computers

MDPI

Article

Reason2Decide-C: Adaptive Cycle-Consistent Training for Clinical Rationales

H M Quamran Hasan [1,*], Housam Khalifa Bashier Babiker [1], Mi-Young Kim [2] and Randy Goebel [1,*]

[1] Department of Computing Science, Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB T6G 2R3, Canada; khalifab@ualberta.ca
[2] Department of Science, Augustana Faculty, University of Alberta, Camrose, AB T4V 2R3, Canada; miyoung2@ualberta.ca
* Correspondence: hmquamra@ualberta.ca (H.M.Q.H.); rgoebel@ualberta.ca (R.G.)

Abstract

Large Language Models (LLMs) used for clinical decision support must not only make accurate predictions but also generate rationales that are consistent with, and sufficient for, those predictions. Building on Reason2Decide, a two-stage rationale-driven multi-task framework, we propose Reason2Decide-C (R2D-C, where C denotes cycle consistency), which augments Reason2Decide's stage 2 training with confidence-adaptive scheduled sampling and cycle-consistent rationale-to-label training. In stage 1, we pretrain our model on rationale generation. In stage 2, we jointlytrain on label prediction and rationale generation, gradually replacing gold labels with model-predicted labels based on confidence. Simultaneously, we feed the rationale logits back into the model to recover the label, thus enforcing explanation sufficiency. We evaluate R2D-C on one proprietary triage dataset, as well as public biomedical QA and reasoning datasets. Across model sizes, R2D-C substantially improves rationale–prediction consistency (where stage 1 and stage 2 predictions agree) and sufficiency (where the rationale alone recovers the ground-truth label) over other baselines while matching or modestly improving predictive performance (F1); in several settings R2D-C surpasses 40× larger foundation models. Ablations confirm that the full combination is optimal, maximizing alignment and LLM-as-a-Judge rationale quality. These results demonstrate that confidence-adaptive scheduled sampling and cycle-consistent rationale-to-label training substantially enhance explanation alignment without sacrificing accuracy.

Keywords: explainability; large language models (LLMs); clinical decision support; rationale generation

Check for updates

Academic Editor: Atsushi Teramoto

Received: 26 March 2026
Revised: 16 April 2026
Accepted: 22 April 2026
Published: 27 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Clinical decision support models are increasingly expected to not only produce accurate predictions but to also provide trustworthy explanations. Rationalization, the generation of free-text rationales, has emerged as a promising way to make models more transparent and trustworthy [1]. Prior work has shown that explanations that sound fluent but which are misaligned with the model's actual decision process can fabricate findings, hide uncertainty, and reinforce existing clinical biases, thus raising serious safety concerns in deployment settings [2,3]. In response, a growing line of work on self-rationalization trains or prompts models to produce explanations alongside their predictions, with the goal that these rationales reflect the internal reasoning used to reach a decision [4,5]. However,

recent studies [5,6] further demonstrate that LLM explanations frequently mention semantically relevant concepts that are not mechanistically important for the model's decision, thus revealing a gap between explanation plausibility and internal reasoning.

At the same time, rationales have emerged as a powerful form of intermediate supervision: by treating explanations as an additional training signal, small task-specific models can inherit part of a large teacher's reasoning behavior while remaining deployable in constrained settings [7,8]. Distilling Step-by-Step (DSS) [7] showed that extracting chain-of-thought [9] rationales from an LLM and training a smaller student model to jointly predict labels and generate rationales can outperform few-shot-prompted LLMs using substantially less data. Reason2Decide (R2D) [8] extended this idea to clinical Natural Language Processing (NLP) with a rationale-driven, two-stage multi-task framework that first pretrains a model for rationale generation and then jointly optimizes prediction and explanation with task-level scheduled sampling.

R2D uses a scheduled sampling [10] paradigm that shifts from conditioning on gold labels to model-generated labels, which is intended to mitigate exposure bias [11]. Exposure bias arises when models are trained to generate rationales conditioned only on ground-truth labels yet must explain their own potentially incorrect predictions during inference. However, R2D transitions from gold labels to model-predicted labels regardless of the model's prediction confidence. This creates an optimization conflict: when the model conditions on an incorrect label, it is still forced to generate a rationale that matches the ground-truth. Consequently, the model is trained to provide "gold" justifications for "wrong" predictions; this can significantly undermine the logical consistency of the generated explanations. Moreover, there is no explicit mechanism to enforce that the generated rationales are sufficient to reproduce the label (i.e., that the explanation alone implies the answer). Without such a constraint, the model may learn to mention clinically relevant evidence without encoding enough information in the rationale to recover the label.

To address these limitations in R2D, we introduce Reason2Decide-C (R2D-C, where C denotes cycle consistency) (Figure 1), a stage 2 training framework that extends R2D to more tightly couple predictions and explanations for clinical NLP tasks. Our approach has two main components. First, we propose a confidence-adaptive scheduler that uses a dynamically calculated confidence threshold to decide whether to condition explanation prompts on model-predicted or gold labels. Like task-level scheduled sampling in R2D, this reduces the train–test mismatch for explanations, while the confidence threshold specifically prevents over-exposing the model to low-confidence predictions. Second, we add a cycle-consistency objective that trains the model to recover labels from its own rationales via a differentiable rationale-to-label pathway. After the model generates an explanation, we softly (via Gumbel-Softmax [12]) feed that rationale back into the model to predict the label. This explicitly encourages rationales to be sufficient for reproducing the decisions they support.

R2D-C is not primarily intended as a method for maximizing raw predictive accuracy. Instead, its goal is to preserve the predictive behavior of rationale-driven clinical models while making their generated rationales more tightly aligned with the decisions they are meant to explain. This distinction is especially important in high-stakes clinical NLP, where a useful model must not only predict well but also provide rationales that reliably justify its predictions. Clinically, this distinction matters because a rationale that is fluent but weakly coupled to the model's recommendation may mislead clinician review, especially in triage settings where different dispositions imply different levels of urgency. By contrast, a rationale that is better aligned with the predicted decision is more useful for transparency, error checking, and decision support, even when predictive accuracy changes only modestly. Throughout this paper, we use prediction–rationale alignment to mean that generated

rationales are consistent with, and sufficiently informative to recover, the model's decision. We do not claim that these properties directly establish mechanistic faithfulness, i.e., that the generated rationale fully reflects the model's true internal causal reasoning process.

The main contributions of our paper are:

- We frame R2D-C as a method for improving prediction–rationale alignment in clinical NLP, rather than primarily for maximizing predictive accuracy, and show that this can be achieved without sacrificing predictive performance.
- We augment R2D's stage 2 with dynamic confidence-adaptive scheduled sampling, which decides when explanations should be conditioned on predicted versus gold labels, thereby reducing exposure bias while preventing conditioning on low-confidence predictions.
- We introduce a cycle-consistency loss that explicitly encourages rationales to be sufficient for reconstructing the labels they condition on.
- We integrate these into a unified framework and evaluate on triage, biomedical QA, diagnosis prediction, and multi-step medical reasoning tasks, demonstrating large gains in prediction–rationale alignment and LLM-judged rationale quality without sacrificing accuracy.

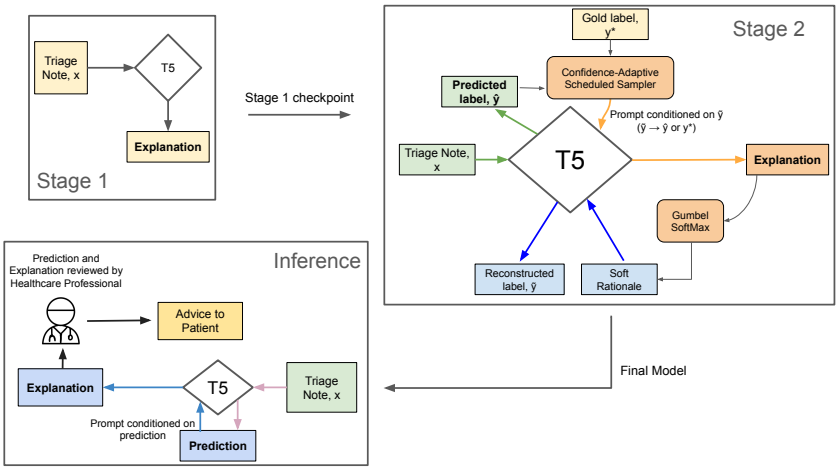


Figure 1. Overview of R2D-C. Stage 1 learns rationale generation. In stage 2, the model predicts a label, adaptively conditions rationale generation on the gold label (y^*) or the predicted label ($\hat{y}$), and applies cycle consistency so that the rationale preserves label information. During inference, the final model outputs both a prediction and an explanation for clinician review. The asterisk $*$ in y^* denotes the ground-truth/gold label, distinguishing it from the model's prediction $\hat{y}$.

2. Related Work

Our work builds on research in explainable NLP, multi-task learning, rationale supervision, and rationale-driven clinical NLP.

2.1. Explainable NLP and Faithful Self-Rationalization

Explainable NLP has produced a broad spectrum of methods, ranging from saliency and attribution techniques to counterfactual interventions and self-explanation models [3,13]. Recent surveys emphasize that faithfulness is distinct from human-perceived plausibility and that faithfulness is frequently overlooked [3]. Recent work has also examined explainability challenges specific to large language models, emphasizing the need for trustworthy explanations and connecting local explainability with mechanistic interpretability [14]. Within this landscape, self-rationalization methods train models to generate natural language explanations alongside predictions, which often leverage explanation datasets or prompting strategies to elicit chain-of-thought reasoning [4,7].

However, a growing body of work has challenged the faithfulness of these self-explanations. For example, Yang et al. [4] conducted a large-scale out-of-distribution study demonstrating that self-rationalizing models frequently produce plausible explanations even when predictions are incorrect, noting that standard metrics generalize poorly beyond the training distribution. Shifting from behavioral to mechanistic analysis, NeuroFaith [5] proposes a concept-level framework that compares self-generated explanations to the model's internal causal structure; the authors find that LLMs often cite "relevant" concepts that are not mechanistically important for the model's decision. This is supported by [6], which highlights a systematic divergence between a model's stated reasoning and its internal logic. Together, these results motivate approaches that explicitly tie explanation content to decision behavior rather than treating explanations as loosely related auxiliary outputs. Our work is motivated by this line of research on explanation faithfulness but operationally focuses on improving prediction–rationale alignment through confidence-adaptive self-conditioning and a cycle-consistency objective.

2.2. Multi-Task Learning and Knowledge Distillation

Multi-task learning leverages shared representation across related tasks to improve generalization [15,16]. In parallel, knowledge distillation trains a smaller student model to mimic a larger teacher model's behavior, typically via logits or intermediate features [17,18].

Rationales serve as a form of intermediate supervision that combines both ideas: instead of only matching logits, the student is trained to reproduce the teacher's explanation, thereby distilling not just predictions but aspects of the teacher's explanatory reasoning [7]. Distilling Step-by-Step (DSS) has approached this combination with a two-stage pipeline: first, a large teacher LLM is prompted to produce chain-of-thought rationales; second, a smaller student is trained in a multi-task fashion to jointly predict labels and generate those rationales [7].

In domain-specific applications, these ideas have been used to build compact, specialized models from general-purpose teachers. For clinical NLP, the ability to distill reasoning from foundation models mitigates the need for extensive clinician-authored explanations and enables smaller models to operate within strict latency and privacy constraints [8].

Our work shares the view that rationales can serve as an informative training signal beyond simple label supervision. However, unlike standard rationale distillation methods such as DSS, which primarily use explanations to improve downstream task performance, R2D-C emphasizes prediction–rationale alignment and develops this idea further within the rationale-driven clinical NLP framework introduced by R2D.

2.3. Rationale-Driven Clinical NLP

R2D [8] was proposed as a rationale-driven two-stage multi-task framework for clinical NLP. In stage 1, a T5 [19] model is trained to generate rationales. In stage 2, the same model is jointly optimized on label prediction and rationale generation with task-specific prompts and a multi-task loss. R2D uses a task-level scheduled sampling paradigm over labels [8]: during stage 2, the explanation prompt gradually transitions from conditioning on gold labels to conditioning on model-predicted labels, in order to address the potential mismatch between training (gold-conditioned explanations) and inference (self-conditioned explanations).

R2D demonstrates that this framework improves both predictive performance and explanation quality relative to standard fine-tuning and DSS-style baselines. However, its scheduled sampling policy is dependent only on the training step, and it does not explicitly enforce that generated rationales preserve enough information to recover the label.

R2D-C builds directly on this framework: it keeps the same two-stage rationale-driven design and prompts but replaces purely global label-level scheduled sampling with a confidence-adaptive variant and augments the loss function with a rationale-to-label cycle-consistency loss. In this way, R2D-C makes prediction–rationale alignment more explicit by encouraging explanations to be conditioned on the model's own predictions and to preserve sufficient information to recover the label.

2.4. Exposure Bias and Scheduled Sampling

Exposure bias [11] arises when models are trained to generate explanations conditioned on gold labels but during inference must condition on their own predictions. Scheduled sampling [10] mitigates this by gradually replacing gold tokens with model outputs during training. This helps align the training and inference conditions. Variants of scheduled sampling have been widely used in neural text generation, including machine translation [20].

R2D adapts scheduled sampling from the token level to the task level, gradually mixing gold and predicted labels as conditioning inputs for the explanation task [8]. This reduces the mismatch between training and inference but still treats all predictions identically regardless of model confidence.

Following the confidence-aware idea of [20], R2D-C extends [8] with a dynamic confidence threshold so that explanation prompts are self-conditioned on predicted labels primarily when the model is confident while resorting to gold labels for other cases. This preserves the exposure-bias benefits of R2D's task-level scheduled sampling but avoids over-training on low-confidence, often incorrect predictions.

3. Materials and Methods

In this section we introduce our methodology and experimental datasets and provide implementation details.

3.1. Methodology

Our methodology considers the combined task of label prediction and rationale generation: given an input x (such as a clinical case or biomedical question), the model must both predict a categorical label $y \in \mathcal{Y}$ and produce a natural language rationale r that justifies this decision.

We adopt an encoder–decoder architecture based on T5 variants, parameterized by θ, using the original model specification without architectural modifications. The attention mechanisms, hidden dimensions, positional encodings, and activation functions remain unchanged from the base implementation. The text-to-text framework of T5 is particularly well-suited for rationale-driven learning: it enables unified autoregressive generation for both short discrete labels (e.g., "Home Care," "Yes"/"No") and longer free-form explanations via task-specific input prefixes. This design allows multi-task training with a shared objective, unlike encoder-only architectures such as BERT [21], which require separate task-specific heads and lack native mechanisms for conditional text generation.

Our training configuration is designed to handle both discrete label prediction and free-text rationale generation. The model first predicts the label as $\texttt{predict:}x \mapsto \hat{y}$ then uses the label to condition the rationale as $\texttt{given label:}\tilde{y}\ \texttt{, explain:}x \mapsto \hat{r}$, where $\tilde{y}$ can be either y^* (the gold label) or $\hat{y}$ (model's predicted label), and $\hat{r}$ is the generated rationale.

3.1.1. Stage 1: Rationale Foundation Training

Domain-adaptive pretraining has been shown to achieve substantial improvements for specialized tasks in biomedical NLP [22,23]. Motivated by this evidence, we adopt a two-stage process in which the model first learns explanation fundamentals before tackling

the joint prediction-explanation problem. This initial phase, inherited from R2D [8], focuses exclusively on rationale generation: given a question x, the model is trained to produce the gold rationale r^*. Let θ denote the parameters of the T5 model. The stage 1 objective is

$$\mathcal{L}_{\text{stage1}} = -\log P_\theta(r^* \mid \texttt{explain:} \quad x) \tag{1}$$

where $P_\theta(r^* \mid \cdot)$ represents the autoregressive probability of the full rationale token sequence. We select the checkpoint that minimizes the validation loss to initialize stage 2.

This two-stage approach implements a curriculum learning design: By first establishing strong clinical explanation fundamentals, the model develops intermediate representations that encode clinical reasoning patterns. These representations then serve as a foundation for the more complex multi-task regime in stage 2, where the model must simultaneously maintain predictive accuracy and generate rationales that are aligned with, and sufficient to reproduce, its own decisions.

3.1.2. Stage 2: Joint Optimization of Prediction, Explanation, and Cycle Consistency

Stage 2 jointly trains the model on label prediction and rationale generation, extending R2D's framework with two key innovations:

- Confidence-adaptive scheduled sampling that selectively conditions explanations on predicted labels based on per-example confidence.
- A differentiable cycle-consistency objective that enforces rationales to be sufficient for reconstructing the conditioning label.

Prediction Task

The model treats label prediction as a text generation task, optimized with a cross-entropy loss objective:

$$\mathcal{L}_{\text{pred}} = -\log P_\theta(y^* \mid \texttt{predict:} \quad x) \tag{2}$$

where $P_\theta(y^* \mid \cdot)$ denotes the autoregressive probability of the full gold label sequence. During training, the predicted label $\hat{y}$ is greedily decoded and used to construct explanation prompts with our confidence-adaptive scheduling mechanism.

Explanation with Confidence-Adaptive Scheduled Sampling

To mitigate exposure bias in self-rationalization, R2D introduced task-level scheduled sampling. Unlike token-level scheduled sampling [10] that replaces tokens within generated sequences, task-level sampling switches the conditioning context from gold labels y^* to model-predicted labels $\hat{y}$ [8]. However, this approach treats all examples uniformly, potentially over-exposing the model to low-confidence (often incorrect) predictions.

We extend this with confidence-adaptive gating that conditions each example on either y^* or $\hat{y}$ based on both a global curriculum and per-example prediction confidence. This determines which label sequence is inserted into the explanation prompt. Specifically, we sample the conditioning label using two independent Bernoulli trials: the first with probability f_t (global curriculum) and the second with probability $\lambda(x)$ (confidence gate). The predicted label $\hat{y}$ is used only if both trials succeed, yielding an effective probability $P(\hat{y}) = f_t \cdot \lambda(x)$:

$$\tilde{y} = \begin{cases} y^* & \text{with probability } 1 - f_t \cdot \lambda(x) \\ \hat{y} & \text{with probability } f_t \cdot \lambda(x) \end{cases} \tag{3}$$

where $f_t \in [0, 0.9]$.

Sequence-level confidence: For each example x, we compute the geometric mean of token log-probabilities over the generated label sequence:

$$c(x) = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\log p(\hat{y}_t \mid \hat{y}_{<t}, x)\right) \tag{4}$$

where $\hat{y}_1, \ldots, \hat{y}_T$ are the greedily decoded label tokens.

Dynamic threshold calibration: We maintain a running estimate of typical confidence via a dynamically updated threshold τ. During the first 5% of training steps (matching the concurrent α_t and cycle warm-ups, both defined later), we collect $c(x)$ across batches. At the end of this warm-up period, we initialize τ to the empirical mean. Post-warm-up, τ is updated at discrete intervals of 1000 steps using the mean confidence of the current non-overlapping window:

$$\tau \leftarrow \text{mean}(\{c(x_i)\}_{i \in \text{current window}}) \tag{5}$$

After each update, the window is reset. This allows τ to adapt to the model's improving calibration while maintaining a stable gating threshold between updates.

Per-example gating: The confidence gate $\lambda(x)$ is computed as

$$\lambda(x) = \sigma\big(k(c(x) - \tau)\big), \quad k = 10 \tag{6}$$

where σ is the sigmoid function. The value of $k = 10$ is a heuristic choice to make the confidence gate sharp enough to distinguish high vs low confidence around τ. This produces a soft threshold: examples with $c(x) \gg \tau$ have $\lambda(x) \approx 1$ (likely use $\hat{y}$), while those with $c(x) \ll \tau$ have $\lambda(x) \approx 0$ (fall back to y^*).

The explanation loss is then

$$\mathcal{L}_{\text{expl}} = -\log P_\theta(r^* \mid \texttt{given label:}\ \ \tilde{y}\texttt{, explain:}\ \ x) \tag{7}$$

where $P_\theta(r^* \mid \cdot)$ is as defined in Equation (1).

Cycle Consistency: Rationale-to-Label Mapping

While confidence-adaptive scheduling ensures the model sees predicted labels during training, it does not explicitly enforce that rationales contain information *sufficient* to recover those labels. We introduce a differentiable cycle that reconstructs $\tilde{y}$ from the generated rationale. Unlike the task-level sampling above, the cycle-consistency module operates on the rationale token sequence itself by applying Gumbel-Softmax to the explanation decoder logits.

Soft rationale construction: From the explanation decoder logits (a tensor of shape [batch, sequence_length, vocabulary]), we apply Gumbel-Softmax with exponentially decaying temperature:

$$Q = \text{GumbelSoftmax}(\text{logits}, \tau_g(t)), \quad \tau_g(t) = \max(\tau_{\min}, \tau_0 e^{-\beta t}) \tag{8}$$

where $\tau_0 = 1.0$, $\tau_{\min} = 0.5$. We set the decay constant $\beta = \frac{\ln 2}{T_total}$, which ensures the temperature reaches the lower bound $\tau_{\min}$ precisely at the final training step of total training steps T_total. This produces soft token distributions over the vocabulary at each sequence position, which becomes sharper (more discrete) as training progresses. We compute soft embeddings by multiplying with the embedding matrix W_{emb} and prepend the task prefix:

$$\text{cycle_input} = [\text{emb}(\texttt{predict:}); Q \cdot W_{\text{emb}}] \tag{9}$$

Cycle loss: We train the model to recover $\tilde{y}$ (the same label used for conditioning) from this soft rationale representation:

$$\mathcal{L}_{\text{cyc}} = -\log P_\theta(\tilde{y} \mid \text{cycle_input}) \tag{10}$$

Gradients flow from the cycle loss back through the soft embeddings to the explanation logits, which encourages rationales to encode decision-relevant information. Crucially, the target is the actual conditioning label $\tilde{y}$ (gold or predicted), ensuring forward and reverse mappings are consistent.

Combined Objective and Training Schedule

Intuitively, confidence-adaptive scheduling decides which label conditions rationale generation, while cycle consistency checks whether the generated rationale preserves enough information to recover that same label. Algorithm 1 summarizes the full stage 2 training procedure.

Algorithm 1: Algorithm for one stage 2 training step in R2D-C

1. Given input x, predict a label sequence $\hat{y}$.
2. Compute the sequence-level confidence $c(x)$ of $\hat{y}$.
3. Use the current curriculum f_t and confidence gate $\lambda(x)$ to choose the conditioning label $\tilde{y} \in \{y^*, \hat{y}\}$.
4. Generate the rationale conditioned on $\tilde{y}$ and compute the explanation loss.
5. Apply Gumbel-Softmax to the rationale decoder logits to form a soft rationale sequence.
6. Feed the soft rationale back through the cycle branch and predict the label again.
7. Compute the cycle-consistency loss against the same conditioning label $\tilde{y}$.
8. Combine prediction, explanation, and cycle-consistency losses into the final training objective.

The total loss at training step t is

$$\mathcal{L}_t = \alpha_t \mathcal{L}_{\text{pred}} + (1 - \alpha_t)\mathcal{L}_{\text{expl}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} \tag{11}$$

where $\lambda_{\text{cyc}} = 0.1$ is kept constant after warm-up. The value 0.1 was initially chosen heuristically to balance cycle consistency against the primary prediction and explanation objectives; we later include a targeted sensitivity analysis over $\lambda_{\text{cyc}} \in \{0, 0.1, 0.2, 0.5\}$.

To smoothly transition from stage 1's explanation-only training to this full multi-task objective while stabilizing the new cycle-consistency constraint, we employ several scheduling mechanisms:

Concurrent α_t and Cycle Warm-ups: During steps $0 \leq t < w$ where $w = 0.05T_total$, we simultaneously:

- Increase the prediction weight α_t from 0 to 0.7 linearly. This is to prioritize explanation loss $\mathcal{L}_{\text{expl}}$ over prediction loss $\mathcal{L}_{\text{pred}}$, to provide a smoother transition from single-task to multi-task optimization [8]. After warm-up, α_t remains constant at 0.7. The choice of 0.7 was made following [8] to balance both tasks while slightly emphasizing prediction since stage 1 already established rationale generation capabilities.
- Disable the cycle loss ($\lambda_{\text{cyc}} = 0$). This allows the model to stabilize on prediction and explanation tasks before enforcing the stricter rationale-to-label consistency constraint.

This dual warm-up ensures that prediction capabilities develop while explanation generation remains prioritized and prevents the stricter cycle constraint from disrupting early training. For $t \geq w$, α_t remains fixed at 0.7 and λ_{cyc} activates at 0.1 and remains constant.

Label-Transition Schedule: With multi-task training stabilized, we then gradually introduce self-conditioning for explanations via the global curriculum f_t, which ramps up linearly for $w \leq t < w + m$ where $m = 0.60T_total$:

$$f_t = \begin{cases} 0, & 0 \leq t < w, \\ \min\left(0.9, \dfrac{t-w}{m}\right), & w \leq t < w+m, \\ 0.9, & t \geq w+m. \end{cases} \tag{12}$$

The 0.9 ceiling was a design choice following [8] to prevent the model from fully relying on self-generated labels during training, in order to help avoid error amplification from incorrect predictions. The hyperparameters $w = 0.05T_total$ for warm-up and $m = 0.60T_total$ for label transition were chosen following [8].

This coordinated schedule ensures: (1) smooth task balancing via α_t, (2) cycle constraint activation only after task stabilization, and (3) controlled self-conditioning exposure governed by $f_t \cdot \lambda(x)$.

3.1.3. Inference

During inference, we first predict the label via greedy decoding

$$\hat{y} = \arg\max_{y} P_\theta(y \mid \texttt{predict:} \quad x) \tag{13}$$

and then generate the rationale conditioned on that prediction:

$$\hat{r} = \arg\max_{r} P_\theta(r \mid \texttt{given label:} \quad \hat{y}\texttt{, explain:} \quad x) \tag{14}$$

3.2. Tasks and Datasets

We evaluate R2D-C across four clinical and biomedical reasoning benchmarks spanning clinical triage, biomedical question answering, diagnosis prediction, and multi-step medical reasoning. The four benchmarks differ not only in task structure but also in the provenance of their rationale supervision. Clinical Triage and DDXPlus use LLM-generated rationales, PubMedQA uses dataset-provided long answers, and MedReason uses rationales extracted from the dataset's reasoning field. We therefore do not treat these rationale sources as fully equivalent, and cross-dataset comparisons should be interpreted with this heterogeneity in mind. As models, we use T5-Small/Base and zero-shot LLMs as non-fine-tuned references.

3.2.1. Clinical Triage Dataset

The Clinical Triage Dataset originates from Alberta Health Link 811, a telephone-based health advice service managed by Alberta Health Services (AHS) in Canada. It comprises nurse-written triage notes from patient phone consultations, paired with recommended care dispositions and explanatory rationales for each decision.

Task: Predict the recommended care pathway (disposition) from a nurse triage note and provide a supporting rationale. The 12-class label space spans low-acuity options (Home Care) to high-urgency directives (e.g., "Go to Emergency Department (ED) immediately"). A full list of disposition classes is given in Appendix B.

Rationales: Since the original version of the rationales consists of atomic facts, following [8], we used a Qwen-3-8B [24] and provided the model with the triage note and disposition and asked it to explain why the disposition was chosen. All prompts used are provided in Appendix A. Because these rationales are LLM-generated rather than clinician-authored, they should be interpreted as a practical supervision signal rather than definitive

ground-truth explanations. This introduces a potential source of bias and circularity since the student model may partly learn to imitate properties of the teacher-generated rationale style rather than purely recover human clinical reasoning. All experiments were conducted using these LLM-generated rationales.

Data Split: The train, validation and test sets consist of roughly 171k, 21k and 9.7k samples, respectively.

3.2.2. PubMedQA

Task: Biomedical question answering with Yes/No/Maybe labels [25]. We concatenated the question and context to form the model input, analogous to the triage note in the clinical triage task.

Rationales: We use the dataset's `long_answer` field as the gold-standard rationale.

Data Split: PubMedQA consists of three subsets: pqa_labeled (1k human-annotated examples), pqa_artificial (automatically generated examples), and pqa_unlabeled (unannotated examples). Because the pqa_artificial subset does not contain any "Maybe" labels, we restricted our experiments to binary "Yes/No" classification. The distribution is highly imbalanced, with "Yes" as the dominant class. To counter this, we included all available "No" instances and constructed a subset in which "No" examples constitute 40% of the data, filling the remaining 60% with "Yes" examples. We then partitioned this subset into 70/20/10 train/validation/test splits. Finally, we augmented the test set with the 1k human-annotated examples from pqa_labeled, resulting in final splits of approximately 26k (train), 7.5k (validation), and 4.6k (test: 3.6k pqa_artificial + 1k pqa_labeled).

3.2.3. DDXPlus

Task: Differential diagnosis prediction from structured clinical cases [26]. The label space consists of 49 classes. We first processed the DDXPlus train set using the official evidence mapping (release_evidences.json) to convert evidence codes (e.g., "E123_@_V2") into human-readable format. The input to the model was the concatenation of the patient age and the converted evidence list: "patient age: [value from AGE column]. [EVIDENCE column]."

Rationales: Rationales were generated using a Qwen-3-8B [24] model following the same protocol as the Clinical Triage dataset. The model was provided the symptoms (EVIDENCE column) and Diagnosis (PATHOLOGY column) and was asked to generate an explanation. As with the Clinical Triage setting, these LLM-generated rationales provide scalable supervision but may also propagate teacher-model artifacts or stylistic regularities, which limit how strongly they can be interpreted as gold-standard explanations.

Data Split: We drew 100k stratified random samples (by PATHOLOGY) using a fixed seed from DDXPlus train set and created a 70k/20k/10k train/validation/test split.

3.2.4. MedReason

Task: Medical question answering with knowledge graph-grounded reasoning [27]. We concatenated the question and multiple-choice options as model input.

Rationales: For rationales, we extracted the reasoning text after the last "Conclusion:" marker from the dataset's reasoning field. We further cleaned the rationales by masking exact label matches with "[LABEL]", converting Unicode arrows (↑↓) to "go up"/"go down" and dropping rows missing conclusions or where extracted label text did not appear in the options string.

Data Split: We used the official train set and performed an 80/10/10 split for train/validation/test, resulting in final splits of roughly 12k (train), 1.5k (validation), and 1.5k (test).

3.3. Implementation Details

We implemented R2D-C (source code available at https://github.com/quamranhasan/Reason2Decide-C, accessed on 12 March 2026) using the T5 Small (77 M), and Base (250 M) architectures on 4× NVIDIA A100 GPUs (NVIDIA, Santa Clara, CA, USA). Training used the AdamW optimizer [28] with a learning rate of 5×10^{-5}, max input length = 1024, and effective batch size = 64. The model-specific configuration is provided in Table 1.

For stage 1, we monitored validation loss and applied early stopping with patience of 3 consecutive evaluations. Stage 2 used delayed early stopping (patience of 5) that activated only after the label-transition phase ($t \geq w + m$) completed. This ensured that the model fully adapted to our confidence-adaptive scheduling and cycle-consistency mechanisms before termination.

During validation, we evaluated only the prediction task using macro F1-score while jointly optimizing prediction and rationale generation during training. We selected the checkpoint with the highest validation F1. The models were trained using publicly available packages from https://github.com/huggingface/transformers (accessed on 2 February 2026) [29].

Table 1. Training configuration by model size. Effective batch size is Per-GPU Batch × #GPUs × Grad Accum.

Model	Per-GPU Batch	Grad Accum
T5-Small	16	1
T5-Base	4	4

3.4. Baselines

We compared R2D-C against the following:

- Distilling Step-by-Step (DSS): Multi-task training following [7]. Model selection was by validation loss (as in [7]). Implementation followed the authors' release.
- Reason2Decide (R2D): Two-stage rationale training with task-level scheduled sampling [8]. Model selection was by validation F1 (as in [8]). Implementation followed the authors' release.
- Zero-shot LLMs: Zero-shot baselines without task-specific fine-tuning, using open-source models including Qwen-3-8B and Qwen-3-32B [24], as well as BioMistral-7B [30] and OpenBioLLM-8B [31].

Protocol: For all fine-tuned baselines, we used the same optimizer, learning rate, effective batch size as our method; per-device batch sizes were set following Table 1. Early stopping was used with patience of 5 validation evaluations. Model selection was by validation Macro F1, except for DSS, where we used validation loss following the author's implementation. We report means over three seeds/runs. We used greedy decoding to ensure run-to-run determinism.

3.5. Evaluation

We report two categories of metrics: label prediction accuracy on the primary task and rationale–label alignment, which measures how well generated rationales support model decisions.

Predictive Performance: We report Macro F1, computed over gold test labels y^*, on the discrete label space (dispositions for Triage; Yes/No for PubMedQA; pathology for DDXPlus; correct option choice for MedReason). Scores are computed on held-out test sets.

Rationale–Label Alignment: R2D [8] was employed to evaluate rationale quality using surface-form generation metrics like BERTScore [32] and BLEU [33]. In this work, we

instead focus on whether generated rationales align with model decisions. To do so, we report the following alignment-oriented proxy metrics:

- **Consistency (P1 = P2)**: This metric is motivated by [34] and measures the agreement between the label predicted from the original input (P1) and the label predicted from the generated rationale alone (P2). High consistency indicates that the rationale preserves the model's decision logic, serving as a proxy for prediction-rationale alignment.
- **Sufficiency (P2 = GT)**: Sufficiency evaluates whether the generated rationale $\hat{r}$ contains sufficient information to recover the gold label (GT) without access to the original input. Given only $\hat{r}$, the model is asked to predict the label P2. This metric is motivated by the ERASER benchmark protocol for evaluating rationalized predictions [35]. This measures label recoverability from the generated rationale but does not by itself establish that the rationale reflects the model's full internal causal reasoning.
- **Gold Sufficiency (GR $\rightarrow$ GT)**: This metric measures whether providing the gold rationale (GR) to the model causes the model to output the gold label (GT). If GR $\rightarrow$ GT is low, the model struggles to recover the label even from a perfect rationale, indicating a model-side limitation rather than a failure of the generated rationale.
- **LLM-as-a-Judge Correctness**: Recent work has demonstrated strong alignment between LLM and human evaluation [36,37]. With this motivation, we employed Qwen-3-8B and Qwen-3-32B [24] models for expert-style assessment on a random sample of 2000 examples from the Clinical Triage test set, without stratification by class, prediction correctness, or model confidence. Because this subset was not stratified, the resulting judge scores should be interpreted as approximate subset-level evidence rather than a controlled robustness analysis across label, error, or confidence strata. We used a 5-point scale for clinical alignment between the rationales and predictions. The motivation is "If the predicted disposition is `Go to L&D now`, does the generated rationale justify the decision, rather than supporting a different disposition (e.g., `homecare?`)" To ensure consistent evaluation, we generated standardized disposition definitions using prompt-based refinement with a Qwen-3-32B model.

We report LLM-as-a-Judge Correctness only for the Triage dataset. We omit this metric for PubMedQA because it is a simple `yes/no` task, similarly for MedReason because its label space is not exhaustive and for DDXPlus because standardized label definitions are not available. Moreover, these metrics are intended to assess prediction-rationale alignment and label recoverability, but they do not directly measure whether a rationale improves clinician decision-making, trust calibration, or workflow usefulness in practice.

4. Results and Discussion

4.1. Predictive Performance and Rationale Alignment

Table 2 reports predictive performance and rationale–label alignment across all datasets and model sizes. Overall, the results support our central hypothesis that R2D-C preserves the predictive behavior of R2D while producing substantially better-aligned rationales. Provided the 95% confidence intervals for Macro F1, we interpret small F1 differences cautiously and focus our main claims on the much larger and more consistent gains in alignment-oriented metrics. The clearest improvements appear in alignment-oriented metrics, suggesting that confidence-adaptive scheduling and cycle-consistent training are most effective at strengthening the connection between predictions and explanations rather than merely improving prediction performance in isolation. Because rationale provenance differs across benchmarks, alignment metrics should be interpreted primarily within each dataset rather than assuming a uniform rationale regime across all four tasks.

Table 2. Model performance comparison across datasets. The best score per dataset is **bold + underlined**. The best fine-tuning strategy per model size is **bold**. For fine-tuned methods, Macro F1 is reported as mean ± standard deviation, with 95% confidence intervals shown in brackets. Instances where rationale generation is non-applicable are shown as –.

Dataset	Model	Method	Macro F1	Consistency	Sufficiency	Gold Sufficiency
Clinical Triage	T5-Small	R2D	55.88 ± 0.01 [55.86, 55.90]	39.64 ± 2.95	36.00 ± 2.00	41.47 ± 2.51
		R2D-C	**56.31 ± 0.30 [55.56, 57.06]**	**81.30 ± 0.35**	**55.10 ± 0.59**	**77.87 ± 0.19**
		DSS	52.73 ± 0.99 [50.27, 55.19]	42.36 ± 0.35	39.48 ± 0.26	47.02 ± 0.23
	T5-Base	R2D	59.92 ± 0.42 [58.88, 60.96]	45.28 ± 5.62	37.59 ± 5.08	46.61 ± 5.89
		R2D-C	**60.25 ± 0.30 [59.50, 61.00]**	**86.26 ± 0.45**	**59.38 ± 0.32**	**82.86 ± 0.16**
		DSS	53.26 ± 0.89 [51.05, 55.47]	45.49 ± 3.10	40.38 ± 2.98	50.13 ± 4.58
	Qwen-3-8B	Zero-Shot	23.88 ± 0.00	–	–	–
	Qwen-3-32B		33.08 ± 0.00	–	–	–
	BioMistral-7B		6.45 ± 0.00	–	–	–
	OpenBioLLM-7B		10.33 ± 0.00	–	–	–
PubMedQA	T5-Small	R2D	85.65 ± 0.21 [85.13, 86.17]	68.86 ± 6.02	68.15 ± 4.53	68.55 ± 2.66
		R2D-C	**85.90 ± 0.12 [85.60, 86.20]**	**90.52 ± 3.15**	**81.94 ± 1.70**	78.55 ± 1.97
		DSS	84.39 ± 0.48 [83.20, 85.58]	81.45 ± 3.00	77.15 ± 1.88	**80.61 ± 5.01**
	T5-Base	R2D	89.67 ± 0.31 [88.90, 90.44]	64.78 ± 2.01	64.03 ± 1.28	67.14 ± 3.20
		R2D-C	**89.80 ± 0.21 [89.28, 90.32]**	**96.42 ± 0.27**	**87.71 ± 0.19**	**83.04 ± 3.48**
		DSS	89.30 ± 0.43 [88.23, 90.37]	61.63 ± 0.42	62.39 ± 0.19	62.27 ± 2.24
	Qwen-3-8B	Zero-Shot	84.24 ± 0.00	–	–	–
	Qwen-3-32B		**90.67 ± 0.00**	–	–	–
	BioMistral-7B		29.44 ± 0.00	–	–	–
	OpenBioLLM-7B		46.22 ± 0.00	–	–	–
DDXPlus	T5-Small	R2D	**99.56 ± 0.05 [99.44, 99.68]**	4.14 ± 3.17	4.12 ± 3.17	4.56 ± 3.40
		R2D-C	99.52 ± 0.00 [99.52, 99.52]	**92.95 ± 5.56**	**92.60 ± 5.55**	**91.96 ± 5.74**
		DSS	99.29 ± 0.14 [98.94, 99.64]	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01
	T5-Base	R2D	**99.53 ± 0.26 [98.88, 100.00]**	2.37 ± 1.64	2.37 ± 1.64	2.58 ± 1.73
		R2D-C	99.39 ± 0.23 [98.82, 99.96]	**99.53 ± 0.17**	**99.32 ± 0.16**	**99.17 ± 0.09**
		DSS	99.40 ± 0.31 [98.63, 100.00]	5.04 ± 5.27	5.08 ± 5.31	5.13 ± 5.17
	Qwen-3-8B	Zero-Shot	36.64 ± 0.00	–	–	–
	Qwen-3-32B		49.44 ± 0.00	–	–	–
	BioMistral-7B		5.64 ± 0.00	–	–	–
	OpenBioLLM-7B		24.34 ± 0.00	–	–	–
MedReason	T5-Small	R2D	16.93 ± 0.79 [14.97, 18.89]	0.02 ± 0.04	0.02 ± 0.04	0.00 ± 0.00
		R2D-C	**16.96 ± 0.81 [14.95, 18.97]**	**2.02 ± 0.34**	**0.97 ± 0.21**	**0.09 ± 0.03**
		DSS	15.92 ± 0.19 [15.45, 16.39]	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	T5-Base	R2D	18.86 ± 0.16 [18.46, 19.26]	0.05 ± 0.04	0.00 ± 0.00	0.00 ± 0.00
		R2D-C	**18.88 ± 0.19 [18.41, 19.35]**	**24.58 ± 13.19**	**8.58 ± 3.96**	**0.31 ± 0.10**
		DSS	15.15 ± 0.34 [14.31, 15.99]	0.04 ± 0.08	0.07 ± 0.07	0.05 ± 0.04
	Qwen-3-8B	Zero-Shot	46.87 ± 0.00	–	–	–
	Qwen-3-32B		**77.29 ± 0.00**	–	–	–
	BioMistral-7B		37.59 ± 0.00	–	–	–
	OpenBioLLM-7B		53.01 ± 0.00	–	–	–

On Clinical Triage, R2D-C provides the strongest overall evidence for this claim. For T5-Small, predictive performance changes from 55.88 to 56.31, but consistency increases significantly from 39.64 to 81.30, sufficiency from 36.00 to 55.10, and gold sufficiency from 41.47 to 77.87. The same pattern holds for T5-Base, where macro F1 changes from 59.92 to 60.25, while consistency rises from 45.28 to 86.26, sufficiency from 37.59 to 59.38, and gold sufficiency

from 46.61 to 82.86. These results indicate that R2D-C improves whether generated rationales actually support the model's decision, while leaving predictive performance broadly intact. From a clinical perspective, these gains are important because they reduce the risk that the model presents a rationale that supports a different care recommendation than the one being predicted. In decision-support settings such as telephone triage, this kind of decision–rationale mismatch can undermine clinician trust and make model outputs harder to review safely.

On PubMedQA, the same pattern holds although in a less clinically rich setting. R2D-C again shows broadly comparable Macro F1 to R2D, with only small changes in predictive performance, while delivering much larger gains in alignment. Relative to R2D, T5-Small changes from 85.65 to 85.90 Macro F1 and T5-Base from 89.67 to 89.80, whereas consistency rises from 68.86 to 90.52 for T5-Small and from 64.78 to 96.42 for T5-Base, and sufficiency rises from 68.15 to 81.94 for T5-Small and from 64.03 to 87.71 for T5-Base. Compared with DSS, R2D-C remains stronger on the alignment-oriented metrics that matter most to our formulation, even when DSS performs slightly better on gold sufficiency with the T5-Small. This again suggests that R2D-C improves rationale–label coupling rather than merely changing surface rationale quality.

On DDXPlus, predictive accuracy is already saturated for all methods, with all models reaching roughly 99.3–99.6 macro F1. In this setting, predictive performance alone would suggest little difference between methods. However, the alignment metrics reveal a significant discrepancy: R2D and DSS remain near zero on consistency, sufficiency, and gold sufficiency, whereas R2D-C reaches 92.95/92.60/91.96 with T5-Small and 99.53/99.32/99.17 with T5-Base. DDXPlus, therefore, highlights why alignment metrics are necessary: when label prediction is near the ceiling, the main contribution of R2D-C appears in whether rationales remain strongly coupled to the underlying decision.

On MedReason, all fine-tuned methods achieve low absolute predictive performance, and we do not view this as evidence against the alignment objective itself. Rather, MedReason appears structurally different from the other benchmarks: its label space is non-exhaustive, and the task more closely resembles zero-shot medical question answering than closed-label classification since the effective answer space varies across questions. In such a setting, larger pretrained models have an inherent advantage because of their broad pretrained medical knowledge. Moreover, because the label space is non-exhaustive, fine-tuned models observe relatively few instances of any single label, making it harder to learn stable input-label associations. Even with this challenging setup, R2D-C still provides directional improvements in rationale–label alignment over both R2D and DSS, improving consistency from 0.02 to 2.02 for T5-Small and from 0.05 to 24.58 for T5-Base. We therefore interpret MedReason as a setting where backbone knowledge is the main bottleneck, while alignment-focused training still offers modest but consistent benefits.

4.2. LLM-as-a-Judge Evaluation

Table 3 presents the LLM-as-a-Judge results on the Clinical Triage dataset. R2D-C achieves the best scores for both T5-Small (4.86 with Qwen-3-8B; 4.44 with Qwen-3-32B) and T5-Base (4.87; 4.48), outperforming both R2D and DSS. Although these improvements are smaller in magnitude than the gains in consistency and sufficiency, they suggest that stronger internal alignment also translates into better judged alignment between rationale and predicted disposition. We note, however, that improved external judgment of justification quality still does not directly establish mechanistic faithfulness; rather, it provides complementary evidence that the rationales are more coherent and decision-supportive.

Taken together, the results show that R2D-C should not be understood primarily as a method for maximizing predictive accuracy. Instead, its main contribution is to preserve the predictive behavior of R2D while making rationales substantially more aligned with the

decisions they are meant to explain. This is precisely the desired outcome in high-stakes settings such as clinical NLP, where a model should not only predict well but also produce rationales that reliably align with, and justify its predictions.

Table 3. LLM-as-a-Judge Correctness on the Clinical Triage dataset. Best score per model is **bold**.

Model	Method	8B Judge	32B Judge
T5-Small	R2D	4.74 ± 0.02	4.18 ± 0.01
	R2D-C	**4.86 ± 0.02**	**4.44 ± 0.01**
	DSS	4.64 ± 0.03	3.99 ± 0.02
T5-Base	R2D	4.79 ± 0.02	4.34 ± 0.01
	R2D-C	**4.87 ± 0.03**	**4.48 ± 0.01**
	DSS	4.66 ± 0.02	4.07 ± 0.02

5. Ablation Studies

We conducted two sets of ablations to better understand the behavior of R2D-C. First, we isolated the contributions of its two core components: confidence-adaptive scheduling (CAS) and the rationale-to-label cycle-consistency loss. Second, we ablated the soft rationale construction used within the cycle branch, comparing the default Gumbel-Softmax formulation against deterministic softmax alternatives.

5.1. Ablation of Core Components

To isolate the contributions of the two main components of R2D-C, we conducted ablations on confidence-adaptive scheduling (CAS) and the rationale-to-label cycle-consistency loss. In particular, *w/o cycle* removes the cycle objective while retaining CAS, and *w/o* CAS removes confidence-adaptive scheduling while retaining the cycle objective. As demonstrated in Table 4 across datasets, predictive performance remains broadly stable across ablations, while the largest differences appear in rationale–label alignment, consistent with our goal of preserving the predictive behavior of R2D while improving prediction–rationale coupling.

On Clinical Triage, the full R2D-C model achieves the strongest overall internal alignment results. Removing the cycle loss causes large drops in consistency, sufficiency, and gold sufficiency relative to both *w/o* CAS and the full model, showing that cycle consistency is the main mechanism enforcing rationale–label coupling. At the same time, removing CAS also weakens alignment relative to the full model, indicating that CAS provides complementary gains on this clinically realistic and uncertainty-sensitive task. The LLM-as-a-Judge results show that variants with CAS consistently outperform *w/o* CAS on Clinical Triage, suggesting that CAS improves externally judged clinical justification quality. Taken together, these results suggest that cycle consistency drives the largest gains in internal alignment metrics, while CAS provides additional benefits that are especially visible in externally judged rationale quality.

On PubMedQA, the ablations indicate that cycle consistency is again the dominant factor. Removing the cycle loss substantially reduces consistency, sufficiency, and gold sufficiency, whereas the *w/o* CAS variant consistently matches or slightly exceeds the full model across all alignment metrics. This suggests that in a simpler binary QA setting, once the rationale is explicitly constrained to recover the label, CAS offers only limited additional benefit. In other words, cycle consistency appears sufficient to obtain most of the alignment gains on PubMedQA.

On DDXPlus, the pattern is even clearer. Removing the cycle loss causes alignment to collapse almost completely despite predictive accuracy remaining near the ceiling, demonstrating that strong task performance alone does not imply that rationales are meaningfully connected to model decisions. By contrast, the *w/o CAS* variant remains

near-perfect on all alignment metrics and slightly exceeds the full model. This indicates that for highly structured, near-deterministic tasks, cycle consistency is essential, whereas CAS is less necessary once rationale-to-label recoverability is enforced.

Table 4. Ablation results for confidence-adaptive scheduling (CAS) and cycle consistency. Best score per model is **bold**.

Dataset	Model	Method	Macro F1	Consistency	Sufficiency	Gold Sufficiency	8B Judge	32B Judge
Clinical Triage	T5-Small	*w/o* cycle	56.41	41.71	36.41	41.56	**4.90**	**4.44**
		w/o CAS	**56.49**	69.62	52.21	74.35	4.78	4.19
		R2D	55.88	39.64	36.00	41.47	4.74	4.18
		R2D-C	56.31	**81.30**	**55.10**	**77.87**	4.86	**4.44**
	T5-Base	*w/o* cycle	59.69	51.42	41.17	50.46	**4.88**	4.47
		w/o CAS	59.75	79.01	57.44	81.53	4.80	4.35
		R2D	59.92	45.28	37.59	46.61	4.79	4.34
		R2D-C	**60.25**	**86.26**	**59.38**	**82.86**	4.87	**4.48**
PubMedQA	T5-Small	*w/o* cycle	85.64	71.62	69.85	69.86	–	–
		w/o CAS	85.87	**91.42**	**82.24**	**78.92**	–	–
		R2D	85.65	68.86	68.15	68.55	–	–
		R2D-C	**85.90**	90.52	81.94	78.55	–	–
	T5-Base	*w/o* cycle	**89.91**	66.07	64.53	67.60	–	–
		w/o CAS	89.64	**96.98**	**87.89**	**84.43**	–	–
		R2D	89.67	64.78	64.03	67.14	–	–
		R2D-C	89.80	96.42	87.71	83.04	–	–
DDXPlus	T5-Small	*w/o* cycle	99.53	6.70	6.67	6.98	–	–
		w/o CAS	99.55	**97.10**	**96.80**	**96.42**	–	–
		R2D	**99.56**	4.14	4.12	4.56	–	–
		R2D-C	99.52	92.95	92.60	91.96	–	–
	T5-Base	*w/o* cycle	99.56	4.09	4.09	4.11	–	–
		w/o CAS	**99.66**	**99.64**	**99.41**	**99.26**	–	–
		R2D	99.53	2.37	2.37	2.58	–	–
		R2D-C	99.39	99.53	99.32	99.17	–	–

Overall, the ablations justify the full R2D-C configuration as the default model, but they also reveal an important asymmetry between its two components. Cycle consistency is the most robust and consistently beneficial contributor to rationale–label alignment across datasets and model sizes. CAS provides complementary gains that are most valuable in harder and more uncertainty-sensitive settings such as Clinical Triage, where externally judged rationale quality also benefits from self-conditioning based on model confidence. For this reason, we adopted the full R2D-C model as our default configuration: it provides the strongest overall balance between predictive performance and rationale–label alignment on the most clinically relevant dataset, while remaining competitive across all settings.

5.2. Ablation on Soft Rationale Construction

We further ablated the differentiable relaxation used in the rationale-to-label cycle branch on the Clinical Triage dataset. In the full R2D-C model, soft rationale embeddings are constructed using Gumbel-Softmax with a decaying temperature schedule, which provides a differentiable approximation to discrete rationale tokens during cycle-consistent training. We compared this against two deterministic alternatives that keep the rest of the training objective unchanged: (i) a standard softmax relaxation and (ii) a temperature-scaled softmax relaxation with a decaying temperature schedule. This isolates whether the gains of R2D-C depend specifically on the Gumbel-Softmax approximation or more generally on differentiable soft rationale reconstruction.

Table 5 presents the results. Overall, the full R2D-C model provides the strongest results across the internal alignment metrics. For T5-Small, R2D-C achieves the best consistency, sufficiency, gold sufficiency, and 8B judge score, while matching the best 32B Judge score; regular softmax is only slightly higher on Macro F1. For T5-Base, R2D-C again achieves the strongest Macro F1 and substantially outperforms both deterministic alternatives on consistency, sufficiency, and gold sufficiency. However, temperature-scaled softmax scores slightly higher on the judge-based metrics. These results suggest that the

main benefit of Gumbel-Softmax lies in improving rationale–label alignment rather than externally judged plausibility alone.

Comparing the two deterministic relaxations, temperature-scaled softmax is generally stronger than regular softmax, especially for T5-Base. This indicates that progressively sharpening the rationale distribution is preferable to using a fixed smooth relaxation, even without stochastic sampling. However, neither deterministic alternative matches the full Gumbel-Softmax formulation on the alignment-oriented metrics that most directly reflect our objective. We therefore retained Gumbel-Softmax in R2D-C as our default differentiable relaxation since it provides the best overall balance between predictive performance and rationale–label alignment.

Table 5. Ablation of differentiable rationale relaxation on the Clinical Triage dataset. Best score per model is **bold**.

Model	Method	Macro F1	Consistency	Sufficiency	Gold Sufficiency	8B Judge	32B Judge
T5-Small	regular_softmax	**56.45**	80.67	54.99	76.95	4.84	**4.44**
	temperature_softmax	56.34	80.83	54.61	77.27	4.84	**4.44**
	R2D-C	56.31	**81.30**	**55.10**	**77.87**	**4.86**	**4.44**
T5-Base	regular_softmax	59.78	76.26	53.84	73.60	4.87	4.48
	temperature_softmax	59.53	78.54	54.67	74.20	**4.88**	**4.50**
	R2D-C	**60.25**	**86.26**	**59.38**	**82.86**	4.87	4.48

5.3. Sensitivity to the Cycle-Consistency Weight

Because the rationale-to-label cycle loss is a central part of R2D-C, we also performed a sensitivity analysis over the cycle-consistency weight λ_{cyc} for T5-Small on Clinical Triage and PubMedQA. We varied $\lambda_{\text{cyc}} \in \{0, 0.1, 0.2, 0.5\}$.

Table 6 shows that adding cycle consistency ($\lambda_{\text{cyc}} > 0$) consistently improves alignment-oriented metrics relative to $\lambda_{\text{cyc}} = 0$, while predictive performance remains broadly stable. On Clinical Triage, larger λ_{cyc} values continue to improve consistency and gold sufficiency slightly, but they also introduce a modest decline in Macro F1, indicating a clearer alignment-performance trade-off. On PubMedQA, the model is relatively insensitive within the tested non-zero range: Macro F1 changes only slightly, while consistency and sufficiency improve monotonically as λ_{cyc} increases.

Taken together, these results support two conclusions. First, the benefits of cycle consistency are robust and do not depend on a single finely tuned weight. Second, $\lambda_{\text{cyc}} = 0.1$ remains a reasonable default because it already captures most of the alignment gains while keeping predictive performance close to the $\lambda_{\text{cyc}} = 0$ setting, especially on the more clinically realistic and uncertainty-sensitive Clinical Triage task.

Table 6. Sensitivity analysis for the cycle-consistency weight λ_{cyc}. Best score per dataset is **bold**.

Dataset	λ_{cyc}	Macro F1	Consistency	Sufficiency	Gold Sufficiency
Clinical Triage	0.0	**56.41**	41.71	36.41	41.56
	0.1	56.31	81.30	55.10	77.87
	0.2	55.78	82.04	**55.40**	78.89
	0.5	55.57	**83.48**	55.28	**79.37**
PubMedQA	0.0	85.64	71.62	69.85	69.86
	0.1	**85.90**	90.52	81.94	78.55
	0.2	85.79	91.63	82.00	78.53
	0.5	85.81	**94.13**	**83.30**	**80.02**

6. Limitations

Although R2D-C achieves its design goal of retaining predictive performance of R2D while improving rationale–label alignment, our study has several limitations. First, the ablation results indicate that the relative importance of confidence adaptive scheduled sampling (CAS) and cycle consistency is task-dependent: cycle consistency is the dominant

factor on PubMedQA and DDXPlus, whereas CAS provides its clearest additional benefits on Clinical Triage.

Second, although we evaluate rationale quality using consistency, sufficiency, gold sufficiency, and LLM-as-a-Judge scores, these metrics do not directly test mechanistic faithfulness and still provide only partial views of explanation quality. More broadly, automatic metrics cannot fully capture whether a rationale is clinically useful or reliable in real-world deployment. For high-stakes domains like clinical NLP, human verification remains essential. We did not conduct clinician- or domain-expert evaluation of the generated rationales in this study and therefore do not claim expert-validated explanation quality.

Third, our approach involved several hyperparameters that were selected heuristically. While we included a targeted sensitivity analysis for λ_{cyc} on Clinical Triage and PubMedQA, a more extensive hyperparameter optimization strategy may yield further improvements in predictive performance and prediction–rationale alignment.

Finally, a substantial portion of our rationale supervision is LLM-generated rather than human-authored. In particular, the Clinical Triage and DDXPlus datasets rely on rationales produced by Qwen-3-8B. While practical, this setup introduces potential circularity: the model may learn to reproduce teacher-style explanatory patterns, and high rationale quality with our metrics does not necessarily imply correspondence to clinician reasoning or to a model's true internal decision process. We therefore interpret these rationales as useful supervisory signals but not as definitive gold explanations.

7. Ethics Statement

This work is situated in the context of clinical NLP, where model outputs may influence high-stakes decisions. The proprietary Clinical Triage dataset used in this study was de-identified, and no personally identifiable information was accessible to the models or researchers during training or evaluation. In addition, all experiments were conducted using open-source models in a controlled local computing environment. Any predictions and rationales generated by R2D-C should be treated as decision-support signals requiring human verification, rather than as replacements for clinical judgment.

8. Conclusions and Future Work

We presented R2D-C, a rationale-driven extension of R2D that combines confidence-adaptive scheduled sampling with rationale-to-label cycle-consistent training to improve prediction–rationale coupling without sacrificing predictive performance. Across the evaluated datasets, the main gains of R2D-C come from improved internal alignment, particularly consistency, sufficiency, and gold sufficiency, while predictive performance remains broadly stable.

Our ablations show that, in our experiments, the full R2D-C formulation provides the strongest overall balance between predictive performance and rationale–label alignment. Taken together, these results support a broader view of rationales as objects that should be explicitly optimized for prediction–rationale alignment, rationale sufficiency, and decision-support usefulness, rather than being treated purely as post hoc textual artifacts.

There are several promising directions for future work. One direction is to make the training strategy more adaptive, for example, by learning when to apply self-conditioning and how strongly to weight the cycle objective. Another is to evaluate the approach with larger and more instruction-tuned backbones, as well as more diverse clinical and biomedical datasets. A final direction is to complement automatic faithfulness metrics with stronger expert-facing evaluation protocols that assess whether generated rationales are clinically meaningful and decision-supportive in practice.

https://doi.org/10.3390/computers15050279

Author Contributions: Conceptualization, H.M.Q.H.; methodology, H.M.Q.H.; software, H.M.Q.H.; validation, H.M.Q.H., H.K.B.B., M.-Y.K. and R.G.; formal analysis, H.M.Q.H.; investigation, H.M.Q.H.; resources, H.M.Q.H., H.K.B.B., M.-Y.K. and R.G.; data curation, H.M.Q.H., M.-Y.K. and R.G.; writing—original draft preparation, H.M.Q.H.; writing—review and editing, H.M.Q.H., H.K.B.B., M.-Y.K. and R.G.; visualization, H.M.Q.H.; supervision, H.K.B.B., M.-Y.K. and R.G.; project administration, H.K.B.B. and M.-Y.K.; funding acquisition, M.-Y.K. and R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Training Experience (CREATE) program "From Data to Decision" (FD2D). This research was also supported by the Alberta Machine Intelligence Institute (Amii), NSERC (including grants DGECR-2022-00369 and RGPIN-2022-0346), and Alberta Innovates (Enabling Better Health through Artificial Intelligence (AI-Better Health) Program).

Institutional Review Board Statement: The use of the Clinical Triage dataset was approved by the University of Alberta Human Research Ethics Board (Project ID: Pro00081689, approved on 19 April 2019).

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to the proprietorial nature of the Clinical Triage dataset, it cannot be made publicly available. The other preprocessed datasets can be found in the source code's GitHub repository.

Acknowledgments: We gratefully acknowledge Alberta Health Services (AHS), and in particular Patricia Chambers and Jane Q. Huang, for providing access to the Alberta Health Link 811 dataset for research purposes.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompts Used

Prompt for Generating Standardized Disposition Definitions:

```
You are building a clinical definition for the disposition: [DISPOSITION_NAME].

Current working definition: """[CURRENT_DEFINITION]"""

You are given new examples of triage notes and rationales for this disposition.
Use them to refine, expand, or correct the working definition. Keep the definition
concise but clinically accurate.

Triage Notes and Rationales: [EXAMPLES]

Update the definition:
- Incorporate any new key symptoms, criteria, or thresholds.
- Remove incorrect parts.
- Keep it as clear and specific as possible.
Output only the revised definition text, nothing else.
```

[DISPOSITION_NAME], [CURRENT_DEFINITION], and [EXAMPLES] were replaced with actual data. This process generated the standardized definitions used for consistent LLM-as-a-Judge Correctness scoring.

Prompt for Zero-Shot Disposition Classification:

> Issue: [ISSUE_ASSESSMENT]
> Dispositions: [CLASSES_TEXT]
>
> Classify the healthcare issue into one of the dispositions above. Return your answer in the following **strict** format:
> Class: [chosen digit]
>
> Do not ask for more information, and do not provide any general statements. Only respond with the digit.

[ISSUE_ASSESSMENT] contained the triage note text, and [CLASSES_TEXT] listed the 12 disposition options with their numerical identifiers. This prompt was used for zero-shot LLM evaluation.

Prompt for LLM-as-a-Judge Correctness Scoring:

> You are an expert clinical trainer for telephone triage nursing.
> [DEFINITIONS]
>
> Task: Given the following rationale and disposition, score the alignment on a scale of 1 to 5, where:
> 5—Excellent Alignment
> 4—Good Alignment
> 3—Moderate Alignment
> 2—Poor Alignment
> 1—Very Poor Alignment
>
> Rationale: [RATIONALE]
> Disposition: [DISPOSITION]
> Output exactly one number (1, 2, 3, 4, or 5) with no other text.

[DEFINITIONS] was replaced with the standardized disposition definitions following Appendix B, [RATIONALE] with the generated rationale, and [DISPOSITION] with the predicted disposition.

Appendix B. Clinical Disposition Definitions

Home Care: Patients suitable for Home Care disposition have mild to moderate, stable, and non-progressive symptoms without signs of immediate or severe complications requiring emergency intervention. They exhibit no respiratory distress, hemodynamic instability, severe pain unresponsive to treatment, significant bleeding, high fever with systemic symptoms, spreading infection, altered mental status, neurological deficits, or other urgent clinical concerns. Their condition allows for safe management, symptom relief, and observation at home with appropriate follow-up. This includes stable minor injuries, controlled localized infections, mild chronic condition issues, and non-urgent questions or concerns. Patients and caregivers should be advised to seek urgent care if symptoms worsen or new concerning signs develop.

See Physician or PCP within 3 days: Patients with mild to moderate, stable but persistent or worsening symptoms that do not pose an immediate threat to life or function, and who require timely medical evaluation to prevent complications. This includes conditions without severe pain, respiratory distress, hemodynamic instability, significant neurological deficits, acute infection, or other urgent signs. The disposition excludes

any patients exhibiting signs of severe or rapidly progressing illness, acute psychiatric emergencies, or other urgent conditions necessitating immediate or emergency care.

Call Pharmacist within 24 h: This disposition is used when a caller has non-emergency medication-related questions or concerns that require timely pharmacist expertise within 24 h to ensure safe, effective, and appropriate medication use. It applies when immediate urgent care is not needed but professional assessment is necessary to clarify dosing, manage side effects, verify medication information, address potential interactions, support adherence, assist with medication access, or provide guidance on special populations and circumstances. This ensures optimized therapy, prevention of harm, and informed patient decisions without delay.

See Physician or PCP within 4 h (or PCP triage): Patients with new or worsening moderate to severe symptoms, signs of infection, or conditions at risk of rapid deterioration that require timely clinical evaluation within hours to prevent complications. This includes significant pain unrelieved by initial treatment, progressive neurological symptoms, signs of systemic infection, post-procedural complications, unstable vital signs, moderate to severe respiratory, abdominal, or urinary symptoms, pregnancy-related concerns, dehydration, metabolic instability, mental health deterioration, and other clinical presentations indicating potential for rapid decline. The disposition ensures prompt physician assessment to guide urgent management and prevent adverse outcomes.

Call EMS 911 Now: Initiate immediate emergency medical services activation for any patient presenting with signs or symptoms of a potentially life-threatening condition. This includes acute neurological deficits, severe chest pain or cardiac symptoms, significant respiratory distress, signs of anaphylaxis, major trauma or uncontrolled bleeding, severe abdominal or back pain with systemic symptoms, acute deterioration in patients with serious underlying conditions, active suicidal intent with risk of harm, critical illness in infants or young children, shock or imminent collapse, severe infection or sepsis, altered mental status or unresponsiveness, and any situation posing an immediate threat to life, safety, or vital functions. Prompt EMS activation is essential whenever there is concern for compromised airway, breathing, circulation, neurological status, or urgent mental health crisis requiring rapid intervention.

Go to ED Now: Immediate emergency department evaluation is required for patients presenting with sudden, severe, or rapidly worsening symptoms that pose an immediate risk to life, limb, or function. This includes active uncontrolled bleeding, significant head or neck injuries, severe respiratory distress or airway compromise, acute neurological deficits, severe or persistent pain suggestive of surgical or obstetric emergencies, signs of shock or altered consciousness, severe allergic reactions, prolonged seizures, suspected serious infections, severe metabolic disturbances, and any other critical conditions requiring urgent assessment and intervention.

See Physician or PCP within 2 Weeks: Patients with new, persistent, worsening, or recurrent symptoms that are stable and do not require emergency care but need timely medical evaluation to diagnose, monitor, or adjust treatment. This includes conditions without signs of acute distress, hemodynamic instability, severe pain, neurological deficits, respiratory compromise, systemic infection, or other urgent symptoms. The disposition applies to a broad range of non-emergent but concerning clinical presentations where prompt follow-up is necessary to prevent progression or complications.

Go to L&D Now: Immediate evaluation in Labor and Delivery is required for pregnant individuals presenting with signs of active labor at or near term, suspected or confirmed rupture of membranes, significant vaginal bleeding, decreased or absent fetal movement, new or worsening moderate to severe abdominal or pelvic pain, signs of preterm labor, pregnancy complications or risk factors combined with concerning symptoms, abdominal or

pelvic trauma, maternal conditions suggestive of serious complications (e.g., preeclampsia, infection, hemodynamic instability), or any other acute symptoms indicating potential maternal or fetal compromise. Prompt assessment is essential to ensure maternal and fetal safety.

Call Poison Center Now: Immediate expert toxicology consultation is required for any suspected or confirmed exposure to potentially harmful substances or situations with risk of significant toxicity, overdose, or complications. This includes exposures involving high-risk medications, chemicals, toxins, unknown or unlabeled agents, vulnerable populations (such as children or pregnant women), or any new or worsening symptoms suggestive of systemic toxicity. Prompt specialist guidance is essential to ensure safe management and appropriate treatment.

See More Appropriate Guideline: Use this disposition when the caller's concerns do not require emergency or urgent care but need assessment, advice, or management under a more specific, condition-focused guideline. It applies to stable, non-urgent symptoms or questions without red flags, including chronic conditions, mild new symptoms, or informational and care coordination needs. Avoid this disposition for any signs of acute deterioration, emergencies, or conditions requiring immediate intervention. Direct callers to the guideline that best matches their specific symptom or concern to ensure appropriate care.

See Physician or PCP within 24 h: Patients with new, worsening, or persistent moderate symptoms that impact daily activities but do not require emergency care. This includes localized signs of infection or inflammation without systemic involvement, moderate injuries without severe complications, mild to moderate neurological, respiratory, gastrointestinal, or mental health symptoms without acute distress, and other conditions needing timely medical evaluation to prevent deterioration, ensure appropriate management, and monitor progression.

Call Dentist when Office is Open: This disposition applies to patients with non-emergent dental issues that do not exhibit signs of serious infection, airway compromise, uncontrolled bleeding, or systemic illness. It includes mild to moderate pain or discomfort manageable with analgesics, stable post-procedure symptoms, minor dental trauma without active bleeding or severe pain, and dental appliance-related discomfort without urgent complications. Patients should seek dental care during regular office hours and be advised to obtain immediate emergency care if symptoms worsen or signs of a dental or medical emergency develop.

References

1. Gurrapu, S.; Kulkarni, A.; Huang, L.; Lourentzou, I.; Batarseh, F.A. Rationalization for explainable NLP: A survey. *Front. Artif. Intell.* **2023**, *6*, 1225093. [CrossRef]
2. Cross, J.L.; Choma, M.A.; Onofrey, J.A. Bias in medical AI: Implications for clinical decision-making. *PLoS Digit. Health* **2024**, *3*, e0000651. [CrossRef]
3. Lyu, Q.; Apidianaki, M.; Callison-Burch, C. Towards Faithful Model Explanation in NLP: A Survey. *arXiv* **2024**. [CrossRef]
4. Yang, J.; Glockner, M.; Rocha, A.; Gurevych, I. Self-Rationalization in the Wild: A Large Scale Out-of-Distribution Evaluation on NLI-related tasks. *arXiv* **2025**. [CrossRef]
5. Bhan, M.; Vittaut, J.N.; Chesneau, N.; Chandar, S.; Lesot, M.J. NeuroFaith: Evaluating LLM Self-Explanation Faithfulness via Internal Representation Alignment. *arXiv* **2026**. [CrossRef]
6. Madsen, A.; Chandar, S.; Reddy, S. Are self-explanations from Large Language Models faithful? In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Ku, L.W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 295–337. [CrossRef]
7. Hsieh, C.Y.; Li, C.L.; Yeh, C.K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.Y.; Pfister, T. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. *arXiv* **2023**. [CrossRef]
8. Hasan, H.M.Q.; Bashier, H.K.; Dai, J.; Kim, M.Y.; Goebel, R. Reason2Decide: Rationale-Driven Multi-Task Learning. *arXiv* **2025**. [CrossRef]

9. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**. [CrossRef]
10. Bengio, S.; Vinyals, O.; Jaitly, N.; Shazeer, N. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS'15), Montreal, QC, Canada, 7–12 December 2015*; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 1171–1179.
11. Schmidt, F. Generalization in Generation: A closer look at Exposure Bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, China, 4 November 2019*; Birch, A., Finch, A., Hayashi, H., Konstas, I., Luong, T., Neubig, G., Oda, Y., Sudoh, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 157–167. [CrossRef]
12. Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization with Gumbel-Softmax. *arXiv* **2017**. [CrossRef]
13. Kunz, J.; Kuhlmann, M. Properties and Challenges of LLM-Generated Explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing, Mexico City, Mexico, 21 June 2024*; Blodgett, S.L., Cercas Curry, A., Dev, S., Madaio, M., Nenkova, A., Yang, D., Xiao, Z., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 13–27. [CrossRef]
14. Atakishiyev, S.; Babiker, H.K.B.; Dai, J.; Farruque, N.; Hayashi, T.; Hriti, N.S.; Rahman, M.A.; Smith, I.; Kim, M.Y.; Zaïane, O.R.; et al. Explainability of Large Language Models: Opportunities and Challenges toward Generating Trustworthy Explanations. *arXiv* **2025**. [CrossRef]
15. Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
16. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv* **2017**. [CrossRef]
17. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**. [CrossRef]
18. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
19. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2023**, *21*, 5485–5551.
20. Liu, Y.; Meng, F.; Chen, Y.; Xu, J.; Zhou, J. Confidence-Aware Scheduled Sampling for Neural Machine Translation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2327–2337. [CrossRef]
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
22. Lewis, P.; Ott, M.; Du, J.; Stoyanov, V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online, 19 November 2020*; Rumshisky, A., Roberts, K., Bethard, S., Naumann, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 146–157. [CrossRef]
23. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef]
24. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report. *arXiv* **2025**. [CrossRef]
25. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019*; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2567–2577. [CrossRef]
26. Tchango, A.F.; Goel, R.; Wen, Z.; Martel, J.; Ghosn, J. DDXPlus: A New Dataset For Automatic Medical Diagnosis. *arXiv* **2022**. [CrossRef]
27. Wu, J.; Deng, W.; Li, X.; Liu, S.; Mi, T.; Peng, Y.; Xu, Z.; Liu, Y.; Cho, H.; Choi, C.I.; et al. MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs. *arXiv* **2025**. [CrossRef]
28. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**. [CrossRef]
29. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2020**. [CrossRef]
30. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.A.; Rouvier, M.; Dufour, R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *arXiv* **2024**. [CrossRef]
31. Ankit Pal, M.S. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. 2024. Available online: https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B (accessed on 27 February 2026).
32. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**. [CrossRef]

33. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002*; Isabelle, P., Charniak, E., Lin, D., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
34. Dasgupta, S.; Frost, N.; Moshkovitz, M. Framework for Evaluating Faithfulness of Local Explanations. In *Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 7–23 July 2022*; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; Proceedings of Machine Learning Research; PMLR: Cambridge, MA, USA, 2022; Volume 162, pp. 4794–4815.
35. DeYoung, J.; Jain, S.; Rajani, N.F.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B.C. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv* **2020**. [CrossRef]
36. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv* **2023**. [CrossRef]
37. Niu, S.; Ma, J.; Lin, H.; Bai, L.; Wang, Z.; Xu, Y.; Song, Y.; Yang, X. Knowledge-Augmented Multimodal Clinical Rationale Generation for Disease Diagnosis with Small Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, 27 July–1 August 2025*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 11011–11024. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.