國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

LAREV：具洩漏感知的推理評估方法——以條件 V-資訊為基礎

LAREV: Leakage-Aware Rationale Evaluation with Conditional V-information

李泓賢

Hong-Sian Li

指導教授: 許永真 博士

陳縕儂 博士

Advisor: Jane Yung-jen Hsu Ph.D.

Yun-Nung Chen Ph.D.

中華民國 115 年 1 月

January, 2026

doi:10.6342/NTU202600527

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

LAREV：具洩漏感知的推理評估方法——以條件 V-資訊為基礎

LAREV: Leakage-Aware Rationale Evaluation with Conditional V-information

本論文係 李泓賢 （學號 R12944062）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國 115 年 1 月 22 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Networking and Multimedia on 22 January 2026 have examined a Master's Thesis entitled above presented by HONG-SIAN LI (student ID: R12944062) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

許永真 陳縕儂

（指導教授 Advisor）

林英嘉

系（所）主管 Director: 鄭卜壬

Acknowledgements

首先，衷心感謝指導教授許永真老師。老師在研究過程中始終給予耐心且細緻的指導，並提供我充足的空間進行探索與嘗試；同時，也在關鍵時刻適時引導方向，使我能重新釐清研究問題，並逐步建立清楚的研究脈絡。特別感謝老師在研究方向上的提醒，記得在一次研究室會議中，我曾報告一篇與 DPO 相關的論文，並一度將其視為研究方向，然而當時對於自己真正想解決的研究問題，其實仍缺乏清楚的認識。老師透過提問提醒我，研究應先明確界定欲解決的問題，再進一步思考適合採用的方法。這樣的提醒使我重新檢視自己的研究動機與思考順序，也成為我後來重新思考並逐步聚焦研究方向的重要轉折點。能在老師的指導下完成本論文，深感幸運與感激。

此外，誠摯感謝共同指導教授陳縕儂老師，以及口試委員陳信希教授、林英嘉教授。各位老師以豐富的學術經驗與專業視角，提供了許多具體且寶貴的建議，使本論文在內容與論述的呈現方式上得以更加完善。在此特別感謝老師們於審閱與口試過程中所給予的細心指正與鼓勵。

感謝一路以來關心與支持我的朋友們。你們或許未必了解研究內容的細節，卻始終願意傾聽、陪伴，並在我需要暫時抽離研究壓力時，給予適時的關心與鼓勵。這些看似平凡的陪伴，對我而言卻是支撐自己持續完成研究的重要力量。

最後，感謝我的家人始終給予我充分的信任與支持，使我能在這段時間專心投入研究，無後顧之憂。你們的理解與包容，是我得以安心前行的重要基石。

也誠實地感謝自己，在面對不確定與壓力時，選擇持續思考與嘗試，而非輕易放棄。回顧這段研究歷程，所獲得的不僅是一篇論文，更是一段重新學習如何面對問題、釐清方向的過程。期許未來的自己，能帶著這份經驗，持續探索與成長。

doi:10.6342/NTU202600527

摘要

隨著大型語言模型在各類推理任務中廣泛採用，模型所產生之自然語言推理（rationale）已成為輔助決策與提升可解釋性的重要工具。然而，如何客觀評估推理文本本身是否真正提供額外、且與標籤相關的有用資訊，仍是一項具挑戰性的研究問題。近期研究提出以條件 V-資訊（Conditional V-information）為基礎的推理評估方法，在給定基準輸入（baseline input）的條件下，比較模型在有無推理文本時的預測行為，以估計推理相對於基準所帶來的可用資訊量。然而，在實務資料集中，此類基準輸入本身往往已包含高度的標籤洩漏（label leakage），使得評估模型可能過度依賴基準訊號，進而錯估推理品質，降低評估結果的可靠性。

本論文提出 LAREV（Leakage-Aware Rationale Evaluation with Conditional V-information），一個具洩漏感知能力的推理評估框架，旨在提升條件式推理評估在基準輸入有洩漏的情境下之穩健性。LAREV 在不改變原有評估指標形式的前提下，透過兩項關鍵設計約束評估模型的學習行為：其一，利用不變風險最小化（Invariant Risk Minimization, IRM）降低模型對基準輸入中特定線索的依賴；其二，引入洩漏探測模型以量化並懲罰評估模型從基準輸入中提取標籤相關資訊的能力，從而引導模型將預測改善歸因於推理文本本身。

實驗結果顯示，在 ECQA 與 e-SNLI 等具代表性的推理資料集中，LAREV 能有效拉大高品質推理與低品質推理之間的評估差距，並展現出較既有方法更穩定且具辨識力的排序行為。此外，分析結果亦顯示，LAREV 在降低由基準輸入中洩漏的依賴性之同時，仍能維持具競爭力的預測表現，驗證其作為一種可靠推理評估框架的實用性。

關鍵字：推理評估、標籤洩漏、洩漏感知、資訊理論、條件 V-資訊、大型語言模型

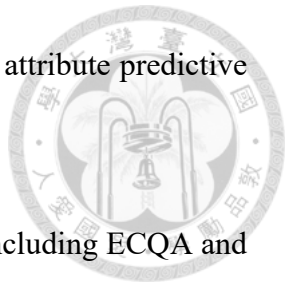
doi:10.6342/NTU202600527

Abstract

With the widespread adoption of large language models in reasoning tasks, natural language rationales generated by models have become an important tool for supporting decision-making and improving interpretability. However, objectively evaluating whether a rationale truly provides additional, label-relevant information remains a challenging research problem. Recent work has proposed rationale evaluation methods based on Conditional V-information, which estimate the usable information contributed by a rationale by comparing model predictions with and without access to the rationale, conditioned on a given baseline input. In practice, however, such baseline inputs often already contain substantial label leakage, causing evaluation models to over-rely on baseline signals and systematically misestimate rationale quality, thereby undermining the reliability of evaluation results.

In this thesis, we propose LAREV (Leakage-Aware Rationale Evaluation with Conditional V-information), a leakage-aware rationale evaluation framework designed to improve the robustness of baseline-conditioned evaluation under baseline leakage. Without altering the original evaluation metric, LAREV introduces two key training constraints on the evaluator model. First, it applies Invariant Risk Minimization (IRM) to reduce the model's reliance on spurious, baseline-specific cues. Second, it incorporates a leakage probing model to quantify and penalize the evaluator's ability to extract label-relevant in-

doi:10.6342/NTU202600527



formation from the baseline input, thereby encouraging the model to attribute predictive improvements to the rationale itself.

Experimental results on representative reasoning benchmarks, including ECQA and e-SNLI, demonstrate that LAREV effectively enlarges the evaluation gap between high-quality and low-quality rationales, and exhibits more stable and discriminative ranking behavior than existing methods. Further analysis shows that LAREV reduces dependence on leaked baseline information while maintaining competitive predictive performance, validating its effectiveness as a reliable framework for rationale evaluation.

Keywords: Rationale Evaluation, Label Leakage, Leakage-Aware Evaluation, Information Theory, Conditional V-Information, Large Language Models

doi:10.6342/NTU202600527

Contents

Page

口試委員審定書 **i**

Acknowledgements **ii**

摘要 **iii**

Abstract **iv**

Contents **vi**

List of Figures **ix**

List of Tables **x**

Denotation **xii**

Chapter 1 Introduction **1**

1.1 Background 1

1.2 Motivation 3

1.3 Proposed Method 4

1.4 Outline of Thesis 5

Chapter 2 Related Work **7**

2.1 Preference-Based Evaluation and LLM-as-a-Judge 7

2.2 Rationale-Specific Automatic Metrics for Reasoning Chains 8

2.3 Behavior-Based Rationale Evaluation 9

2.4 Information-Theoretic Rationale Evaluation 9

Chapter 3 Problem Definition **11**

3.1 Baseline-Conditioned Evaluation via Conditional V-information . . . 11

3.2 Formal Setup (REV-Specific) 12

3.3 Failure Mode: Label Leakage in Baseline Inputs 13

3.4 Problem Statement 14

doi:10.6342/NTU202600527

Chapter 4 Methodology **15**

4.1 Overview of the LAREV Framework 15

4.2 Background and Preliminaries 16

4.2.1 REV 16

4.2.2 Integrated Gradients (IG) 17

4.2.3 Invariant Risk Minimization 18

4.3 Design intuition 19

4.4 Baseline Input Construction 19

4.5 Leakage-Aware Training of the Regular Model 21

4.5.1 Model Roles and Training Setup 21

4.5.2 IRM-Based Control of Baseline Leakage 22

4.5.3 Leakage Probe for Residual Baseline Leakage 25

4.5.4 Leakage Probe Objective 28

4.6 Combined Leakage-Aware Training Objective 29

4.6.1 Unified Training Objective 29

4.6.2 Complementary Roles of IRM and Leakage Probing 30

Chapter 5 Experiments **32**

5.1 Datasets 32

5.1.1 Open-Label Dataset: ECQA 32

5.1.2 Fixed-Label Dataset: e-SNLI 33

5.1.3 Dataset Splits and Statistics 33

5.2 Evaluation Design 34

5.2.1 Rationale Variants 35

5.3 Experimental Setup 36

5.3.1 Model Roles and Training Pipeline 37

5.3.2 IRM Environment Instantiation 38

5.3.3 Implementation Details and Hyperparameters 39

5.4 Results 40

doi:10.6342/NTU202600527

5.4.1 Rationale Evaluation on ECQA 40
5.4.2 Component-wise Analysis of Regularization Effects 43
5.4.3 Generalization of the Leakage-Aware Evaluator 44
5.4.4 Hyperparameter Sensitivity 46
5.4.5 Robustness across Task Models 47
5.4.6 Rationale Evaluation on e-SNLI 48
Chapter 6 Conclusion **51**
6.1 Summary of Contributions 51
6.2 Limitations 52
6.3 Future Work 53
References **54**
Appendix A — Experimental Pipeline **59**
Appendix B — Results with BART-large Backbone **61**
B.1 ECQA Results 61
B.2 e-SNLI Results 63
Appendix C — Additional Illustrative Examples **65**
C.1 Rationale Variants in e-SNLI 65
C.2 LLM Prompt for Antonym Environment Instantiation 66
C.3 Prompt for Rationale Generation from Task Models 66

doi:10.6342/NTU202600527

List of Figures

5.1 Visualization of rationale separation on ECQA using the T5-large backbone. Bars indicate score differences between the Gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales. 42

5.2 Effect of λ_{probe} on the validation loss when training the leakage-aware evaluator Φ_{LA} on ECQA using the T5-large backbone under different λ_{IRM} settings. The x-axis is shown in log scale. 46

5.3 Visualization of evaluator robustness across task models on ECQA using the T5-large backbone. Bars indicate score differences between the Gold rationale and degraded variants (Gold − Leaky, Gold − Gold-Leaky, and Gold − Vacuous). Each group corresponds to a degradation setting, and larger positive values indicate clearer separation between informative and less informative rationales. 48

5.4 Visualization of rationale separation on e-SNLI using the T5-large backbone. Bars indicate score differences between the Gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales. 49

B.1 Visualization of rationale separation on ECQA using the BART-large backbone. Bars indicate score differences between the gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales. 62

B.2 Visualization of rationale separation on e-SNLI using the BART-large backbone. Bars indicate score differences between the gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales. 64

doi:10.6342/NTU202600527

List of Tables

4.1 Representative baseline inputs and label-relevant tokens in ECQA and e-SNLI. The highlighted token in each baseline input corresponds to the leakage term t_{leak} identified via Integrated Gradients. 20

4.2 Illustration of IG-guided baseline environments used for IRM training. For each example, Integrated Gradients (IG) is first applied to identify a localized leakage term in the baseline input. Three environments are then constructed by preserving, masking, or semantically reversing the detected leakage term, while keeping the surrounding context unchanged. 24

4.3 Illustrative examples of distributed baseline leakage. For each baseline input, the IG-identified leakage term t_{leak} is shown in **bold**, while the remaining baseline tokens $b \setminus t_{\text{leak}}$ are highlighted with a light background. The leakage probe objective explicitly regulates the model's sensitivity to $b \setminus t_{\text{leak}}$, which may still encode residual label-relevant information. . . . 26

4.4 Examples of leakage-Masked baseline inputs $\tilde{b}$ used for leakage probe training. Leakage terms identified by Integrated Gradients are masked to remove direct label-relevant cues while preserving the surrounding baseline context. 27

5.1 Dataset statistics for ECQA and e-SNLI. 34

5.2 An illustrative ECQA test example showing the question, answer options, and different rationale variants r used in evaluation. All variants correspond to the same question–answer pair, with the gold answer being happiness. 36

5.3 Absolute rationale scores on ECQA using T5-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution. 41

5.4 Relative sensitivity of different evaluation methods on ECQA (T5-large), measured as score differences between the Gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales. 41

5.5 A representative ECQA example evaluated with T5-large. The upper block shows task information, while the lower block compares Gold and Leaky rationales and their resulting difference. 43

doi:10.6342/NTU202600527

5.6 Analysis on ECQA (T5-large), showing the effect of IRM and leakage probe regularization on relative score separations. Higher values indicate stronger discrimination between informative and degraded rationales. . . 44
5.7 Test-set accuracy (%) comparison between REV and our method on ECQA under different rationale variants. Despite introducing IRM and leakage probe regularization, our method maintains comparable or improved generalization performance. 45
5.8 Evaluator robustness across task models on ECQA. We fix the evaluator backbone and baseline input b, and vary only the rationale source. Columns report score differences between the Gold rationale and degraded variants (Gold − Leaky, Gold − Gold-Leaky, and Gold − Vacuous), with **SUM** denoting the aggregate separation. Higher values indicate clearer discrimination. 47
5.9 Absolute rationale scores on e-SNLI using T5-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution. 48
5.10 Relative sensitivity of different evaluation methods on e-SNLI (T5-large), measured as score differences between the Gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales. 49

B.1 Absolute rationale scores on ECQA using BART-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution. 61
B.2 Relative sensitivity of different evaluation methods on ECQA (BART-large), measured as score differences between the gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales. 61
B.3 Absolute rationale scores on e-SNLI using BART-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution. 63
B.4 Relative sensitivity of different evaluation methods on e-SNLI (BART-large), measured as score differences between the gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales. 63
C.1 An illustrative e-SNLI test example showing different rationale variants constructed for evaluation in a fixed-label setting. All variants correspond to the same premise–hypothesis pair, with the gold label being contradiction. 65

doi:10.6342/NTU202600527

Denotation

Denotation

x	Input instance (e.g., a question or input example).
y	Ground-truth label (target output).
r	Free-text rationale (gold rationale).
b	Baseline input (baseline rationale) used as the reference condition in baseline-conditioned evaluation.
$I_V(r \to y \mid b)$	Conditional V-information measuring the contribution of r to predicting y given baseline input b.
$H_V(y \mid \cdot)$	Predictive V-entropy of y under a constrained predictor family.
$\mathrm{REV}(r)$	Operational REV score computed by contrasting log-likelihoods under baseline-only vs. rationale-augmented inputs.
Φ_{base}	Baseline model that receives only baseline input b (reference evaluator model in REV).
Φ	Regular model that receives rationale-augmented input $r+b$ (standard evaluator model in REV).
Φ_{LA}	Leakage-aware evaluator trained with LAREV constraints, as proposed in this work.

doi:10.6342/NTU202600527

t_{leak}	Token or term in the baseline input identified as contributing to label leakage.
$\tilde{b}$	Modified baseline input obtained by masking or removing leakage-related terms.

doi:10.6342/NTU202600527

Chapter 1

Introduction

1.1 Background

Free-text rationales are widely used to provide human-interpretable explanations for model predictions in natural language processing tasks [1, 2], such as question answering and natural language inference. By accompanying a predicted label with a natural language justification, rationales aim to reveal the reasoning process of the model and improve transparency, trust, and debuggability. As a result, rationales have become a standard mechanism for communicating and justifying model decisions to human users.

At the same time, modern large language models are increasingly capable of generating fluent and human-like rationales to accompany their predictions [3, 4]. These rationales often appear coherent, plausible, and well aligned with human reasoning patterns, further strengthening their appeal as explanations of model behavior. Consequently, free-text rationales are widely adopted to support interpretability claims and to justify model decisions [5–8].

However, the apparent plausibility of a rationale does not constitute evidence that the model actually relies on it when making a prediction. This gap arises because a model's

doi:10.6342/NTU202600527

internal decision process is not directly observable: in general, it is impossible to inspect how different components of the input influence the final output [9, 10]. As a result, a generated rationale may appear reasonable to human readers while playing no functional role in the model's decision process [11].

This lack of direct observability poses a fundamental challenge for rationale evaluation. Human judgments of coherence or plausibility cannot determine whether a rationale provides additional task-relevant information beyond what is already available to the model, nor whether it is actually relied upon by the model during prediction. While human-centered evaluation protocols can provide reliable assessments, they are expensive, domain-specific, and difficult to scale [12].

In response to these limitations, prior work has explored automatic approaches to explanation evaluation. Many such methods aim to reduce reliance on human annotation by introducing proxy-based evaluation signals. However, a large portion of existing approaches focus on surface-level similarity between generated explanations and reference texts, and do not explicitly assess whether an explanation influences the model's predictive behavior or contributes label-relevant information [13].

These shortcomings have motivated the development of information-based rationale evaluation methods that assess explanations through their effect on model behavior rather than surface-level similarity. In particular, recent work seeks to quantify how much a rationale influences a model's predictive behavior, with the goal of estimating the additional label-relevant information contributed by the explanation.

One influential example of this line of work is Rationale Evaluation with Conditional V-information (REV) [14], which formalizes rationale quality as the additional usable in-

doi:10.6342/NTU202600527

formation provided by a rationale conditioned on a baseline input. In the REV framework, rationale quality is approximated by comparing the log probability assigned to the gold label when the model is given access to both the rationale and the baseline input, versus the baseline input alone. This formulation provides a principled, model-based approximation to conditional information gain.

1.2 Motivation

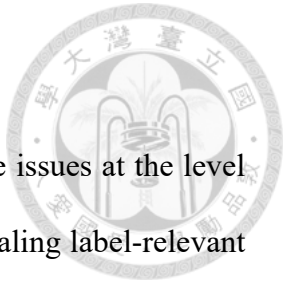
Despite the strong theoretical grounding of REV, baseline-conditioned rationale evaluation faces a critical practical challenge: the baseline input itself often contains substantial label-relevant information, a phenomenon commonly referred to as label leakage. Prior work has also observed that such leakage-prone explanations often take a repetitive or tautological form, echoing information that is already sufficient to determine the label [15].

However, in many real-world datasets, baseline input constructions—such as templated questions, answer choices, or partial contexts—can already provide strong cues for predicting the correct label. As a result, models may achieve high accuracy even when rationales are absent or uninformative.

In such cases, the baseline-conditioned rationale evaluation framework REV can yield unreliable estimates of rationale quality. A rationale may appear to improve model predictions simply because the baseline already reveals the answer, rather than because the rationale contributes meaningful reasoning. Consequently, the estimated additional contribution of the rationale in the rationale-plus-baseline setting can be systematically inflated, making it difficult to distinguish genuinely informative rationales from superfi-

doi:10.6342/NTU202600527

cial or vacuous ones.



Some prior work on rationale evaluation has considered leakage issues at the level of the rationale itself, aiming to prevent rationales from directly revealing label-relevant information [16, 17]. However, the REV framework is proposed as a baseline-conditioned evaluation paradigm and does not explicitly account for leakage arising from the baseline input itself. As a result, when label-relevant information is already present in the baseline, the resulting rationale scores may become unreliable, making it difficult to faithfully assess the additional contribution of the rationale. This limitation motivates the need for evaluation frameworks that explicitly consider baseline leakage while retaining the advantages of baseline-conditioned rationale evaluation.

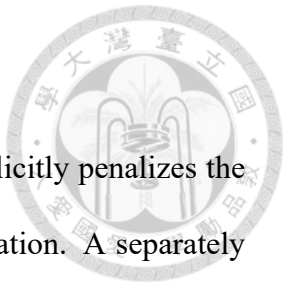
1.3 Proposed Method

To address the baseline leakage problem, we propose **LAREV (Leakage-Aware Rationale Evaluation with Conditional V-information)**, a training framework designed to improve the reliability of baseline-conditioned rationale evaluation. LAREV adopts the conditional V-information objective of REV, but modifies the training procedure of the evaluator model to reduce sensitivity to label-relevant signals present in the baseline input.

The framework introduces two complementary mechanisms. First, LAREV applies an invariance regularization based on Invariant Risk Minimization (IRM), encouraging the evaluator to make consistent predictions across multiple baseline environments derived from the same input. By enforcing invariance across these environments, the model is discouraged from relying on environment-specific shortcut cues and is instead driven to

doi:10.6342/NTU202600527

focus on information that generalizes across baseline variations.



Second, LAREV incorporates a leakage probing signal that explicitly penalizes the model for retaining predictive sensitivity to residual baseline information. A separately trained leakage probe is used to estimate how much label-relevant information remains accessible through a masked baseline input. During training, this signal serves as a regularization term that discourages the evaluator from encoding or exploiting such leakage.

Importantly, LAREV does not alter the evaluation metric of REV, nor does it require changes to the form or content of the rationales. Instead, it reshapes the evaluator model so that the resulting rationale scores more faithfully reflect the information contributed by the rationale, rather than being influenced by leakage induced by the baseline input.

1.4 Outline of Thesis

The remainder of this thesis is organized as follows.

Chapter 2 reviews prior work on rationale evaluation, with particular attention to evaluation frameworks that assess rationale usage through model behavior.

Chapter 3 formalizes the rationale evaluation problem under limited observability and introduces the baseline-conditioned evaluation framework based on conditional V-information. It further analyzes how label leakage in the baseline input can undermine the reliability of proxy-based evaluation signals and presents the resulting problem statement.

Chapter 4 presents the proposed LAREV framework in detail. This chapter describes the construction of baseline inputs, the attribution-guided identification of leakage terms, and the leakage-aware training procedure for the evaluator model. Two complementary

doi:10.6342/NTU202600527

mechanisms—IRM-based invariance regularization and leakage probing—are introduced to control different sources of baseline leakage, while preserving the original evaluation objective of REV.

Chapter 5 reports experimental evaluations on both an open-label dataset (ECQA) and a fixed-label dataset (e-SNLI). This chapter describes the experimental setup and evaluation design, and presents quantitative results and ablation studies that analyze the effectiveness and generalization behavior of the proposed method.

Finally, Chapter 6 concludes the thesis by summarizing the main contributions, discussing the limitations of the current approach, and outlining directions for future research on leakage-aware rationale evaluation.

doi:10.6342/NTU202600527

Chapter 2

Related Work

We review prior work on rationale evaluation, with an emphasis on how rationale quality is assessed for modern LLM-generated, free-form explanations. We first summarize how step-by-step rationales are used to support reasoning and interpretability. We then discuss three major families of rationale evaluation: (i) preference-based evaluation (often via LLM-as-a-judge), (ii) rationale-specific automatic metrics, and (iii) behavior- and information-theoretic evaluation frameworks.

2.1 Preference-Based Evaluation and LLM-as-a-Judge

A widely adopted approach to evaluating free-form model outputs is preference evaluation, where annotators (humans or models) compare two responses and select the preferred one. This paradigm has become especially prominent for LLM evaluation due to its simplicity and scalability, and is often operationalized via strong LLM judges [18].

Despite its practicality, binary preference signals are often coarse-grained for rationale evaluation: they provide limited diagnostic insight into which rationale attributes drive preferences and may conflate reasoning quality with style, verbosity, or presenta-

doi:10.6342/NTU202600527

tion. Recent work has therefore argued for decomposing rationale quality into fine-grained attributes and analyzing how such attributes explain preference outcomes [19]. This line of work highlights a growing need for evaluation signals that are both interpretable and sensitive to rationale-specific properties.

2.2 Rationale-Specific Automatic Metrics for Reasoning Chains

Another line of work develops automatic metrics tailored to step-by-step reasoning chains, aiming to move beyond generic text similarity measures. ROSCOE [20] proposes a suite of interpretable, reference-free scores designed for step-by-step rationales, targeting properties such as semantic consistency, logicality, and informativeness. ReCEval [21] evaluates reasoning chains through two key axes: step-wise correctness and informativeness, viewing reasoning chains as informal proofs that should both make valid inferences and contribute new information toward deriving the answer.

In parallel, work on training or tailoring rationalizers often relies on multiple rewards aligned with desirable rationale attributes (e.g., plausibility, consistency, diversity), and validates improvements via both automatic and human evaluation [22]. These efforts collectively reflect a shift toward more structured and property-aware evaluation of rationales.

Nevertheless, automatic rationale metrics can be sensitive to formatting assumptions (e.g., explicit step segmentation) and may not directly measure the marginal contribution of a rationale relative to information already present in the input. This motivates evaluation frameworks that explicitly ground rationale quality in changes in predictive behavior under

doi:10.6342/NTU202600527

controlled access to input.

2.3 Behavior-Based Rationale Evaluation

Behavior-based rationale evaluation assesses explanations by measuring how model predictions change when rationales are provided, removed, or perturbed. Representative approaches in this category include Rationale Quality (RQ) [23], which measures whether conditioning on a rationale improves a proxy model's ability to predict the gold label, and Leakage-Adjusted Simulatability (LAS) [16], which evaluates whether a simulator model can predict the task model's output given the input and a natural-language explanation while adjusting for trivial label leakage.

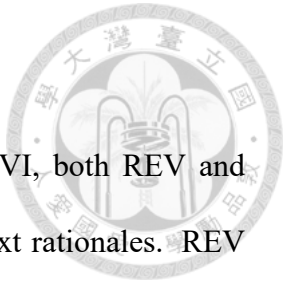
Although effective at detecting whether rationales influence model behavior, these behavior-based proxies do not explicitly quantify the amount of label-relevant usable information contributed by a rationale.

2.4 Information-Theoretic Rationale Evaluation

Information-theoretic approaches to rationale evaluation further formalize behavior-based evaluation by quantifying the amount of usable, label-relevant information contributed by a specific input under computational constraints. Pointwise V-information (PVI) [24] follows this line of work by measuring the usable information between an input and its label at the level of individual samples, and has been used to characterize dataset difficulty and sample-level hardness. While PVI provides a fine-grained, information-theoretic view of input–label relationships, it is not designed as a rationale evaluation metric and does not isolate the marginal contribution of free-text explanations beyond the

doi:10.6342/NTU202600527

original input.



Building on the notion of usable information introduced by PVI, both REV and RORA adapt information-theoretic ideas to the evaluation of free-text rationales. REV [14] is a representative behavior-grounded, information-theoretic approach that frames rationale quality in terms of conditional V-information, enabling a principled comparison between rationale-conditioned and baseline-only predictions.

RORA [17] addresses a key failure mode of rationale evaluation in which rationales achieve high scores by directly revealing label information. To mitigate this issue, RORA also adopts an information-theoretic perspective and regularizes the evaluator to be insensitive to identified leakage cues in the rationale.

Overall, while both RORA and REV are motivated by information-theoretic notions of usable information, they operationalize rationale quality through different comparative formulations. RORA contrasts predictions with and without access to the rationale, whereas REV measures the marginal contribution of a rationale conditioned on a baseline input, leading to different sensitivities to baseline information.

doi:10.6342/NTU202600527

Chapter 3

Problem Definition

This chapter formalizes the rationale evaluation problem under limited observability. We focus on baseline-conditioned rationale evaluation as instantiated by Rationale Evaluation with Conditional V information (REV) [14], and analyze the conditions under which its proxy-based estimates of rationale contribution become unreliable. In particular, we examine how label leakage in the baseline input confounds baseline-conditioned comparisons and motivates the need for leakage-aware evaluation models.

3.1 Baseline-Conditioned Evaluation via Conditional V-information

REV provides a principled formulation of rationale contribution based on predictive V-information. Specifically, rationale contribution is formalized using conditional V-information [25], which quantifies the reduction in predictive uncertainty about the label when a rationale is provided in addition to a baseline input. From an information-theoretic perspective, this formulation defines the evaluation target as the amount of usable information contributed by the rationale beyond the baseline [26].

Under this framework, the contribution of a rationale is expressed as the conditional

doi:10.6342/NTU202600527

V-information from r to y given a baseline input b:

$$I_V(r \to y \mid b) = H_V(y \mid b) - H_V(y \mid r + b), \tag{3.1}$$

where $H_V(\cdot)$ denotes predictive V-entropy under a constrained predictor family.

In practice, REV operationalizes this quantity by estimating the two terms using the log-likelihoods assigned to the ground-truth label by two models:

- a baseline model Φ_{base}, which receives only the baseline input b;
- a regular model Φ, which receives the rationale-augmented input $r + b$.

The resulting REV score is defined as:

$$\text{REV}(r) = [-\log p_{\Phi_{\text{base}}}(y \mid b)] - [-\log p_{\Phi}(y \mid r + b)]\,. \tag{3.2}$$

Under this formulation, larger REV values indicate greater reductions in predictive uncertainty when the rationale is provided, and are interpreted as more substantial evidence that the rationale contributes usable information for prediction.

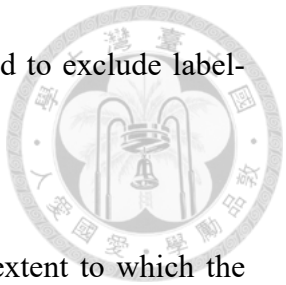
3.2 Formal Setup (REV-Specific)

To formalize baseline-conditioned rationale evaluation under the REV framework, we consider a supervised prediction setting where, for each instance, we are given:

- an input instance x,
- a ground-truth label y,
- a free-text rationale r,

doi:10.6342/NTU202600527

- and a baseline input b, constructed from x and y, and intended to exclude label-relevant information.



In the REV setting, rationale evaluation aims to estimate the extent to which the rationale r contributes task-relevant information for predicting y, beyond what is already available in the baseline input b. Since a model's internal decision process is not directly observable, such contributions cannot be measured directly and must instead be inferred through observable model behavior.

3.3 Failure Mode: Label Leakage in Baseline Inputs

The reliability of baseline-conditioned evaluation critically depends on the baseline input serving as a clean reference point. Ideally, the baseline b should exclude label-relevant information, such that any observed change in predictive behavior can be attributed to the rationale itself.

In practice, however, baseline inputs often encode substantial label-relevant information. When such label leakage occurs, baseline-conditioned comparisons can become confounded. Intuitively, when label-relevant information is already accessible through the baseline input, the evaluator becomes less sensitive to the presence of the rationale. Even if the rationale itself provides useful task-relevant information, its contribution may not induce a clear or measurable change in the model's predictive behavior under baseline-conditioned comparisons.

Consequently, proxy signals derived from baseline-conditioned performance differences can become unreliable. In such cases, REV scores may reflect the presence of label-relevant information in the baseline rather than the actual contribution of the rationale,

doi:10.6342/NTU202600527

leading to distorted estimates of rationale quality.

3.4 Problem Statement

The analysis above highlights a fundamental limitation of proxy-based rationale evaluation under baseline leakage. While REV provides a principled objective for measuring rationale contribution, its reliability is undermined when the baseline input is not a clean reference point.

This work, therefore, addresses the following problem:

> How can we obtain reliable estimates of rationale contribution within the REV framework when the baseline input itself contains label-leaking information that confounds proxy-based evaluation signals?

The goal is not to redefine the evaluation objective, but to design evaluation procedures that preserve the baseline-conditioned comparison central to REV while explicitly accounting for and mitigating the effects of label leakage in the baseline input.

doi:10.6342/NTU202600527

Chapter 4

Methodology

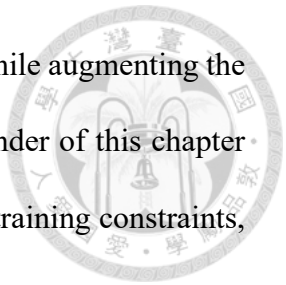
4.1 Overview of the LAREV Framework

We propose a methodology for evaluating the contribution of free-text rationales under baseline-conditioned evaluation settings, as instantiated by the REV framework. In such settings, a baseline input b serves as a reference condition, and rationale contribution is estimated by contrasting model behavior with and without access to the rationale.

We introduce **LAREV (Leakage-Aware Rationale Evaluation with Conditional V-information)**, a framework designed to improve the reliability of REV-style evaluation when the baseline input itself contains label-relevant information. If the baseline provides strong predictive signals that are independent of the rationale, the baseline-conditioned comparison may obscure whether the rationale contributes additional usable information.

Rather than redefining the evaluation objective underlying baseline-conditioned evaluation, LAREV modifies the training procedure of the predictive model used in the comparison. By regulating the extent to which the model can exploit predictive information from the baseline, LAREV aims to mitigate distortions caused by baseline leakage and to yield more faithful estimates of rationale contribution.

doi:10.6342/NTU202600527



At a high level, LAREV follows the REV evaluation structure while augmenting the pipeline with leakage-aware control over baseline usage. The remainder of this chapter presents the background concepts, baseline construction procedures, training constraints, and estimation steps that together constitute the LAREV framework.

4.2 Background and Preliminaries

We first introduce the methodological components underlying the proposed LAREV framework. The goal is to clarify how prior tools are instantiated and adapted to support leakage-aware rationale evaluation.

4.2.1 REV

REV, an information-based rationale evaluation method, seeks to assess whether a free-text rationale contributes to a model's prediction beyond what is already available in a baseline input. Rather than relying on the plausibility or fluency of a rationale, this line of work evaluates rationales based on their effects on model predictive behavior.

As introduced in Chapter 3, REV formalizes rationale contribution using conditional V-information [25], which measures the reduction in predictive uncertainty when a rationale is provided in addition to a baseline input. This formulation defines the evaluation target from an information-theoretic perspective [26].

In the REV framework, the conditional V-information is operationalized by comparing the log probability of two models under different input conditions. Specifically, given an input instance x, a ground-truth label y, a baseline input b constructed from x and y,

doi:10.6342/NTU202600527

and a free-text rationale r, the REV score of the rationale r is

$$\mathrm{REV}(r) = [-\log p_{\Phi_{\text{base}}}(y \mid b)] - [-\log p_{\Phi}(y \mid r + b)] . \tag{4.1}$$

where Φ_{base} denotes a baseline model that receives only the baseline input, and Φ denotes another model that receives the rationale-augmented input. This operational form makes explicit that REV relies on two distinct models and a comparison between baseline-only and rationale-augmented predictions.

LAREV builds upon this estimation framework. Rather than redefining the evaluation metric, LAREV improves the reliability of REV-based estimation by modifying the training procedure of the model Φ in the second term of the estimator, while leaving the baseline model Φ_{base} unchanged.

4.2.2 Integrated Gradients (IG)

Baseline inputs may contain tokens whose presence alone provides strong predictive signals for the target label. Identifying such tokens is essential for understanding how predictive information is distributed within the input, particularly when evaluating the contribution of free-text rationales.

Attribution methods provide a means of quantifying the influence of individual input tokens on model predictions. Among these methods, Integrated Gradients (IG) [27] offers a widely used approach for computing token-level attribution in differentiable models.

Given a predictive model f and an input representation x, IG attributes the model's prediction to each input dimension by integrating gradients along a straight-line path from

doi:10.6342/NTU202600527

a reference input x' to x:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha. \tag{4.2}$$

The resulting attribution scores indicate how strongly individual tokens contribute to the model's output relative to the reference input. Importantly, attribution scores produced by IG are not interpreted as evidence of causal model reasoning. Instead, IG is used as a diagnostic tool to identify input tokens that the model relies on most strongly when making predictions.

4.2.3 Invariant Risk Minimization

Invariant Risk Minimization (IRM) [28] is a learning principle proposed to encourage models to learn predictive relationships that remain stable across multiple environments. The central motivation of IRM is to reduce reliance on spurious correlations that vary across environments, thereby improving robustness to distributional shifts.

In IRM, an environment refers to a setting in which the prediction task remains fixed, while the correlations between input features and labels vary across environments. For example, in image classification, different backgrounds correlated with object labels can be treated as distinct environments, even though the object identity remains the target.

Formally, given a set of environments $\mathcal{E}$, IRM seeks a representation Φ such that there exists a single predictor w that is optimal across all environments, while jointly minimizing empirical risk:

$$\min_{\Phi, w} \sum_{e \in \mathcal{E}} \mathcal{L}^e(w \circ \Phi) \quad \text{s.t. } w \in \arg\min_{w'} \mathcal{L}^e(w' \circ \Phi), \quad \forall e \in \mathcal{E}. \tag{4.3}$$

doi:10.6342/NTU202600527

where $\mathcal{L}^e$ denotes the loss function evaluated on environment e.

IRM has been widely applied to tasks where predictive shortcuts often stem from environment-dependent signals. By encouraging invariance across environments, IRM provides a general mechanism for discouraging models from exploiting such shortcuts, without requiring explicit identification or removal of spurious features.

4.3 Design intuition

LAREV aims to encourage the evaluator to base its predictions primarily on information provided by the rationale, rather than on label-relevant cues that are already present in the baseline input. In doing so, LAREV reduces the tendency of the model to exploit baseline leakage and to ignore potentially informative rationale content during training. To achieve this, we introduce two complementary constraints that target different forms of baseline leakage. First, IRM discourages the evaluator from relying on localized shortcut cues in the baseline. Second, leakage probing complements IRM by discouraging the encoder from encoding residual label-relevant information that remains recoverable from the baseline. Together, these constraints bias predictive gains in the rationale-plus-baseline setting to be driven by the rationale content, while preserving the REV-style estimator.

4.4 Baseline Input Construction

Following the REV framework, the baseline input b represents task-relevant information that is available independently of the free-text rationale r. It serves as the reference condition in baseline-conditioned evaluation, against which the contribution of an additional rationale is assessed. Accordingly, baseline inputs encode essential task context

doi:10.6342/NTU202600527

while excluding explicit explanatory content.

Baseline construction is task-specific. For ECQA, a commonsense question answering task, the baseline is formed by declaratively combining the input question with the ground-truth answer, using a T5-based infilling model fine-tuned on (question, answer, declarative statement) tuples [29] following prior work [30] [1]. For e-SNLI, a natural language inference task, each (premise, hypothesis, label) tuple is first converted into a templated declarative sentence reflecting the target relation, and subsequently paraphrased using a pretrained T5-based model [2] to avoid fixed template patterns.

Although baseline inputs are designed to be vacuous with respect to reasoning, they may still contain substantial label-relevant information. Table 4.1 illustrates representative examples from the ECQA and e-SNLI datasets. This observation motivates the leakage-aware mechanisms introduced in the following sections.

Dataset	**Input Question**	**Baseline Input (b)**	**Label (y)**	**Label-Relevant Token (t_{leak})**
ECQA	The man didn't like getting out of bed and stepping on the cold tile, so where did he put carpeting?	The man didn't like getting out of bed and stepping on the cold tile, so he put carpeting in the **bedroom**.	bedroom	**bedroom**
e-SNLI	Bicyclists waiting at an intersection. The bicycles are on a road.	The presence of cyclists waiting at a crossroads **implies** that they are on a road.	entailment	**implies**

Table 4.1: Representative baseline inputs and label-relevant tokens in ECQA and e-SNLI. The highlighted token in each baseline input corresponds to the leakage term t_{leak} identified via Integrated Gradients.

[1] https://github.com/jifan-chen/QA-Verification-Via-NLI
[2] https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

doi:10.6342/NTU202600527

4.5 Leakage-Aware Training of the Regular Model

We describe a leakage-aware training procedure applied exclusively to the regular model Φ, while keeping the baseline model Φ_{base} used for comparison fixed. The resulting leakage-aware model is denoted as Φ_{LA}. Importantly, Φ_{LA} is trained as a separate model instance under a modified training objective, rather than obtained via fine-tuning from Φ. Rather than modifying the evaluation estimator itself, this procedure intervenes at training time to regulate how baseline information is utilized when the model predicts from rationale-augmented inputs.

Specifically, LAREV augments the training of Φ with additional constraints that discourage reliance on spurious or leaked cues present in the baseline input b. By regulating baseline usage during training, the resulting model Φ_{LA} yields more reliable baseline-conditioned behavior when predicting under inputs of the form $r + b$. The following subsections describe the mechanisms used to implement this leakage-aware training.

4.5.1 Model Roles and Training Setup

LAREV distinguishes between two predictive models with asymmetric roles. The baseline model Φ_{base} is trained using standard procedures following REV and receives only the baseline input b. It serves as a fixed reference in baseline-conditioned evaluation and is not subject to leakage-aware constraints.

In contrast, the regular model Φ receives the rationale-augmented input $r + b$ and is the only model whose training procedure is modified under LAREV. After applying leakage-aware constraints, this model is denoted as Φ_{LA}.

doi:10.6342/NTU202600527

By intervening solely on the training of Φ_{LA}, the framework preserves the baseline-conditioned comparison underlying REV, while explicitly limiting the extent to which predictions under $r+b$ can exploit signals originating from the baseline input. Under these model roles, the proposed LAREV score can be approximated by the following difference in negative log-likelihoods:

$$\text{LAREV}(r) = [-\log p_{\Phi_{\text{base}}}(y \mid b)] - [-\log p_{\Phi_{\text{LA}}}(y \mid r+b)]\,. \tag{4.4}$$

4.5.2 IRM-Based Control of Baseline Leakage

Baseline rationales often contain individual tokens that directly reveal label-relevant information, resulting in leakage terms, as illustrated in Table 4.1. When such tokens are present, a regular model Φ can minimize prediction loss by relying on these shortcuts rather than on information provided by the rationale.

To control this behavior, LAREV applies IRM using environments that are explicitly constructed to vary these shortcut signals. Specifically, Integrated Gradients (IG) is first used to identify leakage terms in the baseline input b. These terms are then manipulated to generate multiple baseline variants, which are treated as distinct environments in the IRM objective.

Leakage term detection via Integrated Gradients. To ensure that leakage terms reflect information available solely from the baseline, Integrated Gradients (IG) is computed with respect to the baseline-conditioned model Φ_{base}, using the baseline input b as the model input. Given this model and input, IG is used to compute token-level attribution scores with respect to the negative log-likelihood of the target label predicted by Φ_{base}. Tokens that

doi:10.6342/NTU202600527

consistently exhibit large attribution magnitudes are interpreted as localized baseline cues that the model relies on disproportionately. These tokens are designated as leakage terms, denoted as t_{leak}, and form the basis for subsequent environment construction. Importantly, IG is not used for post-hoc explanation in this setting, but as an operational tool for identifying candidate shortcut features in the baseline input. Examples of label-relevant tokens identified as leakage terms by IG are provided in Table 4.1.

IG-guided construction of IRM environments. Once a leakage term has been identified, LAREV constructs multiple variants of the baseline input by systematically manipulating this term while keeping the surrounding context unchanged. Each variant corresponds to a distinct environment in the IRM formulation. Specifically, three environments are constructed: (i) an original baseline environment in which the leakage term is preserved, (ii) a mask environment in which the leakage term is replaced by a mask symbol, and (iii) an antonym environment in which the semantic polarity of the leakage term is reversed. All three environments are defined at the level of controlled interventions on the identified leakage term, and are treated symmetrically in the IRM objective. This design ensures that correlations between the leakage term and the target label vary across environments, while other aspects of the baseline remain largely invariant. Table 4.2 provides illustrative examples of the resulting environments for ECQA and e-SNLI.

IRM objective under leakage-aware environments. The IG-guided baseline variants are treated as distinct environments in the IRM objective. The IRM constraint encourages the regular model Φ_{LA} to produce consistent predictions across these environments, thereby reducing reliance on leakage tokens. Importantly, the model is not forced to ignore the baseline input altogether; baseline information that remains consistently predic-

doi:10.6342/NTU202600527

Dataset	Environment	Baseline Variant
ECQA	Baseline	The man didn't like getting out of bed and stepping on the cold tile, so he put carpeting in the bedroom.
	Mask	The man didn't like getting out of bed and stepping on the cold tile, so he put carpeting in the <mask>.
	Antonym	The man didn't like getting out of bed and stepping on the cold tile, so he put carpeting in the attic.
e-SNLI	Baseline	The presence of cyclists waiting at a crossroads implies that they are on a road.
	Mask	The presence of cyclists waiting at a crossroads <mask> that they are on a road.
	Antonym	The presence of cyclists waiting at a crossroads denies that they are on a road.

Table 4.2: Illustration of IG-guided baseline environments used for IRM training. For each example, Integrated Gradients (IG) is first applied to identify a localized leakage term in the baseline input. Three environments are then constructed by preserving, masking, or semantically reversing the detected leakage term, while keeping the surrounding context unchanged.

tive across environments is retained.

Concretely, in our setting with three environments (baseline, mask, and antonym), a leakage term is considered environment-specific if relying on it yields low loss in only a subset of environments, but leads to substantially higher loss in others. IRM discourages such environment-dependent dependencies, while favoring predictive behavior that achieves consistently low risk across all environments.

IRM loss formulation. Formally, let $\mathcal{E}$ denote the set of environments constructed from IG-guided baseline variants. For each environment $e \in \mathcal{E}$, we define the empirical risk of the regular model Φ_{LA} as

$$\mathcal{L}^e(\Phi_{\text{LA}}) = \mathbb{E}_{(r,b,y)\sim e}\left[-\log p_{\Phi_{\text{LA}}}(y \mid r + b)\right], \tag{4.5}$$

where the expectation is taken over rationale–baseline pairs (r, b) instantiated under environment e.

doi:10.6342/NTU202600527

Following the IRMv1 formulation, invariance is enforced by encouraging the optimal prediction behavior to be shared across all environments. In our setting, this is implemented using the standard IRMv1 first-order surrogate, which introduces a scalar scaling variable applied to the model outputs. Specifically, the IRM penalty is defined as

$$\mathcal{L}_{\text{IRM}}(\Phi_{\text{LA}}) = \sum_{e \in \mathcal{E}} \|\nabla_w \mathcal{L}^e(w \circ \Phi_{\text{LA}})\|_2^2 \Big|_{w=1}. \tag{4.6}$$

where w is a scalar multiplier applied to the model logits. Evaluating the gradient at $w = 1$ corresponds to the IRMv1 scale surrogate and encourages invariant predictive behavior across environments, thereby discouraging reliance on environment-specific shortcut features.

The overall training objective for IRM-based leakage control is then given by

$$\mathcal{L}_{\text{IRM}}^{\text{total}} = \mathbb{E}_{e \in \mathcal{E}} \left[\mathcal{L}^e(\Phi_{\text{LA}})\right] + \lambda_{\text{IRM}} \cdot \mathbb{E}_{e \in \mathcal{E}} \left[\mathcal{L}_{\text{IRM}}^e(\Phi_{\text{LA}})\right], \tag{4.7}$$

where λ_{IRM} is a tunable regularization weight controlling the strength of the invariance constraint.

4.5.3 Leakage Probe for Residual Baseline Leakage

While IRM controls reliance on identified leakage terms, it does not explicitly constrain how predictive information is encoded in the remaining baseline tokens. To complement IRM, LAREV introduces an auxiliary leakage probe objective that regulates the model's sensitivity to the baseline tokens $b \setminus t_{\text{leak}}$, where t_{leak} is the IG-identified leakage term.

doi:10.6342/NTU202600527

Motivation and intuition. Table 4.3 illustrates the motivation for introducing the leakage probe. For each baseline input, a localized leakage term t_{leak} is first identified using Integrated Gradients (e.g., trying or implies). Even when this explicit leakage term is controlled, the surrounding baseline context $b \setminus t_{\text{leak}}$ may still encode label-relevant information. For example, in ECQA, even when the explicit answer token (e.g., **bedroom**) is masked, semantically related context such as "bed" may still provide residual label-relevant signals, motivating the need to regulate model sensitivity to $b \setminus t_{\text{leak}}$. The leakage probe, therefore, regulates the model's sensitivity to $b \setminus t_{\text{leak}}$, encouraging predictive gains under the rationale-augmented input $r + b$ to be driven by the rationale r rather than by residual baseline cues.

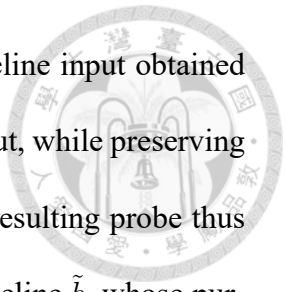
Dataset	**Baseline Input (b)**
ECQA	The man didn't like getting out of bed and stepping on the cold tile, so he put carpeting in the **bedroom** .
e-SNLI	The presence of cyclists waiting at a crossroads **implies** that they are on a road.

Table 4.3: Illustrative examples of distributed baseline leakage. For each baseline input, the IG-identified leakage term t_{leak} is shown in **bold**, while the remaining baseline tokens $b \setminus t_{\text{leak}}$ are highlighted with a light background. The leakage probe objective explicitly regulates the model's sensitivity to $b \setminus t_{\text{leak}}$, which may still encode residual label-relevant information.

Leakage probe model construction. The leakage probe ψ is not implemented as a separate network, but is derived directly from the regular evaluator model Φ in the REV framework. Specifically, Φ corresponds to the standard regular model trained on the original task without any leakage-aware constraints (e.g., IRM or leakage probing), and is used as the initialization for constructing ψ under a constrained training setup.

Under this construction, ψ shares the same architecture and initialization as Φ, but is trained to operate solely on baseline inputs $\tilde{b}$, without access to rationale information, with

doi:10.6342/NTU202600527



restricted parameter updates. Here, $\tilde{b}$ denotes a leakage-masked baseline input obtained by masking the IG-identified leakage term in the original baseline input, while preserving the remaining context. Examples of $\tilde{b}$ are shown in Table 4.4. The resulting probe thus functions as a probing model conditioned on the leakage-masked baseline $\tilde{b}$, whose purpose is to quantify the predictive strength of baseline information under a controlled input $\tilde{b}$ and to provide a regularization signal for leakage-aware training.

Dataset	**Leakage-Masked Baseline ($\tilde{b}$)**
ECQA	The man didn't like getting out of bed and stepping on the cold tile, so he put carpeting in the `<mask>`.
e-SNLI	The presence of cyclists waiting at a crossroads `<mask>` that they are on a road.

Table 4.4: Examples of leakage-Masked baseline inputs $\tilde{b}$ used for leakage probe training. Leakage terms identified by Integrated Gradients are masked to remove direct label-relevant cues while preserving the surrounding baseline context.

Frozen encoder and decoder-only adaptation. The objective of the leakage probe is not to suppress baseline leakage, but to measure and expose the predictive strength of the leakage-masked baseline input $\tilde{b}$ in order to identify residual label leakage encoded in the baseline.

To this end, the encoder of Φ is kept fixed during leakage probe training, while only the decoder is optimized. This design ensures that baseline-conditioned representations remain unchanged throughout probing, and that any reduction in probe loss arises solely from how these fixed representations are utilized at the decision stage.

If the encoder were allowed to update, the probe could reduce loss by reshaping the representation space itself, for example, by amplifying or reorganizing label-relevant baseline cues. Such representational adaptation would confound interpretation, as probe performance would no longer reflect the predictive strength of baseline information learned

doi:10.6342/NTU202600527

by the pretrained model.

4.5.4 Leakage Probe Objective

Having specified the construction and training constraints of the leakage probe ψ, we now define its learning objective and clarify what information this objective is designed to measure.

Leakage measurement via probing. The leakage probe ψ is trained to predict the target label y using only the leakage-masked baseline input $\tilde{b}$. Its learning objective is defined as the negative log-likelihood of the target label, as defined in Eq. (4.8).

$$\mathcal{L}_{\text{probe}}(\psi) = \mathbb{E}_{(\tilde{b},y)}\left[-\log p_\psi\left(y \mid h_{\tilde{b}}^{\Phi}\right)\right], \qquad h_{\tilde{b}}^{\Phi} = \text{Enc}_\Phi(\tilde{b}). \tag{4.8}$$

Minimizing $\mathcal{L}_{\text{probe}}$ encourages the probe to extract as much label-relevant signal as possible from the baseline input, subject to the constraints imposed during probe training, such as frozen encoder representations.

Under these constraints, the probe loss serves as a quantitative measure of how much label-relevant information remains decodable from the leakage-masked baseline $\tilde{b}$ under a fixed pretrained representation space. A lower value of $\mathcal{L}_{\text{probe}}$ indicates that baseline information alone remains highly predictive through $h_{\tilde{b}}^{\Phi}$, whereas a higher probe loss suggests weaker decodable label support from $\tilde{b}$.

Importantly, the leakage probe is not designed to suppress leakage by itself; instead, it provides a measurement signal that can be reused as a regularization objective in subsequent stages. In Section 4.6, we describe how the trained probe ψ is kept fixed and used to regularize the encoder representations learned by the leakage-aware evaluator Φ_{LA}.

doi:10.6342/NTU202600527

4.6 Combined Leakage-Aware Training Objective

Having introduced IRM-based invariance and leakage probing as complementary mechanisms, we now describe how these components are integrated into a unified leakage-aware training objective for the regular model Φ_{LA}.

4.6.1 Unified Training Objective

Let $\mathcal{L}_{\text{ERM}}(\Phi_{\text{LA}})$ denote the standard empirical risk minimization objective of the evaluator model on rationale-augmented inputs $(r + b)$. To control baseline leakage, LAREV augments this objective with two complementary regularization signals: an IRM-based invariance penalty and a leakage probing signal.

Formally, the overall leakage-aware training objective is given by

$$\mathcal{L}_{\text{LA}} = \mathcal{L}_{\text{ERM}}(\Phi_{\text{LA}}) + \lambda_{\text{IRM}} \cdot \mathcal{L}_{\text{IRM}}(\Phi_{\text{LA}}) - \lambda_{\text{probe}} \cdot \mathcal{L}_{\text{probe}}(\psi), \tag{4.9}$$

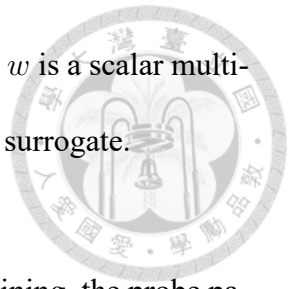
where $\mathcal{L}_{\text{IRM}}$ denotes the IRMv1 invariance penalty computed across IG-guided baseline environments, and $\mathcal{L}_{\text{probe}}$ is the leakage probe loss. The coefficients λ_{IRM} and λ_{probe} control the strength of the IRM regularization and the leakage probing signal, respectively.

Expanding each component, the unified objective can be written as

$$\begin{aligned}\mathcal{L}_{\text{LA}}(\Phi_{\text{LA}}, \psi) = {} & \mathbb{E}_{(r,b,y)} \left[-\log p_{\Phi_{\text{LA}}}\big(y \mid r + b\big) \right] \\ & + \lambda_{\text{IRM}} \sum_{e \in \mathcal{E}} \left\| \nabla_w \, \mathbb{E}_{(r,b,y) \in e} \left[-\log p_{w \circ \Phi_{\text{LA}}}\big(y \mid r + b\big) \right] \right\|_2^2 \\ & - \lambda_{\text{probe}} \, \mathbb{E}_{(\tilde{b},y)} \left[-\log p_{\psi}\big(y \mid h_{\tilde{b}}^{\Phi_{\text{LA}}}\big) \right], \qquad h_{\tilde{b}}^{\Phi_{\text{LA}}} = \text{Enc}_{\Phi_{\text{LA}}}(\tilde{b}). \end{aligned} \tag{4.10}$$

doi:10.6342/NTU202600527

Here, $\mathcal{E}$ denotes the set of IG-guided baseline environments, and w is a scalar multiplier applied to the model logits, following the standard IRMv1 scale surrogate.



Interpretation of the probe regularizer. During leakage-aware training, the probe parameters ψ are kept fixed. The probe term is subtracted so that optimizing Φ_{LA} maximizes the probe loss, i.e., it encourages the encoder representations $h_{\tilde{b}}^{\Phi_{\text{LA}}}$ to contain as little decodable label information as possible under the frozen probe. Intuitively, if ψ can easily predict y from $h_{\tilde{b}}^{\Phi_{\text{LA}}}$, then $\mathcal{L}_{\text{probe}}$ is small and the subtraction provides little regularization benefit. Conversely, when the baseline-induced representations become less predictive for ψ, $\mathcal{L}_{\text{probe}}$ increases, and the negative probe term more strongly reinforces the preference that predictive gains under $r + b$ should be driven by rationale information rather than residual baseline cues.

Operationally, gradients from the probe loss are back-propagated through $h_{\tilde{b}}^{\Phi_{\text{LA}}}$ into the encoder of Φ_{LA}, while ψ remains frozen.

4.6.2 Complementary Roles of IRM and Leakage Probing

IRM and leakage probing address complementary sources of baseline leakage by targeting different components of the baseline input. IRM focuses on leakage arising from explicitly identified leakage terms t_{leak}. By constructing multiple baseline variants that differ in the realization or presence of these terms, IRM encourages Φ_{LA} to learn predictions that are invariant to t_{leak}. As a result, IRM reduces reliance on shortcuts that are directly attributable to specific label-revealing tokens.

Leakage probing, in contrast, targets the residual predictive signal in the baseline after such leakage terms are controlled for. By operating on the leakage-masked baseline

doi:10.6342/NTU202600527

input $\tilde{b}$, the probe measures how much label-relevant information remains in the baseline beyond the identified leakage terms. This signal discourages Φ_{LA} from exploiting baseline information that remains predictive even when explicit leakage tokens are removed.

Together, IRM and leakage probing regulate complementary aspects of baseline leakage. IRM suppresses reliance on explicit leakage terms t_{leak}, while leakage probing limits dependence on residual information in $b \setminus t_{\text{leak}}$. Empirically, ablation results with either $\lambda_{\text{IRM}} = 0$ or $\lambda_{\text{probe}} = 0$ indicate that combining both mechanisms is more effective than either alone; detailed results are reported in Chapter 5 (Table 5.6).

doi:10.6342/NTU202600527

Chapter 5

Experiments

5.1 Datasets

We describe the datasets used in our experiments. We focus on datasets that provide free-text human rationales alongside task labels, as such datasets are particularly suitable for studying the contribution and potential leakage of rationales under baseline-conditioned evaluation.

5.1.1 Open-Label Dataset: ECQA

ECQA (Explainable Commonsense Question Answering) [31] is a commonsense reasoning dataset built upon the CommonsenseQA (CQA) benchmark [32], augmented with human-written explanations. Each instance consists of a question, a set of candidate answers, the correct answer label, and a natural language rationale explaining why the answer is correct.

We categorize ECQA as an open-label dataset. Although the task is formulated as multiple-choice question answering, the answer space is effectively open-ended at the token level. In our experimental setting, we do not provide the candidate answer choices to

doi:10.6342/NTU202600527

the model. Instead, the model is required to directly generate the predicted answer as free-form text, rather than selecting from fixed categorical indices. As a result, answer identity is represented by surface-level tokens, which makes the task particularly susceptible to answer leakage through rationales.

5.1.2 Fixed-Label Dataset: e-SNLI

e-SNLI [12] is an extension of the SNLI [33] natural language inference dataset, augmented with human-written explanations for each inference decision. Each instance consists of a premise, a hypothesis, a discrete inference label (entailment, contradiction, or neutral), and a free-text rationale that justifies the label.

We categorize e-SNLI as a fixed-label dataset, as the output space is restricted to a small, closed set of categorical labels. In this setting, the prediction task is formulated as selecting among predefined inference categories, rather than generating free-form answer tokens.

This structural constraint leads to a different role for rationales. Instead of encoding answer identity at the token level, e-SNLI rationales are intended to support or justify a categorical inference decision by articulating the underlying reasoning. At the same time, the presence of lexical or stylistic cues in the input may still introduce forms of label leakage, even when the label space itself is fixed.

5.1.3 Dataset Splits and Statistics

For ECQA, we strictly follow the original data splits provided by the dataset authors. All samples are assigned to training, validation, and test sets according to the original split

doi:10.6342/NTU202600527

identifiers.

For e-SNLI, the whole dataset contains a substantially larger number of instances, with the original training split comprising approximately 550,000 samples. To ensure computational feasibility, we construct reduced splits by subsampling from the original dataset. Specifically, we select the first 10,000 training instances, the first 2,000 validation instances, and the first 2,000 test instances from the corresponding original splits.

Across both datasets, the test splits are used exclusively for evaluation, and all reported results are computed on the held-out test sets. Training and validation splits are used to fit the evaluator models and to tune hyperparameters.

Table 5.1 summarizes the resulting dataset statistics.

Dataset	#Train	#Validation	#Test	Task Type
ECQA	7,598	1,090	2,194	QA
e-SNLI	10,000	2,000	2,000	NLI

Table 5.1: Dataset statistics for ECQA and e-SNLI.

5.2 Evaluation Design

We describe the evaluation design used to assess the robustness of the leakage-aware evaluator Φ_{LA}. We focus on how Φ_{LA} responds to different forms of rationale inputs under a fixed baseline condition, in order to examine whether its behavior remains stable when rationales vary in quality, relevance, and leakage.

doi:10.6342/NTU202600527

5.2.1 Rationale Variants

To systematically analyze the effect of rationales on the behavior of the final trained evaluator model Φ_{LA} on the test split, we construct multiple rationale variants for each original instance, following the evaluation practice in RORA [17]. We use r to denote the Gold rationale provided by the dataset, and $\tilde{r}$ to denote a generic rationale variant derived from r. Each variant is paired with the same baseline input b and serves as an evaluation input to observe how the model Φ_{LA} responds to different informational properties of rationales.

These rationale variants are designed to isolate specific characteristics, such as relevance, redundancy, and explicit answer leakage, while holding the baseline condition fixed. In particular, we distinguish the Gold rationale r provided by the dataset from several modified or degraded variants $\tilde{r}$ derived from it.

Specifically, we consider the following types of rationales:

- **Gold rationale**: the original human-written explanation provided in the dataset, which corresponds to the rationale r in our work.
- **Leaky rationale**: a rationale that explicitly reveals the correct answer or contains answer-identifying cues.
- **Gold-Leaky rationale**: a combination of the Gold rationale r with additional answer-revealing content.
- **Vacuous rationale**: a minimally informative or generic rationale that provides little task-relevant information. This variant corresponds to the baseline input b used in baseline-conditioned evaluation.

doi:10.6342/NTU202600527

Each rationale variant $\tilde{r}$ is concatenated with the baseline input b and provided as input to the trained evaluator model Φ_{LA}. In parallel, the baseline input b alone is provided to the baseline model Φ_{base}. By comparing model behavior under these two conditions, we obtain a LAREV score as defined in Eq. (4.1), which reflects how Φ_{LA} responds to a given rationale variant relative to the baseline condition.

Table 5.2 provides an illustrative ECQA test example, demonstrating how different rationale variants are constructed from the same underlying instance for evaluation. An additional illustrative example for the e-SNLI dataset is provided in Appendix C.1.

Field	**Content (ECQA Example)**
Question	What is likely the mood of those going to a party?
Answer Options	(op1) stress relief; (op2) have fun; (op3) happiness; (op4) babies; (op5) laughter
Label	happiness
Gold rationale	A party is a social gathering of invited guests involving eating, drinking, and entertainment which generally gives happiness. All the other options are not moods.
Leaky rationale	The answer is happiness.
Gold-Leaky rationale	A party is a social gathering of invited guests involving eating, drinking, and entertainment which generally gives happiness. All the other options are not moods. The answer is happiness.
Vacuous rationale	The mood of those going to a party is likely to be happiness.

Table 5.2: An illustrative ECQA test example showing the question, answer options, and different rationale variants r used in evaluation. All variants correspond to the same question–answer pair, with the gold answer being happiness.

5.3 Experimental Setup

We describe the experimental design used to evaluate the proposed leakage-aware rationale evaluation framework. We detail the roles of the different models involved, outline the training pipeline, and explain the evaluation procedure used to assess the robustness

doi:10.6342/NTU202600527

of rationales. For clarity, an overview of the end-to-end experimental pipeline is provided in Appendix A.1.

5.3.1 Model Roles and Training Pipeline

Our experimental pipeline involves multiple models with distinct roles, including a baseline model Φ_{base}, a regular rationale-augmented model Φ, and a final leakage-aware evaluator Φ_{LA}.

Baseline model Φ_{base}. The baseline model Φ_{base} is trained using only the baseline input b, which is constructed to exclude substantive explanatory information. This model captures task-relevant behavior under the baseline condition and serves as the reference model in baseline-conditioned evaluation. In addition, Φ_{base} is used as the attribution target for integrated gradients, which are computed under the baseline-only input to identify influential tokens associated with potential spurious correlations.

Regular model Φ. The regular model Φ is trained on inputs formed by concatenating the Gold rationale r with the baseline input b, i.e., $r + b$. Its role is to establish a fixed encoder representation that serves as a reference for assessing residual label predictability from the baseline input.

After training Φ, its encoder is frozen and used to produce representations of the masked baseline input $\tilde{b}$. A decoder-based leakage probe ψ is then trained on top of these frozen representations to predict task labels, thereby quantifying how much label-relevant information remains accessible through $\tilde{b}$.

Importantly, Φ and ψ are used exclusively to construct the leakage probing signal

doi:10.6342/NTU202600527

and are not used to initialize, fine-tune, or otherwise optimize the leakage-aware evaluator Φ_{LA}.

Leakage-aware evaluator Φ_{LA}. The final evaluator model Φ_{LA} is trained using leakage-aware objectives that incorporate multiple training environments and probe-based signals. The role of Φ_{LA} is to assess the contribution of rationales while reducing reliance on spurious or leaked cues present in the baseline input.

Training and evaluation protocol. The baseline model Φ_{base} and the regular model Φ are trained on the training split using inputs b and $r+b$, respectively. The validation split is used for hyperparameter tuning and model selection. All evaluation results are computed exclusively on the held-out test split, and no test instances are used during training or model selection.

5.3.2 IRM Environment Instantiation

Following the IG-guided environment construction described in Section 4.5.2, we instantiate three environments for IRM training: baseline (E1), mask (E2), and antonym (E3) environments.

The original and masked environments are instantiated by directly preserving or masking the identified leakage term in the baseline input, respectively. For the antonym environment, we apply a localized rewrite to the leakage term t_{leak} while keeping the surrounding baseline context unchanged. In practice, the antonym environment is instantiated using a pretrained large language model (`gemini-2.5-flash-lite`) as an automated text rewriting engine. The full prompting details are deferred to Appendix C.2.

doi:10.6342/NTU202600527

5.3.3 Implementation Details and Hyperparameters

Backbone models. All models in the pipeline—including the baseline model Φ_{base}, the regular model Φ, the leakage probe ψ, and the leakage-aware evaluator Φ_{LA}—are instantiated from the same backbone architecture. We conduct experiments using two sequence-to-sequence backbones: **T5-large** and **BART-large**.

Optimization. All models are trained using the AdamW optimizer. We use a fixed learning rate of 3e-5 across all training stages.

Batch size. Batch sizes are selected based on the role and training objective of each model. For training the baseline model Φ_{base} and the regular model Φ, we use a batch size of 8. For training the leakage probe ψ, we use a larger batch size of 16.

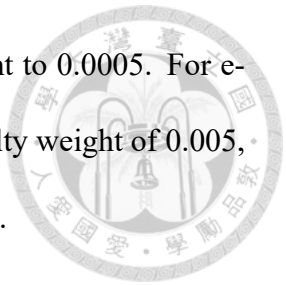
For training the leakage-aware evaluator Φ_{LA}, we use a batch size of 3. This choice reflects the IRM training setup, where each training batch corresponds to a single instance observed under three distinct environments. Using a batch size of 3 ensures that gradients are computed jointly across all environments associated with the same underlying example, as required by the IRM objective.

Training epochs. The baseline model Φ_{base}, the regular model Φ, and the leakage probe ψ are each trained for 8 epochs. The leakage-aware evaluator Φ_{LA} is trained for fewer epochs (2 epochs).

IRM and leakage penalty weights. Penalty weights for IRM and leakage probing are selected separately for each dataset when using the T5-large backbone. For ECQA, we

doi:10.6342/NTU202600527

set the IRM penalty weight to 5 and the leakage probe penalty weight to 0.0005. For e-SNLI, we use a stronger IRM penalty of 25 and a leakage probe penalty weight of 0.005, reflecting the more structured and label-constrained nature of the task.



Annealing schedule. Both the IRM penalty and the leakage probe penalty follow a linear warm-up schedule. Specifically, each penalty weight is linearly increased from zero to its full value during the first one-third of the total training steps, and remains fixed thereafter. This annealing strategy stabilizes early training by allowing the model to first learn task-relevant behavior before being strongly constrained by invariance and leakage-aware objectives.

5.4 Results

5.4.1 Rationale Evaluation on ECQA

We begin by comparing LAREV with prior information-based rationale evaluation methods, including REV[14] and RORA[17] [1], which assess rationales by measuring changes in model behavior under different input conditions. All methods considered here operate under baseline-conditioned or perturbation-based evaluation settings, and their scores should therefore be interpreted as proxies for usable information from rationale r.

Following the evaluation design described in Section 5.2.1, Table 5.3 reports the absolute evaluation scores on ECQA using the T5-large backbone. Across all rationale variants, LAREV (Ours) consistently yields lower absolute scores than REV and RORA. These lower scores should not be interpreted as inferior performance. Instead, they re-

[1] Although the RORA paper reports results on ECQA, we were unable to reproduce the reported numbers using the released code. We therefore implemented RORA following the methodology described in the paper and applied this implementation consistently across all dataset–backbone combinations in this work.

doi:10.6342/NTU202600527

flect a more stringent information-based evaluation criterion of LAREV that explicitly discounts predictive gains attributable to baseline leakage. As a result, LAREV is less likely to overestimate the contribution of rationales when baseline inputs already encode label-relevant signals. For this reason, absolute score magnitudes are not the primary indicator of evaluation quality in this setting; instead, we assess method effectiveness by examining whether score differences between the Gold rationale r and other rationale variants are meaningfully separated.

Methods $\downarrow$	**Rationale Variants** $\rightarrow$			
	Gold	**Gold-Leaky**	**Vacuous**	**Leaky**
Ours	-5.0796	-5.6513	-6.8136	-6.4818
RORA	-3.9256	-3.9449	-4.7805	-3.3221
REV	-0.2675	-0.4077	-0.6049	-0.3443

Table 5.3: Absolute rationale scores on ECQA using T5-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution.

Methods $\downarrow$	**Gold** $-$ **Degraded Rationales** $\rightarrow$			**SUM**
	Gold $-$ **Leaky**	**Gold** $-$ **Gold-Leaky**	**Gold** $-$ **Vacuous**	
Ours	**1.4022**	**0.5717**	**1.7340**	**3.7079**
RORA	-0.6035	0.0193	0.8549	0.2707
REV	0.0768	0.1402	0.3374	0.5544

Table 5.4: Relative sensitivity of different evaluation methods on ECQA (T5-large), measured as score differences between the Gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales.

Table 5.4 reports relative score differences between the Gold rationale r and various rationale variants. These differences directly reflect the extent to which a method distinguishes informative rationales from Gold-Leaky, Leaky, or Vacuous ones. For ease of comparison, we additionally report **SUM**, defined as the sum of relative score differences across all degraded rationale variants, which serves as an aggregate measure of overall discriminative strength.

doi:10.6342/NTU202600527

LAREV exhibits consistently larger positive separations between the Gold rationale r and all other rationale variants. In particular, the separation between Gold and Leaky rationales is substantially larger than that observed for RORA and REV. LAREV also produces larger separations for Vacuous rationales, indicating that its evaluation scores are more sensitive to changes in rationale content under the baseline-conditioned setting. Figure 5.1 provides a visual summary of the results in Table 5.4.

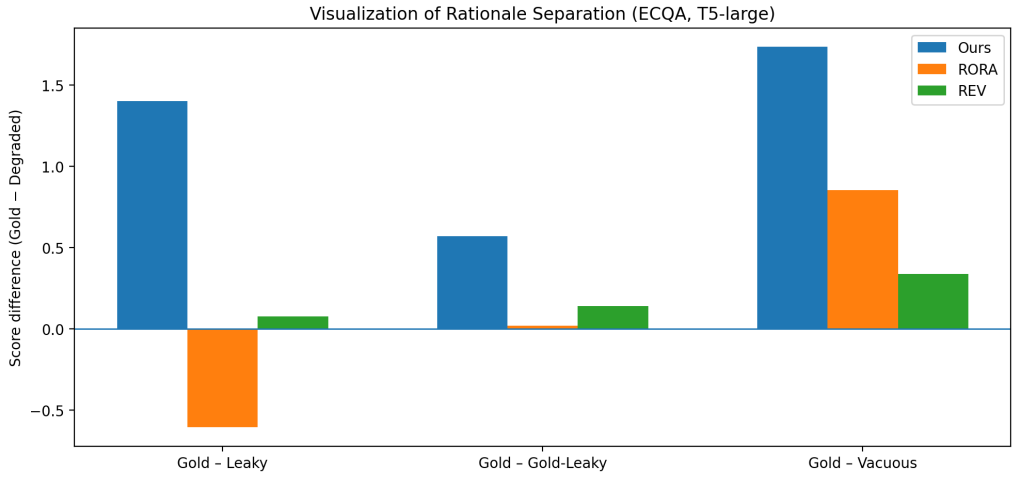


Figure 5.1: Visualization of rationale separation on ECQA using the T5-large backbone. Bars indicate score differences between the Gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales.

Table 5.5 presents a representative ECQA example illustrating how different evaluation methods behave under baseline leakage at the instance level. The Gold rationale provides a coherent explanation aligned with the correct choice, whereas the Leaky rationale explicitly states the answer. Although all methods assign higher scores to the Gold rationale, the resulting separation between Gold and Leaky rationales differs substantially across methods.

LAREV yields a markedly larger separation between Gold and Leaky rationales than both RORA and REV. By contrast, REV produces only a slight difference between the

doi:10.6342/NTU202600527

two rationales, indicating a limited ability to distinguish them in this example.

For completeness, we report the absolute and relative evaluation results on ECQA using the BART-large backbone in Appendix B.1, which show qualitatively similar patterns to those observed with T5-large.

Input Question
Where would you put a light?
Answer Options
(op1) ocean; (op2) desk; (op3) universe; (op4) attic; (op5) chair
Label
desk
Gold
Desk is a piece of furniture with a flat or sloping surface and typically with drawers, at which one can read, write or do other work. We would put a light on desk. Putting light in ocean and Universe is not useful as light is needed where we live or work. Attic is the room partly inside the roof of a building and sparsely used so we don't need a light there. Chairs are for sitting and you can't put a light on chair.
Leaky
The answer is desk.
Gold − Leaky
Ours (LAREV): **2.4955** RORA: 0.9648 REV: 0.2298

Table 5.5: A representative ECQA example evaluated with T5-large. The upper block shows task information, while the lower block compares Gold and Leaky rationales and their resulting difference.

5.4.2 Component-wise Analysis of Regularization Effects

Table 5.6 analyzes the individual and joint effects of the IRM loss and the leakage probe loss on ECQA using T5-large. The objective of this experiment is to assess whether the two regularization components provide complementary benefits for baseline-conditioned rationale evaluation.

doi:10.6342/NTU202600527

Methods ↓	λ_{IRM}	λ_{probe}	Gold − Degraded Rationales →			
			Gold − Leaky	Gold − Gold-Leaky	Gold − Vacuous	SUM
REV	N/A	N/A	0.0768	0.1402	0.3374	0.5544
Ours	0	0.0005	-0.4092	-0.0798	-0.0283	-0.5173
Ours	5	0	-0.0423	0.5249	0.1402	0.6228
Ours	5	0.0005	**1.4022**	**0.5717**	**1.7340**	**3.7079**

Table 5.6: Analysis on ECQA (T5-large), showing the effect of IRM and leakage probe regularization on relative score separations. Higher values indicate stronger discrimination between informative and degraded rationales.

When either regularization is applied in isolation, the improvements over REV remain limited. Using the leakage probe alone ($\lambda_{\text{IRM}} = 0$) results in unstable and often negative separations across rationale variants, suggesting that penalizing residual baseline leakage without enforcing invariance is insufficient. Similarly, applying IRM alone ($\lambda_{\text{probe}} = 0$) yields modest gains in specific comparisons, such as Gold − Gold-Leaky, but fails to consistently distinguish Gold rationale from the other degraded ones.

In contrast, jointly applying IRM and the leakage probe results in substantially larger and more consistent separations across all rationale variants. Across all comparisons, this combined setting achieves the strongest performance and yields a marked increase in the overall SUM metric. These results indicate that IRM and leakage probe regularization address distinct yet complementary failure modes of baseline-conditioned evaluation, and that their joint application is essential for obtaining a reliable and discriminative rationale evaluation signal.

5.4.3 Generalization of the Leakage-Aware Evaluator

Table 5.7 reports the test-set answer accuracy on ECQA when predictions are generated by the leakage-aware evaluator Φ_{LA} under different rationale variants. This analysis examines whether enforcing leakage-aware constraints on Φ_{LA} affects its generalization

doi:10.6342/NTU202600527

Methods ↓	Rationale Variants →			
	Gold	**Gold-Leaky**	**Vacuous**	**Leaky**
REV	**77.16**	92.75	67.18	92.02
Ours	76.57	**93.48**	**73.43**	**93.44**

Table 5.7: Test-set accuracy (%) comparison between REV and our method on ECQA under different rationale variants. Despite introducing IRM and leakage probe regularization, our method maintains comparable or improved generalization performance.

performance.

At first glance, some degree of performance degradation might be expected. Since Φ_{LA} is trained on inputs of the form $r + b$ while explicitly discouraging reliance on baseline information b through IRM and leakage probe regularization, the model is constrained from exploiting label-relevant cues that may be present in the baseline. Such restrictions could potentially hinder prediction accuracy, especially when the baseline itself is informative.

However, the results do not support this concern. Across all rationale variants, Φ_{LA} achieves accuracy that is comparable to that of the REV evaluator. In particular, Φ_{LA} slightly outperforms REV on the Gold-Leaky, Vacuous, and Leaky variants, while exhibiting only a marginal decrease on the Gold condition.

These results suggest that enforcing leakage-aware constraints does not meaningfully harm the evaluator's predictive performance. Instead, Φ_{LA} maintains competitive generalization behavior while reducing its reliance on baseline shortcuts, indicating that leakage-aware training can improve the robustness of rationale evaluation without sacrificing answer accuracy.

doi:10.6342/NTU202600527

5.4.4 Hyperparameter Sensitivity

The leakage probe penalty λ_{probe} controls the degree to which the evaluator Φ_{LA} is restricted from relying on residual label-relevant information in the baseline input. From a conventional perspective, strengthening this restriction would be expected to reduce the amount of readily accessible predictive information, and thus to make optimization more difficult.

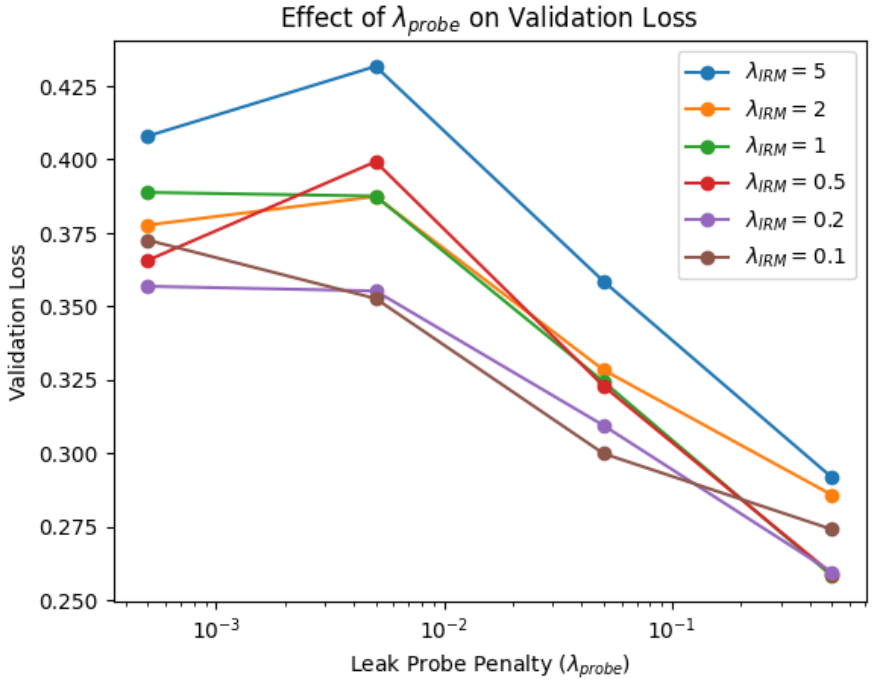


Figure 5.2: Effect of λ_{probe} on the validation loss when training the leakage-aware evaluator Φ_{LA} on ECQA using the T5-large backbone under different λ_{IRM} settings. The x-axis is shown in log scale.

Figure 5.2 reports the validation loss achieved during training on ECQA with the T5-large backbone as λ_{probe} varies across several orders of magnitude under different λ_{IRM} settings. Contrary to the above intuition, increasing λ_{probe} does not lead to higher validation loss. Instead, validation loss consistently decreases across all values of λ_{IRM}.

This behavior suggests that constraining access to residual baseline information does

doi:10.6342/NTU202600527

not impair the learning process of Φ_{LA}. Rather than reducing usable signal, the probe-based penalty appears to steer the evaluator toward alternative information sources, most notably the rationale r, resulting in improved validation performance. Importantly, this trend is observed consistently across different IRM strengths, indicating that the effect of λ_{probe} is robust and not tied to a specific invariance setting.

5.4.5 Robustness across Task Models

In addition to evaluating rationales constructed from human annotations, we further examine whether the evaluation signal remains stable when rationales are generated by different task models. Specifically, for each test instance, we replace the rationale r with a generated rationale from one of four task models (GPT-4, Gemini-2.5-Pro, Llama-3.1-8B-Instruct, and Flan-T5-Large), using a fixed prompting template described in Appendix C.3, while keeping the baseline input b and the evaluator backbone fixed. We report results for both REV and our method (LAREV) under the same rationale sources.

Methods ↓	**Task Models**	**Gold − Degraded Rationales** →			
		Gold − Leaky	**Gold − Gold-Leaky**	**Gold − Vacuous**	**SUM**
Ours	GPT-4	0.9459	0.3505	1.2777	2.5741
Ours	Gemini-2.5-Pro	0.8002	0.4446	1.1320	2.3768
Ours	Llama-3.1-8B-Instruct	0.8224	0.3465	1.1542	2.3231
Ours	Flan-T5-Large	1.2365	0.9695	1.5683	3.7743
REV	GPT-4	0.0322	0.1071	0.2928	0.4321
REV	Gemini-2.5-Pro	-0.0907	0.1378	0.1699	0.2170
REV	Llama-3.1-8B-Instruct	-0.0299	0.1075	0.2307	0.3083
REV	Flan-T5-Large	-0.2401	0.0640	0.0205	-0.1556

Table 5.8: Evaluator robustness across task models on ECQA. We fix the evaluator backbone and baseline input b, and vary only the rationale source. Columns report score differences between the Gold rationale and degraded variants (Gold − Leaky, Gold − Gold-Leaky, and Gold − Vacuous), with **SUM** denoting the aggregate separation. Higher values indicate clearer discrimination.

Table 5.8 shows that our evaluator yields consistently larger and strictly positive sep-

doi:10.6342/NTU202600527

arations across all task models, whereas REV exhibits much smaller separations and becomes negative under some rationale sources. This indicates that the impact of baseline leakage on REV-style evaluation persists across different rationale generators. Figure 5.3 provides a visual summary of these trends. By explicitly constraining reliance on the baseline input, LAREV produces a more stable and reliable discriminative signal under varying rationale sources.

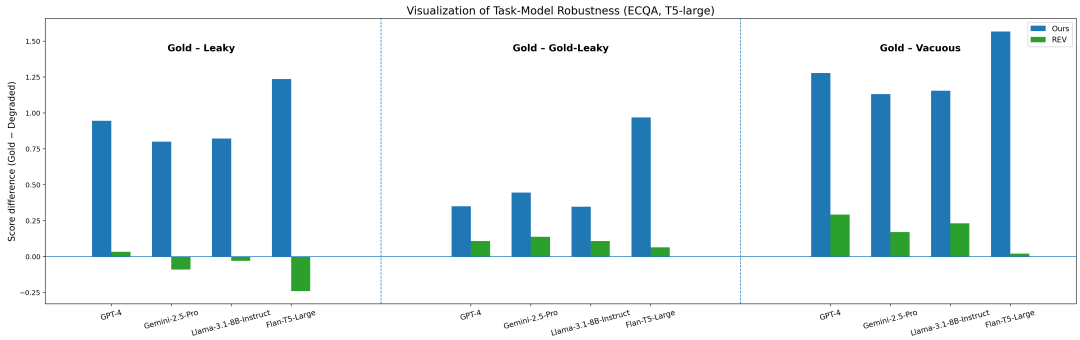


Figure 5.3: Visualization of evaluator robustness across task models on ECQA using the T5-large backbone. Bars indicate score differences between the Gold rationale and degraded variants (Gold − Leaky, Gold − Gold-Leaky, and Gold − Vacuous). Each group corresponds to a degradation setting, and larger positive values indicate clearer separation between informative and less informative rationales.

5.4.6 Rationale Evaluation on e-SNLI

Methods ↓	**Rationale Variants →**			
	Gold	**Gold-Leaky**	**Vacuous**	**Leaky**
Ours	-2.9755	-3.9342	-2.9889	-3.4332
RORA	-3.5932	-3.9095	-3.5167	-3.5391
REV	-1.9038	-1.6886	-1.0780	-1.2272

Table 5.9: Absolute rationale scores on e-SNLI using T5-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution.

We additionally evaluate information-based rationale evaluation methods on the e-SNLI dataset to assess whether the observed trends generalize beyond ECQA. As shown in Table 5.9, across all rationale variants, LAREV assigns lower absolute scores than REV

doi:10.6342/NTU202600527

Methods ↓	Gold − Degraded Rationales →			SUM
	Gold − Leaky	Gold − Gold-Leaky	Gold − Vacuous	
Ours	**0.4577**	**0.9587**	**0.0134**	**1.4298**
RORA	-0.0541	0.3163	-0.0765	0.1857
REV	-0.6766	-0.2152	-0.8258	-1.7176

Table 5.10: Relative sensitivity of different evaluation methods on e-SNLI (T5-large), measured as score differences between the Gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales.

because it more strictly limits the influence of baseline information on model predictions. As discussed in Section 5.4.1, absolute score magnitudes are therefore not directly comparable across methods.

Relative score differences between the Gold rationale r and degraded variants are reported in Table 5.10. LAREV consistently produces positive separations across all comparisons, whereas REV fails to distinguish Gold rationales from Leaky or Vacuous ones under this setting. These results indicate that explicitly controlling baseline leakage leads to more reliable information-based rationale evaluation signals on e-SNLI as well. Figure 5.4 provides a visual summary of the results in Table 5.10.

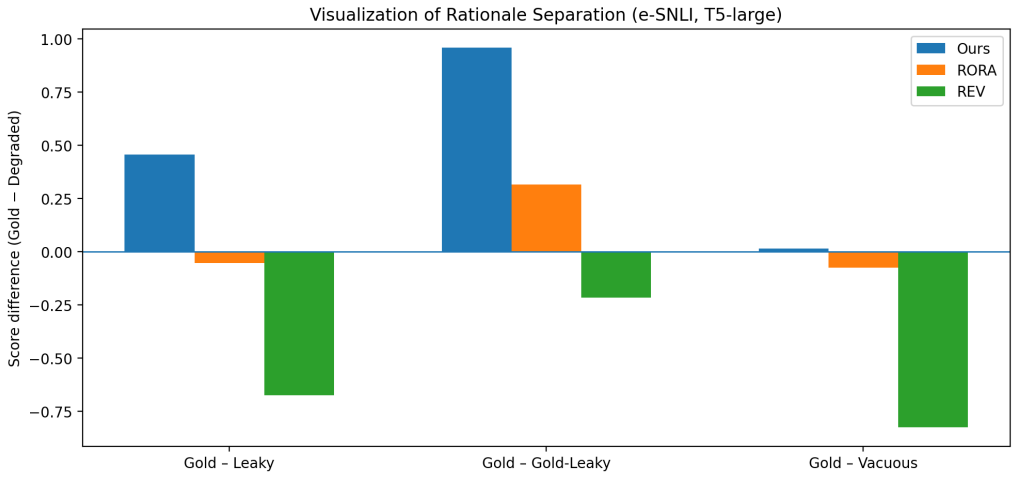
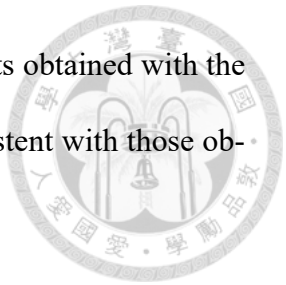


Figure 5.4: Visualization of rationale separation on e-SNLI using the T5-large backbone. Bars indicate score differences between the Gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales.

doi:10.6342/NTU202600527

We additionally report the absolute and relative evaluation results obtained with the BART-large backbone in Appendix B.2. The overall trends are consistent with those observed for T5-large.



doi:10.6342/NTU202600527

Chapter 6

Conclusion

Baseline-conditioned rationale evaluation, as defined in the REV [14] framework, aims to quantify the additional usable information contributed by a free-text rationale r beyond a baseline input b. However, when b itself contains residual label-relevant cues, baseline leakage can cause evaluation signals to be inflated and unreliable. We address the reliability issue by proposing LAREV, a leakage-aware training framework for rationale evaluation that preserves the baseline-conditioned objective while improving robustness to leakage in b.

6.1 Summary of Contributions

We make three primary contributions to information-based rationale evaluation. We first characterize baseline leakage as a critical failure mode under baseline-conditioned evaluation, showing how label-relevant signal in the baseline input b can distort the measured contribution of a rationale r. Building on this analysis, we propose LAREV, a leakage-aware training framework for the evaluator that incorporates two complementary regularization mechanisms. An IRM-based invariance constraint discourages reliance on environment-specific shortcut cues derived from b, while a probe-based penalty explic-

doi:10.6342/NTU202600527

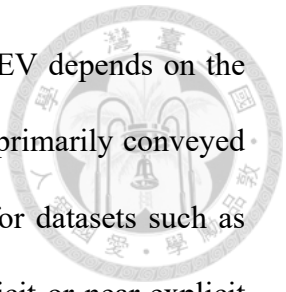
itly controls dependence on residual label-relevant information that remains accessible through masked baseline inputs. Together, these components encourage the evaluator to ground its predictions more strongly in the information provided by the rationale r, rather than in leakage from the baseline input b. Finally, extensive empirical evaluations on both an open-label dataset (ECQA) and a fixed-label dataset (e-SNLI) demonstrate that leakage-aware training substantially improves the discriminative strength of information-based rationale evaluation, while preserving the predictive performance of the evaluator.

6.2 Limitations

The construction of the antonym environment (E3) in LAREV relies on attribution-guided identification of leakage terms, followed by automated rewriting to produce semantically altered baseline inputs. In practice, integrated gradients may occasionally assign high attribution scores to function words or pronouns, such as articles (e.g., “a”, “the”) or gendered pronouns (e.g., “he”, “she”). This behavior is counterintuitive from a human perspective, as such tokens are not typically regarded as meaningful sources of label-relevant information in the baseline input.

When attribution highlights these function words or pronouns, the resulting antonym environment is constructed by modifying tokens that lack clear semantic content. As a result, rewriting such tokens—regardless of the specific rewriting strategy—may have limited impact on the underlying predictive signal used by the model. This mismatch between attribution signals and human-interpretable semantics can reduce the effectiveness of the antonym environment, thereby weakening the invariance constraint imposed during training.

doi:10.6342/NTU202600527



More broadly, the effectiveness of IRM-based training in LAREV depends on the assumption that spurious label correlations in the baseline input are primarily conveyed through surface-level lexical cues. This assumption is reasonable for datasets such as ECQA and e-SNLI, where baseline leakage often manifests as explicit or near-explicit answer tokens. However, this assumption may not hold for tasks that rely on structured or algorithmic reasoning, such as mathematical problem solving or code understanding. In such cases, token-level environment construction may be insufficient to produce meaningful differences across environments. In such cases, the resulting invariance constraints may be weak or ineffective, limiting the applicability of the proposed training strategy.

6.3 Future Work

Further improving the separation between Gold and Leaky rationales remains an important direction for extending leakage-aware rationale evaluation. While LAREV focuses on reducing reliance on baseline leakage, Gold rationales may still contain answer-identifying cues that partially overlap with leaky variants. Introducing additional constraints to control label leakage within the rationale itself may therefore provide further benefits.

In particular, incorporating invariance constraints, such as IRM-style regularization, directly into the rationale r could encourage the evaluator to rely more on explanation-level reasoning rather than on surface-level label cues embedded in the rationale text. Such an extension may further sharpen the distinction between informative and leaky rationales, leading to more fine-grained and reliable rationale evaluation signals.

doi:10.6342/NTU202600527

References

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

[2] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36:11809–11822, 2023.

[3] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. arXiv preprint arXiv:1606.04155, 2016.

[4] Mahdi Dhaini, Juraj Vladika, Ege Erdogan, Zineb Attaoui, and Gjergji Kasneci. Can llm-generated textual explanations enhance model classification performance? an empirical study. In International Conference on Artificial Neural Networks, pages 192–204. Springer, 2025.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM

doi:10.6342/NTU202600527

SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

[6] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. arXiv preprint arXiv:1906.02361, 2019.

[7] Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. Learning to rationalize for nonmonotonic reasoning with distant supervision. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 12592–12601, 2021.

[8] Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing. arXiv preprint arXiv:2102.12060, 2021.

[9] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7:49–72, 2019.

[10] Sawan Kumar and Partha Talukdar. Nile: Natural language inference with faithful natural language explanations. arXiv preprint arXiv:2005.12116, 2020.

[11] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages 610–623, 2021.

[12] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. Advances in Neural Information Processing Systems, 31, 2018.

doi:10.6342/NTU202600527

[13] Miruna Clinciu, Arash Eshghi, and Helen Hastie. A study of automatic metrics for the evaluation of natural language explanations. arXiv preprint arXiv:2103.08545, 2021.

[14] Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. Rev: Information-theoretic evaluation of free-text rationales. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2007–2030, 2023.

[15] Ivan Aslanov and Ernesto Guerra. Tautological formal explanations: does prior knowledge affect their satisfiability? Frontiers in Psychology, 14:1258985, 2023.

[16] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? arXiv preprint arXiv:2010.04119, 2020.

[17] Zheng Ping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. Rora: Robust free-text rationale evaluation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1070–1087, 2024.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36:46595–46623, 2023.

[19] Ziang Li, Manasi Ganti, Zixian Ma, Helena Vasconcelos, Qijia He, and Ranjay Krishna. Rethinking human preference evaluation of llm rationales. arXiv preprint arXiv:2509.11026, 2025.

doi:10.6342/NTU202600527

[20] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. arXiv preprint arXiv:2212.07919, 2022.

[21] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness. arXiv preprint arXiv:2304.10703, 2023.

[22] Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. Tailoring self-rationalizers with multi-reward distillation. arXiv preprint arXiv:2311.02805, 2023.

[23] Sarah Wiegreffe, Ana Marasović, and Noah A Smith. Measuring association between labels and free-text rationales. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10266–10284, 2021.

[24] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In International Conference on Machine Learning, pages 5988–6008. PMLR, 2022.

[25] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. arXiv preprint arXiv:2002.10689, 2020.

[26] John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D Manning. Conditional probing: measuring usable information beyond a baseline. arXiv preprint arXiv:2109.09234, 2021.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep

doi:10.6342/NTU202600527

networks. In International conference on machine learning, pages 3319–3328. PMLR, 2017.

[28] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

[29] Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. arXiv preprint arXiv:1809.02922, 2018.

[30] Jifan Chen, Eunsol Choi, and Greg Durrett. Can nli models verify qa systems' predictions? arXiv preprint arXiv:2104.08731, 2021.

[31] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for commonsenseqa: New dataset and models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3050–3065, 2021.

[32] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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), pages 4149–4158, 2019.

[33] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 632–642, 2015.

doi:10.6342/NTU202600527

Appendix A
Experimental Pipeline

doi:10.6342/NTU202600527

Algorithm A.1: LAREV Experimental Pipeline

Input: Dataset $\mathcal{D}$ with input questions, labels, and gold rationales
Output: Absolute LAREV scores on $\mathcal{D}_{\text{test}}$ for rationale variants $\tilde{r}$ (e.g., Table 5.3)

Stage I: Data preparation;
(1) Prepare dataset splits;
Split $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$;
(2) Construct baseline rationales b;
Generate baseline rationales for all instances in $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{test}}$;
(3) Construct evaluation rationale variants $\tilde{r}$;
On $\mathcal{D}_{\text{test}}$, create rationale variants (Gold, Leaky, Gold-Leaky, Vacuous);

Stage II: Model training;
(4) Train baseline model Φ_{base};
Train Φ_{base} on baseline inputs b only from $\mathcal{D}_{\text{train}}$ (validate on $\mathcal{D}_{\text{val}}$);
(5) Train regular model Φ;
Train Φ on concatenated inputs $r + b$ from $\mathcal{D}_{\text{train}}$ (validate on $\mathcal{D}_{\text{val}}$);

Stage III: Leakage-aware refinement;
(6) Compute integrated gradients on Φ_{base};
For each instance in $\mathcal{D}_{\text{train}}$, run IG under the baseline-only input b to identify influential baseline tokens;
(7) Generate multiple environments;
Using the IG-identified token(s), construct multiple environments (e.g., E1, E2, E3) for instances in $\mathcal{D}_{\text{train}}$ to induce variation in spurious baseline cues;
(8) Train leakage probe model ψ;
Initialize ψ from Φ, freeze encoder parameters, and train only the decoder to predict labels from encoder representations induced by the masked baseline input $\tilde{b}$ on $\mathcal{D}_{\text{train}}$ (validate on $\mathcal{D}_{\text{val}}$);
(9) Train leakage-aware evaluator Φ_{LA};
Optimize leakage-aware objectives using the constructed environments and probe signals on $\mathcal{D}_{\text{train}}$ (validate on $\mathcal{D}_{\text{val}}$);

Stage IV: Evaluation;
(10) Score rationale variants;
For each r on $\mathcal{D}_{\text{test}}$, evaluate Φ_{LA} on $r + b$ and Φ_{base} on b, compute the LAREV score (Eq. (4.4)), and aggregate over $\mathcal{D}_{\text{test}}$;

doi:10.6342/NTU202600527

Appendix B
Results with BART-large Backbone

B.1 ECQA Results

Methods ↓	Rationale Variants →			
	Gold	**Gold-Leaky**	**Vacuous**	**Leaky**
Ours	-1.2284	-1.6152	-1.4902	-1.8737
RORA	-1.0633	-1.2763	-1.0721	-1.1169
REV	-0.4378	-0.5099	-0.4377	-0.4288

Table B.1: Absolute rationale scores on ECQA using BART-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution.

Methods ↓	Gold - Degraded Rationales →			SUM
	Gold − Leaky	**Gold − Gold-Leaky**	**Gold − Vacuous**	
Ours	**0.6453**	**0.3868**	**0.2618**	**1.2939**
RORA	0.0536	0.2130	0.0088	0.2754
REV	-0.0090	0.0721	-0.0001	0.0630

Table B.2: Relative sensitivity of different evaluation methods on ECQA (BART-large), measured as score differences between the gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales.

doi:10.6342/NTU202600527

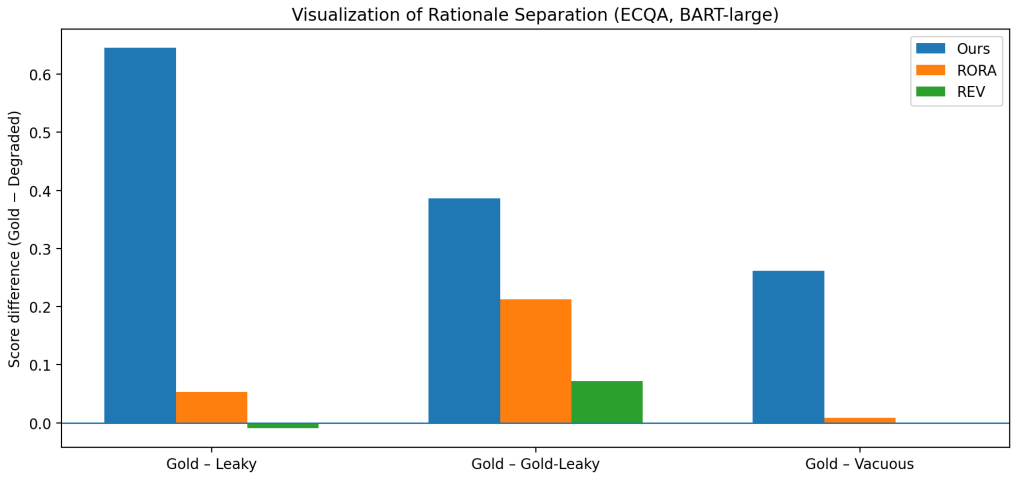


Figure B.1: Visualization of rationale separation on ECQA using the BART-large backbone. Bars indicate score differences between the gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales.

doi:10.6342/NTU202600527

B.2 e-SNLI Results

Methods ↓	Rationale Variants →			
	Gold	**Gold-Leaky**	**Vacuous**	**Leaky**
Ours	-1.7695	-2.8048	-2.2582	-2.4910
RORA	-1.0963	-1.1981	-0.7277	-0.4705
REV	-0.6327	-0.7356	-0.5316	-0.6450

Table B.3: Absolute rationale scores on e-SNLI using BART-large. Columns correspond to different rationale variants, while rows indicate evaluation methods. More negative values indicate lower estimated informational contribution.

Methods ↓	Gold - Degraded Rationales →			SUM
	Gold − Leaky	**Gold − Gold-Leaky**	**Gold − Vacuous**	
Ours	**0.7215**	**1.0353**	**0.4887**	**2.2455**
RORA	-0.6258	0.1018	-0.3686	-0.8926
REV	0.0123	0.1029	-0.1011	0.0141

Table B.4: Relative sensitivity of different evaluation methods on e-SNLI (BART-large), measured as score differences between the gold rationale and degraded variants. Larger values indicate clearer separation between informative and less informative rationales.

doi:10.6342/NTU202600527

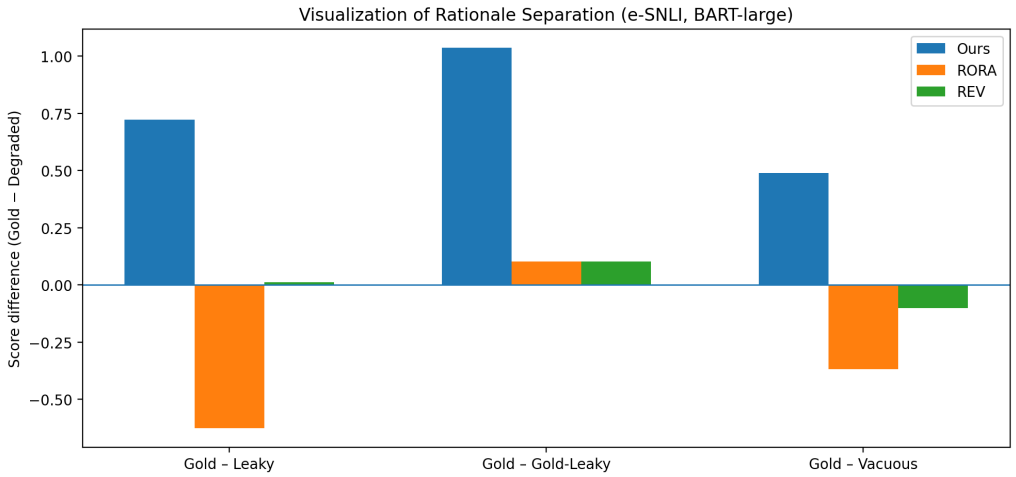


Figure B.2: Visualization of rationale separation on e-SNLI using the BART-large backbone. Bars indicate score differences between the gold rationale and degraded variants under different evaluation methods. Degradation settings are grouped, and larger positive values correspond to clearer separation between informative and less informative rationales.

doi:10.6342/NTU202600527

Appendix C
Additional Illustrative Examples

C.1 Rationale Variants in e-SNLI

Field	Content (e-SNLI Example)
Premise	A couple walk hand in hand down a street.
Hypothesis	A couple is sitting on a bench.
Label	contradiction
Gold rationale	The couple cannot be walking and sitting at the same time.
Leaky rationale	The answer is contradiction.
Gold-Leaky rationale	The couple cannot be walking and sitting at the same time. The answer is contradiction.
Vacuous rationale	A pair of individuals walk hand-in-hand down a street, while another couple sit on a bench.

Table C.1: An illustrative e-SNLI test example showing different rationale variants constructed for evaluation in a fixed-label setting. All variants correspond to the same premise–hypothesis pair, with the gold label being contradiction.

doi:10.6342/NTU202600527

C.2 LLM Prompt for Antonym Environment Instantiation

The following prompt template is used to instantiate the antonym environment (E3) by querying the `gemini-2.5-flash-lite` model as a deterministic text transformation tool.

```
Please change the masked part of the following sentence with an antonym.
Do not change any other part of the sentence except the masked part.
Output only a single sentence.

Question context: <QUESTION>
Original answer: <ANSWER>
Masked sentence: <MASKED_SENTENCE>
The masked term to be changed: <LEAKAGE_TERM>

Antonym version:
```

C.3 Prompt for Rationale Generation from Task Models

The following prompt template is used to generate free-text rationales from task models in our experiments. Given a question–answer pair, we use the same prompt to generate rationales from multiple task models, including GPT-4, Gemini-2.5-Pro, Llama-3.1-8B-Instruct, and Flan-T5-Large.

```
Question: <QUESTION>
Answer: <ANSWER>

Explain why the answer is correct in detail.
Assume the given answer is correct. Do not challenge or change the answer.
Please finish in one sentence.
```


doi:10.6342/NTU202600527