score against the whole report, findings and impression together. It must report at least one of clinical F1 or GREEN. It must read abnormalities with the official released RadBERT weights and score with the official GREEN weights. We report the numbers each paper published itself. 10 works pass this screen and are included in our comparison.

Implementation. The prompt mixture that trains M_{cond} is 30% plain, 35% ground-truth and 35% ground-truth with the probe's own error rates applied. SelfCue has two hyperparameters, the contrast strength α and the threshold τ at which the readout becomes a condition, both chosen on the validation split at $\alpha = 1.0$ and $\tau = 0.75$. We train the projector and a rank-128 LoRA adapter on the language model, with the visual encoder and the compressor frozen. Optimisation uses AdamW (Loshchilov & Hutter, 2019) at a learning rate of 2×10^{-5} with a cosine schedule and an effective batch size of 16. The MedGemma baseline uses 5×10^{-6}, which a sweep selected for it. Every model is trained and run on a single NVIDIA B200 GPU.

5.2 Main results

BLEU, ROUGE-L and METEOR reflect fluency rather than whether the abnormalities are right (Liang & Xu, 2026). We report them for comparability and read clinical quality from F1 and GREEN.

Clinical efficacy. SelfCue improves clinical efficacy on all three encoder–backbone pairs in Table 1. M_{cond} already improves on M_{base} in F1@0.5, $\text{F1}_{\max}$ and AUROC, and SelfCue improves on M_{cond} in each of them. On COLIPRI with Llama-3.1-8B, F1@0.5 rises from 0.395 to 0.430 and then to 0.481, and $\text{F1}_{\max}$ from 0.482 to 0.514 and then to 0.542. M_{cond} and SelfCue use the same model and the same condition and differ only in how they decode, so the gain from M_{cond} to SelfCue comes from our decoding rule.

The two F1 scores measure different things. $\text{F1}_{\max}$ takes the best threshold for each abnormality and so reflects how well the reports rank the volumes, while F1@0.5 scores the reports as they are written. Precision and recall in Appendix C.3 show how F1@0.5 improves. On COLIPRI recall rises from 0.365 to 0.452 while precision moves from 0.590 to 0.577. On VISD+CT-CLIP and MedGemma-27B recall rises by 0.065 and 0.105 with precision within 0.006. SelfCue mentions abnormalities that M_{cond} left out at nearly the same precision. Appendix C.1 follows one volume through the four systems. On COLIPRI a report mentions 2.4 of the 3.4 abnormalities present, so there is room above what it reaches here. SelfCue recovers 59%, 61% and 64% of the $\text{F1}_{\max}$ gap between the report of M_{base} and the probe that reads its hidden state. It adds the most where the report leaves out the most. Sorted by prevalence, the 18 abnormalities fall into three groups of six. Against the report of M_{base}, $\text{F1}_{\max}$ rises from 0.359 to 0.463 on the rarest group and from 0.598 to 0.635 on the most common. The rarest group is also where the probe exceeds the report of M_{base} by the widest margin.

GREEN. SelfCue holds GREEN at the level of M_{cond} on all three pairs. A higher τ raises GREEN at the cost of clinical efficacy. GREEN weighs every statement equally, and its judge counts about ten findings per report against the 3.43 of the 18 abnormalities present in a volume. SelfCue moves a fraction of those, mentioning 2.40 abnormalities per volume against M_{cond}'s 2.01, so a large gain in F1 reaches only a small part of what GREEN scores.

SelfCue-KD. SelfCue-KD is trained on the reports SelfCue writes and decodes with a single forward pass under the plain prompt. On COLIPRI with Llama-3.1-8B it reaches 0.462 F1@0.5 and 0.534 $\text{F1}_{\max}$, within 0.020 and 0.008 of its teacher, and its GREEN is the same. On VISD+CT-CLIP with Llama-3.1-8B the gap to its teacher is wider, 0.028 and 0.014, yet it stays above M_{cond} on both and ends 0.007 ahead of its teacher on GREEN. On COLIPRI with MedGemma-27B it is within 0.009 and 0.013 of its teacher, stays above M_{cond} on both, and is 0.011 ahead of its teacher on GREEN. It therefore runs unchanged in a standard inference engine. On one B200, SelfCue decoding one report at a time takes 6.3 hours for the 1564 validation volumes, while SelfCue-KD under vLLM (Kwon et al., 2023) writes all of them in 71 seconds.

Eligible works. With Llama-3.1-8B, SelfCue is on par with the strongest eligible works on F1@0.5 and GREEN and above them on $\text{F1}_{\max}$. CLarGen (Maye-Lasserre et al., 2026) reaches an F1@0.5 of 0.472 with Llama-3.1-8B as its writer, against 0.481 for SelfCue. With MedGemma-27B it reports 0.486 against 0.481 for SelfCue on the same backbone. SelfCue thus depends less on the backbone than CLarGen does. SliceWorld (Tian & Song, 2026) reports a GREEN of 0.507, against 0.510 for both SelfCue and SelfCue-KD on COLIPRI. It gets there with a world model pretrained on DeepLesion, a second CT dataset, whereas both of ours are trained on CT-RATE alone. CT-AGRG (Di Piazza et al., 2025) is the only eligible work that reports $\text{F1}_{\max}$, 0.501, against 0.542 for SelfCue

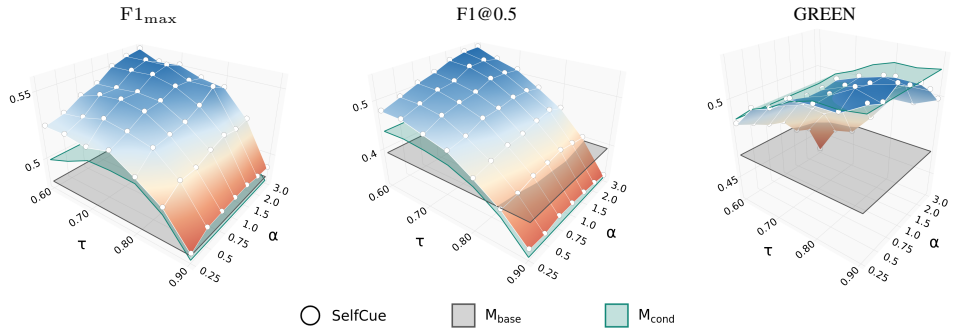

Figure 2: SelfCue with the gate on empty conditions over α and τ on COLIPRI with Llama-3.1-8B. M_{cond} is decoded with the condition at the same τ and no contrast.

and 0.534 for SelfCue-KD. The other eligible works score below SelfCue on every clinical metric they report.

5.3 Ablations

5.3.1 Contrast strength and condition threshold

SelfCue has two hyperparameters. The contrast strength α sets how far the decoding moves away from the plain branch, and the threshold τ sets how many abnormalities the condition mentions. Figure 2 sweeps both on COLIPRI with Llama-3.1-8B, and Appendix C.4 sweeps the other two pairs. At every one of the 28 cells, SelfCue exceeds on $\text{F1}_{\max}$ both M_{base} and M_{cond} decoded with the same condition, and it exceeds M_{cond} on F1@0.5 as well, so the gain holds wherever the two are set. The threshold τ moves $\text{F1}_{\max}$ most. $\text{F1}_{\max}$ stays flat up to $\tau = 0.75$ and falls beyond it, and M_{cond} falls with it. A higher threshold puts fewer abnormalities in the condition, and at $\tau = 0.90$ the condition mentions none on 55% of volumes. F1@0.5 falls over the whole range instead. A lower threshold makes the reports mention each abnormality more often. F1@0.5 counts these mentions at a fixed cut of 0.5. $\text{F1}_{\max}$ picks the best cut for each abnormality and so barely moves.

At $\alpha = 0$, SelfCue reduces to M_{cond}. Every positive α lifts $\text{F1}_{\max}$ above it, so the contrast itself helps. F1@0.5 follows the same course. GREEN moves the other way. At $\tau = 0.75$ it stays at the level of M_{cond} up to $\alpha = 1.0$ and drops at larger α, and it rises with τ. At $\tau = 0.60$ the contrast costs GREEN rather than gaining it, 0.475 against M_{cond}'s 0.491. Because a higher threshold keeps only the abnormalities the probe is most confident about, the reports contain fewer false findings, which GREEN counts as errors (Ostmeier et al., 2024).

	$\tau = 0.70$			$\tau = 0.75$			$\tau = 0.90$		
α	F1@0.5	$\text{F1}_{\max}$	GREEN	F1@0.5	$\text{F1}_{\max}$	GREEN	F1@0.5	$\text{F1}_{\max}$	GREEN
0.25	0.000	0.000	−0.001	0.000	0.000	0.000	+0.001	+0.002	−0.002
1.0	0.000	0.000	+0.002	0.000	0.000	0.000	+0.004	+0.003	+0.003
2.0	+0.003	+0.003	+0.022	+0.004	+0.004	+0.029	+0.008	+0.009	+0.038
3.0	+0.007	+0.006	+0.057	+0.007	+0.005	+0.070	+0.015	+0.012	+0.097

Table 2: **Gate on empty conditions.** It sets α to 0 on volumes whose condition is empty, so they are decoded without contrast. Cells are gated minus ungated SelfCue on COLIPRI with Llama-3.1-8B.

5.3.2 Gate on empty conditions

When the condition mentions no abnormality, SelfCue sets α to 0 on that volume. Without this gate the contrast would amplify the sentence *no significant abnormality* against a plain branch that carries no condition at all. Table 2 compares gated and ungated decoding on COLIPRI with Llama-3.1-8B. The gate leaves F1@0.5 and $\text{F1}_{\max}$ nearly unchanged everywhere. Its effect on GREEN grows with α, and with τ, since a higher threshold leaves more conditions empty. How much the gate matters depends on the pair. On COLIPRI with MedGemma-27B it already raises GREEN by 0.027 at the operating point.

	Condition source			Training set condition		
Where it is used	F1@0.5	$\mathrm{F1}_{\max}$	GREEN	F1@0.5	$\mathrm{F1}_{\max}$	GREEN
Training M_{cond}	$+0.008$	$+0.004$	-0.001	-0.026	-0.021	-0.017
SelfCue decoding	$+0.008$	0.000	-0.003	–	–	–
Teacher corpus	$+0.005$	-0.003	-0.004	-0.151	-0.063	-0.013

Table 3: **Condition source.** It is either the LLM hidden states or the visual embedding, and the training set conditions are ground truth with injected noise or 5-fold cross-fitted predictions. Each cell is the change when one choice is replaced by the other, on COLIPRI with Llama-3.1-8B.

5.3.3 CONDITION SOURCE

The abnormality condition is used in three places, to train M_{cond}, to decode with SelfCue and to write the teacher corpus. We implement two condition sources, a probe on the LLM hidden states and a probe on the visual embedding. On the training set a probe fit on that set cannot be applied to it directly, so the conditions there are either ground truth with injected probe-level noise or 5-fold cross-fitted predictions. We read the probe off the hidden states, train M_{cond} under injected noise and write the teacher corpus under cross-fitted conditions. Table 3 swaps each choice for the other option, one at a time. Reading the probe off the visual embedding instead changes every measure by at most 0.008. How the training set conditions are produced matters more, and in opposite directions for the two uses. Training M_{cond} on cross-fitted conditions lowers its $\mathrm{F1}_{\max}$ by 0.021. Writing the teacher corpus under injected noise lowers the student's $\mathrm{F1}_{\max}$ by 0.063, and the student mentions 1.53 abnormalities per volume instead of 2.25. The student sees only the image. Injected noise is independent of the volume, so nothing in the image tells the student where the condition errs, whereas cross-fitted errors fall on the volumes the probe finds hard.

5.3.4 SEQUENCE-LEVEL DISTILLATION

SelfCue-KD is trained on the teacher's text. The alternative is token-level distillation (Hinton et al., 2015), which matches the student's next-token distribution to the teacher's at every position of the reference reports. Table 4 compares the two students, which start from the same checkpoint and train for the same number of steps. Appendix C.6 varies that checkpoint and the teacher reports the student sees. The two losses produce nearly the same student. The sequence-level student is ahead by 0.010 F1@0.5 and 0.002 $\mathrm{F1}_{\max}$ and behind by 0.002 GREEN. What the student ends up doing is set by the corpus it is trained on far more than by the loss. We use the sequence-level loss because it is ordinary fine-tuning on text and never holds the teacher in memory beside the student.

The mention rates show what the teacher writes and what the student keeps. The contrast pushes the abnormalities the condition lists into the report. Given the condition in its prompt, M_{cond} mentions 0.70 of the listed abnormalities that are present, and SelfCue raises this to 0.90. It pushes the listed abnormalities that are absent almost as far, from 0.61 to 0.86, so the probe's errors pass into the teacher's reports. The students inherit far less of them. They decode without the condition and read only the image, and SelfCue-KD is more selective than M_{cond}, which reads the condition. It mentions more of the listed abnormalities that are present, 0.73, fewer of those that are absent, 0.58, and three times as many of the present abnormalities the condition leaves out, 0.11 against 0.04. So besides speed, distillation leaves the deployed model less bound to the probe.

	M_{base}	M_{cond}, plain prompt	M_{cond}	SelfCue	SelfCue-KD	Token-level KD
F1@0.5	0.395	0.299	0.430	0.481	0.462	0.452
$\mathrm{F1}_{\max}$	0.482	0.438	0.514	0.542	0.534	0.532
AUROC	0.739	0.709	0.762	0.775	0.768	0.765
GREEN	0.466	0.495	0.510	0.510	0.510	0.512
In condition, present	0.505	0.364	0.697	0.901	0.729	0.695
In condition, absent	0.386	0.259	0.610	0.855	0.578	0.554
Outside condition, present	0.148	0.095	0.039	0.018	0.112	0.121
Outside condition, absent	0.055	0.030	0.009	0.006	0.019	0.020

Table 4: **Sequence-level distillation.** The last four rows are mention rates, split by whether the condition lists the abnormality and whether it is present, averaged over the 18 abnormalities. Results are on COLIPRI with Llama-3.1-8B.

	F1@0.5	$\mathrm{F1}_{\max}$	AUROC	GREEN
M_{base}	0.395 0.424 ± 0.008	0.482 0.487 ± 0.006	0.739 0.750 ± 0.006	0.466 0.431 ± 0.004
M_{cond}	0.430 0.430 ± 0.006	0.514 0.510 ± 0.007	0.762 0.757 ± 0.001	0.510 0.510 ± 0.006
SelfCue	0.481 0.489 ± 0.012	0.542 0.541 ± 0.005	0.775 0.774 ± 0.005	0.510 0.503 ± 0.007
SelfCue-KD	0.462 0.463 ± 0.001	0.534 0.529 ± 0.002	0.768 0.763 ± 0.002	0.510 0.510 ± 0.008

Table 5: **Held-out selection.** COLIPRI with Llama-3.1-8B on the validation split. Each cell gives the mean and standard deviation over three seeds, after the reported value in grey.

5.4 Selection on held-out data

CT-RATE has no test split, so Table 1 is scored on the same validation split its settings were chosen on. A held-out experiment in Table 5 confirms that the reported setting and its gains remain consistent without the validation split. We carve 3,000 training volumes by patient and train on the remaining 21,128 at three seeds, with the recipe unchanged. All three seeds land on that setting. The models here see 12% less data and still perform at the level of Table 1, with M_{base}'s GREEN the one deviation. SelfCue gains 0.059 F1@0.5 and 0.031 $\mathrm{F1}_{\max}$ over M_{cond}, against 0.052 and 0.028 there. Across seeds, the standard deviation is at most 0.012 on every measure. SelfCue therefore holds its gain both without the split its setting was chosen on and across independent training runs.

6 Discussion

The limit is the readout. SelfCue closes about 60% of the distance between the report and the probe that reads the generator's hidden state, so some of what is left is still open to the decoding rule. The larger part is not. Given the ground-truth labels as its condition, the same rule reaches 0.869 macro F1@0.5 against SelfCue's 0.480 on COLIPRI with Llama-3.1-8B (Appendix C.5), so what a report can say is bounded by what the readout gets right.

F1 against GREEN. The two metrics count different things. Macro F1 weighs the 18 abnormalities equally, and SelfCue gains most on the rarest of them, the abnormalities a report is most likely to leave out. GREEN counts everything a report says, of which the 18 abnormalities are a small part, and a condition that is right about every one of them leaves it where it was. SelfCue thus raises the clinical content of a report while keeping its agreement with the reference level.

Knowing and saying. Earlier work reads the gap between what a model holds and what it says out of its activations (Burns et al., 2023; Orgad et al., 2025; Goel et al., 2026). We close part of it instead of measuring it, in a setting where the text is the clinical product rather than an answer to be checked. The rule needs only a readout of that state and a generator that accepts a condition, so it carries to other report generators. Our evidence comes from CT-RATE and its 18 labelled abnormalities, and other imaging modalities and reader studies are the next step.

Future work. What SelfCue writes into the condition is limited to the 18 abnormalities that CT-RATE labels, a small part of what GREEN counts. The decoding rule is not tied to these labels. A probe trained on richer annotations, such as where an abnormality lies, how large it is or findings outside the 18, could fill the same condition sentence. A condition of that kind is what it would take for the gain to reach GREEN as well as F1. A second direction is iterative training. SelfCue-KD already mentions three times as many of the present abnormalities that the condition leaves out as M_{cond} does. The student can take the place of M_{base}, and the whole pipeline of probing, conditioning, decoding and distilling can run again on it. Whether each round adds to the last is open.

7 Conclusion

A 3D CT report generator holds more about the abnormalities in a volume than its report says. The information survives every layer of the model and is lost only when the hidden state becomes text. SelfCue reads it back from the generator's own hidden state and uses it to steer decoding, raising clinical efficacy, and SelfCue-KD moves that correction into the weights so the deployed model decodes in a single pass. Surfacing what a model already knows is a route to better reports that complements stronger encoders and larger datasets.

AI USE STATEMENT

In this work, we used AI tools to assist with the implementation of software code used in our experiments. We also used Claude for brainstorming and language editing to improve the readability of the manuscript. Generative AI tools were not used to design the research methodology or experiments, interpret the experimental results, or conduct literature review and citation selection. All AI-assisted code and text were reviewed and verified by the authors. We take full responsibility for the final content of this work.

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, and Ling Zhang. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23041–23050, October 2025.

Huayu Chen, Hang Su, Peize Sun, and Jun Zhu. Toward guidance-free AR visual generation via condition contrastive alignment. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 65862–65881, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a4a697a6922ff56a93fd27e9a0842986-Paper-Conference.pdf.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edc36117f795ca52a0cbf6a7b3882859-Paper-Conference.pdf.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations*, 2021.

Théo Di Piazza, Carole Lazarus, Olivier Nempont, and Loïc Boussel. CT-AGRG: Automated abnormality-guided report generation from 3D chest CT volumes. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 01–05, 2025. doi: 10.1109/ISBI60581.2025.10981073.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010. URL https://www.sciencedirect.com/science/article/pii/S016786550500303X. ROC Analysis in Pattern Recognition.

Jyotin Goel, Ipshita Bandyopadhyay, and Justin Shenk. The knowing-saying gap: When probes see errors that confidence misses, 2026. URL https://arxiv.org/abs/2608.07528.

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, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent,

Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3D medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012 of *Lecture Notes in Computer Science*, pp. 476–486. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72390-2_45. URL https://doi.org/10.1007/978-3-031-72390-2_45.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3D: Advancing vision-language modeling in 3D medical imaging. 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. 135074–135102. Curran Associates, Inc., 2025. doi: 10.52202/085713-4507. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c4861cc81b768dc36a676fceed8f7672-Paper-Conference.pdf.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, February 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01599-y.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747. URL https://doi.org/10.1148/radiology.143.1.7063747.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial*

Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i3.28038. URL https://doi.org/10.1609/aaai.v38i3.28038.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Anvar Kurmukov, Valeria Chernina, Regina Gareeva, Maria Dugova, Ekaterina Petrash, Olga Aleshina, Maxim Pisov, Boris Shirokikh, Valentin Samokhin, Vladislav Proskurov, Stanislav Shimovolos, Maria Basova, Mikhail Goncahrov, Eugenia Soboleva, Maria Donskova, Farukh Yaushev, Alexey Shevtsov, Alexey Zakharov, Talgat Saparov, Victor Gombolevskiy, and Mikhail Belyaev. The impact of deep learning aid on the workload and interpretation accuracy of radiologists on chest computed tomography: a cross-over reader study, 2024. URL https://arxiv.org/abs/2406.08137.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023a. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Siyou Li, Pengyao Qin, Huanan Wu, Dong Nie, Arun J. Thirunavukarasu, Juntao Yu, and Le Zhang. μ^2tokenizer: Differentiable multi-scale multi-modal tokenizer for radiology report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pp. 3–12. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04971-1_1. URL https://doi.org/10.1007/978-3-032-04971-1_1.

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/.

Renjie Liang and Zijian Xu. When do cheap probes predict expensive training? probing 3D-CT encoders for text generation, 2026. URL https://arxiv.org/abs/2607.22771.

Renjie Liang, Yiling Ma, Yang Xing, Zhengkang Fan, Jinqian Pan, Chengkun Sun, Li Li, Kuang Gong, and Jie Xu. Beyond the embedding bottleneck: Adaptive retrieval-augmented 3D CT report generation, 2026a. URL https://arxiv.org/abs/2603.15822.

Renjie Liang, Zijian Xu, Jinqian Pan, Chengkun Sun, Zhengkang Fan, Shawn Li, You Qin, Mei Liu, and Jie Xu. ORCA: ORgan-Centroid aggregation for training-free 3D CT visual token compression, 2026b. URL https://arxiv.org/abs/2608.00345.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, Peng LU, Yueting Zhuang, Ling Zhang, Beng Chin Ooi, and Yingda Xia. OmniCT: Towards a unified slice-volume LVLM for comprehensive CT analysis. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 41691–41716, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/45419fc1c6edac1e2abb42644c9a183c-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Ming Liu. Decodable but not corrected by fixed residual-stream linear steering: Evidence from medical LLM failure regimes, 2026. URL https://arxiv.org/abs/2605.05715.

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

Tom Maye-Lasserre, Yitong Li, Bailiang Jian, Morteza Ghahremani, Benedikt Wiestler, and Christian Wachinger. Generating reports or repeating templates? measuring and mitigating template collapse in 3D CT report generation, 2026. URL https://arxiv.org/abs/2605.30984.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, 2023. doi: 10.1109/CVPR52729.2023.01374.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a712d461e57201efe35d429a6f1731c1-Paper-Conference.pdf.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Mélanie Roschewitz, Kenneth Styppa, Yitian Tao, Jiwoong Sohn, Jean-Benoit Delbrouck, Benjamin Gundersen, Nicolas Deperrois, Christian Bluethgen, Julia E. Vogt, Bjoern Menze, Farhad Nooralahzadeh, Michael Krauthammer, and Michael Moor. RadAgent: A tool-using AI agent for stepwise interpretation of chest computed tomography, 2026. URL https://arxiv.org/abs/2604.15231.

Guillaume Sanchez, Alexander Spangher, Honglu Fan, Elad Levi, and Stella Biderman. Stay on topic with classifier-free guidance. 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. 43197–43234. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sanchez24a.html.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. MedGemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Pengcheng Shi, Minghui Zhang, Kehan Song, Jiaqi Liu, Yun Gu, and Xinglin Zhang. U-VLM: Hierarchical vision language modeling for report generation, 2026. URL https://arxiv.org/abs/2603.00479.

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/.

Rebecca Smith-Bindman, Philip W. Chu, Hana Azman Firdaus, Carly Stewart, Matthew Malekhedayat, Susan Alber, Wesley E. Bolch, Malini Mahendra, Amy Berrington de González, and Diana L. Miglioretti. Projected lifetime cancer risks from current computed tomography imaging. *JAMA Internal Medicine*, 185(6):710–719, 06 2025. ISSN 2168-6106. doi: 10.1001/jamainternmed.2025.0505. URL https://doi.org/10.1001/jamainternmed.2025.0505.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context, 2022. URL https://arxiv.org/abs/2209.15189.

Shantam Srivastava, Mahesh Bhosale, David Doermann, and Mingchen Gao. CWCD: Category-wise contrastive decoding for structured medical report generation. In Yuankai Huo, Mingchen Gao, Chang-Fu Kuo, Yueming Jin, and Ruining Deng (eds.), *Proceedings of The 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 868–893. PMLR, 2026. URL https://proceedings.mlr.press/v315/srivastava26a.html.

Yuanhe Tian and Yan Song. SliceWorld: A predictive and controllable world-state model for CT report generation, 2026. URL https://arxiv.org/abs/2605.24371.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL http://jmlr.org/papers/v16/vapnik15b.html.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle,

and Fernando Pérez-García. Comprehensive language-image pre-training for 3D medical image understanding, 2025. URL https://arxiv.org/abs/2510.15042.

Ning Wu, Rui Liu, Xinkun Lin, Weixing Chen, Jinxi Xiang, Tao Wei, Lina Yao, and Mingjie Li. Not all tokens matter equally: Dynamic in-context vector distillation with decisive-token supervision for long-form medical report generation, 2026. URL https://arxiv.org/abs/2605.27194.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. doi: 10.1148/ryai.210258. URL https://doi.org/10.1148/ryai.210258.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. CCD: Mitigating hallucinations in radiology MLLMs via clinical contrastive decoding. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35174–35200, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1755. URL https://aclanthology.org/2026.findings-acl.1755/.

Xiaodan Zhang, Yanzhao Shi, Junzhong Ji, Chengxin Zheng, and Liangqiong Qu. MEPNet: Medical entity-balanced prompting network for brain CT report generation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34788. URL https://doi.org/10.1609/aaai.v39i24.34788.