

Progressive Training for Explainable Citation-Grounded Dialogue: Reducing Hallucination to Zero in English-Hindi LLMs

Vedant Pandya

School of Artificial Intelligence and Data Engineering (SAIDE)
Indian Institute of Technology Jodhpur
m25ai1132@iitj.ac.in

Abstract

Knowledge-grounded dialogue systems aim to generate informative, contextually relevant responses by conditioning on external knowledge sources. However, most existing approaches focus exclusively on English, lack explicit citation mechanisms for verifying factual claims, and offer limited transparency into model decision-making. We present XKD-DIAL, a progressive four-stage training pipeline for **explainable, knowledge-grounded dialogue generation** in a bilingual (English–Hindi) setting, comprising: (1) multilingual adaptation, (2) English dialogue SFT with citation grounding, (3) bilingual dialogue SFT, and (4) GRPO alignment with citation-aware rewards. We evaluate six models spanning encoder-decoder (250M–3B) and decoder-only (1B–7B) architectures at every pipeline stage. Our key contributions are: (i) three post-hoc explainability analyses - cross-attention alignment, Integrated Gradients attribution, and occlusion-based causal grounding - applied systematically across the training trajectory to reveal *how* citation behaviour is learned, not only *whether* it is learned; (ii) citation-grounded SFT reduces hallucination to 0.0% for encoder-decoder models from Stage 2 onward; (iii) the progressive pipeline prevents catastrophic forgetting while improving Hindi capabilities; (iv) smaller models match larger models on English after SFT; and (v) GRPO provides marginal improvement over well-designed SFT for structured citation tasks. We evaluate across six automatic metrics (BLEU, ROUGE, BERTScore, FactScore, Citation-F1, hallucination rate).

Keywords: Knowledge-Grounded Dialogue, Multilingual NLP, Explainability, Large Language Models, Citation Generation, Hindi, GRPO, Hallucination Reduction

1 Introduction

Knowledge-grounded dialogue generation has emerged as a critical research direction for building conversational AI systems that produce factually accurate, informative responses [Dinan et al., 2019, Rashkin et al., 2021]. By conditioning response generation on retrieved knowledge passages, these systems can mitigate the chronic hallucination problem of large language models (LLMs) - where models generate plausible-sounding but factually incorrect information [Ji et al., 2023]. However, current approaches suffer from three fundamental limitations.

First, the monolingual bottleneck. The vast majority of knowledge-grounded dialogue research focuses exclusively on English [Dinan et al., 2019, Rashkin et al., 2021, Kim et al., 2020]. For languages like Hindi - spoken by over 600 million people - there exists no standard benchmark, no established training methodology, and no systematic study of how knowledge-grounded dialogue systems perform in a bilingual setting. Extending such systems to Hindi is particularly challenging due to: (a) limited availability of Hindi dialogue corpora with knowledge annotations, (b) morphological richness and free word order that complicate both generation and evaluation, and (c) the need to handle code-switching and cross-lingual knowledge transfer.

Second, the absence of verifiable citations. While retrieval-augmented generation (RAG) systems retrieve relevant passages [Lewis et al., 2020, Guu et al., 2020], the generated response typically does not indicate *which* passage supports *which* claim. Without explicit citation markers (e.g., “According to [1], ...”), users cannot verify factual claims against their sources, undermining trust and transparency. Recent work on

attributed text generation [Rashkin et al., 2021] has highlighted this gap, but citation-grounded training for dialogue remains underexplored.

Third, the opacity of model decisions. Even when a model generates a correct, grounded response, it provides no insight into *why* it selected particular knowledge passages or *how* it composed the response. This opacity is especially problematic for citation-grounded systems: a model may produce the correct citation marker [1] without genuinely conditioning its output on passage 1, making citation accuracy an unreliable quality signal on its own. Interpretability methods - cross-attention visualization [Jain and Wallace, 2019, Wiegrefe and Pinter, 2019], Integrated Gradients [Sundararajan et al., 2017], and occlusion-based causal grounding [Lei et al., 2016] - can expose this dissociation, but their systematic application to knowledge-grounded dialogue generation across an entire training trajectory has not been attempted.

1.1 Our Approach

We propose XKD-DIAL (**EX**plainable **K**nowledge-Grounded **D**ialogue), a progressive four-stage training pipeline designed to address all three limitations simultaneously. Our key insight is that complex multilingual, knowledge-grounded generation capabilities can be built incrementally, where each training stage adds a specific skill while preserving previously learned capabilities:

- 1. Stage 1: Multilingual Adaptation.** English-Hindi translation training to build bilingual representations, particularly for models with limited Hindi pretraining exposure.
- 2. Stage 2: English Dialogue SFT.** Supervised fine-tuning on English knowledge-grounded dialogue with explicit citation markers, teaching the model to generate responses that attribute claims to specific knowledge passages.
- 3. Stage 3: Bilingual Dialogue SFT.** Extension to Hindi dialogue with citations, leveraging cross-lingual transfer from Stage 2.
- 4. Stage 4: GRPO Alignment.** Reinforcement learning via Group Relative Policy Optimization [Shao et al., 2024] with a composite reward function that incentivizes citation accuracy, factual consistency, and penalizes hallucination.

1.2 Contributions

Our main contributions are as follows:

- 1. A progressive training pipeline for bilingual knowledge-grounded dialogue.** We introduce a four-stage methodology that incrementally builds multilingual, citation-grounded dialogue capabilities while preventing catastrophic forgetting. To our knowledge, this is the first systematic pipeline for English-Hindi knowledge-grounded dialogue with citations.
- 2. Comprehensive cross-architecture empirical study.** We evaluate six models across two architecture families (encoder-decoder and decoder-only) spanning 250M to 7B parameters, with each model evaluated at every pipeline stage (30 total evaluation runs). This provides fine-grained ablation of which stage contributes which capability.
- 3. Citation-grounded hallucination reduction.** We observe that training with explicit citation format substantially reduces hallucination rates (reaching 0.0% under automatic NLI-based evaluation from Stage 2 onward for encoder-decoder models), suggesting citation-grounded SFT as a promising anti-hallucination strategy warranting further investigation including human evaluation.
- 4. Empirical analysis of GRPO for structured tasks.** We provide an empirical characterisation of GRPO behaviour in our experimental configuration ($\beta=0.04$, 500 steps), finding marginal improvement over SFT. This contributes to the broader discussion of when RL alignment is beneficial, though comprehensive hyperparameter exploration remains future work.
- 5. Explainability analysis.** We apply attention visualization and token attribution methods to analyze how models attend to knowledge passages during generation, providing interpretability insights for knowledge-grounded dialogue.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 details our four-stage training pipeline. Section 4 describes the experimental setup including datasets, models, and evaluation metrics. Section 5 presents results and analysis. Section 6

discusses key findings and their implications. Section 7 concludes with future directions.

2 Related Work

Our work intersects several active research areas: knowledge-grounded dialogue systems, multilingual language models, reinforcement learning from human feedback, and explainability in natural language generation. We review each in turn, highlighting the gaps that motivate our approach.

2.1 Knowledge-Grounded Dialogue

The seminal Wizard of Wikipedia [Dinan et al., 2019] established the paradigm of conditioning dialogue responses on retrieved Wikipedia passages, demonstrating that access to external knowledge significantly improves informativeness and factual accuracy. Subsequent work addressed the critical problem of *faithfulness*: FaithDial [Rashkin et al., 2021] introduced a benchmark specifically targeting hallucination in knowledge-grounded dialogue, showing that standard models frequently generate claims unsupported by the provided knowledge. The DSTC9 shared task [Kim et al., 2020] extended the challenge to unstructured knowledge access, requiring models to identify relevant knowledge snippets from FAQs and reviews before generating responses.

On the modeling side, BlenderBot 2.0 [Shuster et al., 2022] combined internet search with long-term memory for open-domain conversation, while Atlas [Izacard et al., 2023] demonstrated that retrieval-augmented few-shot learning can match much larger models. The RAG framework [Lewis et al., 2020] and REALM [Guu et al., 2020] established end-to-end training of retriever-generator systems, and Izacard and Grave [2021] showed that Fusion-in-Decoder approaches effectively aggregate multiple retrieved passages.

A critical gap in this literature is the absence of **explicit citation mechanisms**. While these systems retrieve and condition on knowledge, the generated responses do not indicate which passage supports which claim. Our work addresses this by training models to produce inline citations (e.g., “According to [1], ...”), enabling users to verify factual claims against their sources.

2.2 Multilingual and Hindi Language Models

The development of multilingual pretrained models has progressed rapidly. mT5 [Xue et al., 2021] extended the T5 text-to-text framework to 101 languages, while BLOOM [BigScience Workshop, 2023] provided an open-access 176B-parameter multilingual model. For Indian languages specifically, IndicBART [Dabre et al., 2022] offered a pretrained seq2seq model covering 11 Indic languages, MuRIL [Khanuja et al., 2021] provided BERT-style representations for Indian languages, and IndicTrans2 [Gala et al., 2023] achieved state-of-the-art machine translation across all 22 scheduled Indian languages.

On the instruction-tuned front, Flan-T5 [Chung et al., 2024] demonstrated that multi-task instruction tuning dramatically improves zero-shot and few-shot performance. Decoder-only models such as Mistral-7B [Jiang et al., 2023] with its sliding window attention, LLaMA-3 [Meta AI, 2024] with its expanded multilingual training data, and Gemma-2 [Google DeepMind, 2024] have pushed the boundaries of efficient, high-quality generation.

Despite these advances, **knowledge-grounded dialogue in Hindi remains unexplored**. No existing work combines Hindi dialogue generation with citation grounding. Our work addresses this gap by constructing a bilingual English–Hindi pipeline that leverages cross-lingual transfer through progressive training stages.

2.3 Reinforcement Learning for Language Model Alignment

InstructGPT [Ouyang et al., 2022] pioneered the use of Reinforcement Learning from Human Feedback (RLHF) for aligning language models with human preferences, establishing the SFT → Reward Model → PPO pipeline that has become standard practice. However, PPO suffers from training instability and high computational cost due to the need for a separate reward model.

Group Relative Policy Optimization (GRPO) [Shao et al., 2024], introduced by DeepSeek for mathematical reasoning, offers an alternative that eliminates the need for a critic model. GRPO generates multiple outputs per prompt, ranks them by reward, and uses the relative ranking as the training signal. This approach is more computationally efficient and

has been shown to be effective for tasks with well-defined reward signals.

Our work applies GRPO to knowledge-grounded dialogue with a **composite citation-aware reward function** that combines factual consistency (NLI-based), entity overlap, citation attribution accuracy, and hallucination penalties. A key finding of our study is that GRPO provides marginal contribution over well-designed SFT for this task - suggesting that when the output format is highly structured (citation-grounded responses), SFT alone may be sufficient.

2.4 Explainability in Neural Text Generation

The interpretability of neural models has been a subject of active debate. Jain and Wallace [2019] argued that attention weights are unreliable explanations, while Wiegrefe and Pinter [2019] showed that attention can be a useful, if imperfect, explanation signal under certain conditions. Tang et al. [2020] specifically studied attention faithfulness in neural machine translation, finding that faithful attention improves both translation quality and interpretability.

Beyond attention, gradient-based methods offer complementary interpretability. Integrated Gradients [Sundararajan et al., 2017] provides axiomatic attribution by accumulating gradients along a path from a baseline to the input, while SHAP [Lundberg and Lee, 2017] offers game-theoretic attribution values. For text generation, Lei et al. [2016] proposed extracting rationales - minimal subsets of input that suffice for the prediction - as a form of explanation.

In the context of knowledge-grounded dialogue, explainability is particularly important: users need to understand not just *what* the model says, but *which knowledge passage* influenced *which part* of the response. Our work applies attention visualization and token attribution to analyze how models attend to knowledge passages during citation-grounded generation, providing the first such analysis for multilingual knowledge-grounded dialogue.

2.5 Position of Our Work

Table 1 summarizes the positioning of our work relative to existing approaches. To our knowledge, XKD-DIAL is the first system that simultaneously addresses all four dimensions: knowledge grounding with citations, multilingual (English-

Table 1: Comparison with related work across key dimensions. XKD-DIAL is the first to address all four dimensions simultaneously.

System	Cite	Hindi	RL	XAI
Wizard of Wikipedia	×	×	×	×
FaithDial	×	×	×	×
BlenderBot 2.0	×	×	×	×
RAG	×	×	×	×
Atlas	×	×	×	×
InstructGPT	×	×	✓	×
DeepSeekMath	×	×	✓	×
XKD-Dial (Ours)	✓	✓	✓	✓

Hindi) support, RL-based alignment, and model explainability.

3 Methodology

We present a progressive four-stage training pipeline that incrementally builds multilingual, citation-grounded dialogue capabilities. The overall system architecture is illustrated in Figure 7 (see Appendix). The key design principle is *skill composition*: each stage adds a specific capability while preserving those learned in previous stages.

3.1 Problem Formulation

Given a user query q (in English or Hindi) and a set of retrieved knowledge passages $\mathcal{K} = \{k_1, k_2, \dots, k_n\}$, the task is to generate a response r that:

1. Is factually consistent with \mathcal{K} ,
2. Contains explicit citation markers $[i]$ linking claims to specific passages $k_i \in \mathcal{K}$,
3. Is fluent in the query language (English or Hindi), and
4. Does not hallucinate information absent from \mathcal{K} .

The input to the model is a structured prompt:

```

Query: {q}
Knowledge:
[1] {k1}
[2] {k2}
...
Respond using the knowledge above
with citations [1], [2], etc.
```

The expected output is a natural language response with inline citations, e.g., “According to [1], the Eiffel Tower was completed in 1889. It was designed by Gustave Eiffel [2].”

Table 2: Model architectures used in our study.

Model	Type	Params	Layers
Flan-T5-Base	Enc-Dec	250M	12+12
Flan-T5-Large	Enc-Dec	780M	24+24
Flan-T5-XL	Enc-Dec	3B	24+24
LLaMA-3.2-1B	Dec-Only	1B	16
Gemma-2-2B	Dec-Only	2B	26
Mistral-7B	Dec-Only	7B	32

3.2 Model Selection

We select six models spanning two architecture families and a parameter range from 250M to 7B, enabling systematic analysis of how architecture type and model scale affect knowledge-grounded dialogue. Table 2 summarizes the architectures.

The Flan-T5 family [Chung et al., 2024] provides encoder-decoder models instruction-tuned on 1,800+ tasks, offering strong baseline zero-shot performance. The decoder-only models - LLaMA-3.2-1B-Instruct [Meta AI, 2024], Gemma-2-2B-IT [Google DeepMind, 2024], and Mistral-7B-Instruct [Jiang et al., 2023] - represent the more recent autoregressive paradigm. This selection enables three key comparisons: (i) encoder-decoder vs. decoder-only at similar scale, (ii) scaling behavior within architecture families, and (iii) architecture-specific failure modes (Section 5).

3.3 Stage 1: Multilingual Adaptation

The first stage adapts pretrained models to bilingual English-Hindi representations through translation training. This is particularly important for models with limited Hindi exposure in their pre-training corpora.

Training objective. For encoder-decoder models, we train on parallel English-Hindi sentence pairs from the IIT Bombay parallel corpus [Kunchukuttan et al., 2018], using the standard seq2seq cross-entropy loss:

$$\mathcal{L}_{\text{Stage1}} = - \sum_{t=1}^T \log P_{\theta}(y_t | y_{<t}, x) \quad (1)$$

where x is the source sentence and y is the target translation. For decoder-only models, we format translation as an instruction-following task using model-specific chat templates and train with causal language modeling loss on the target portion only.

Training protocol. We train bidirectionally (EN→HI and HI→EN) for a **single epoch** with cosine learning rate scheduling. All models use BFloat16 precision.

Design rationale. Stage 1 is deliberately limited to one epoch to provide *broad bilingual exposure* rather than deep convergence on the translation objective. Over-training on translation risks overwriting the instruction-following capabilities acquired during pretraining, which are essential for Stages 2-4. A single pass through the parallel corpus is sufficient to shift model representations toward bilingual alignment without catastrophic interference with pretrained knowledge. Our ablation (Section 5) confirms that this lightweight adaptation strategy is effective: Stage 1 provides the largest Hindi improvement for the smallest model (Flan-T5-Base: +0.130 Hindi BERTScore), while larger models with stronger multilingual pre-training show smaller but consistent gains.

3.4 Stage 2: English Dialogue SFT

Stage 2 introduces the core dialogue generation capability with citation grounding through supervised fine-tuning on English knowledge-grounded dialogue data.

Data format. Each training example consists of:

- **Input:** A structured prompt containing the user query and numbered knowledge passages.
- **Output:** A natural language response with inline citation markers referencing the knowledge passages.
- **Metadata:** Source dataset, language, and knowledge passage identifiers.

This format is *model-agnostic* - the same JSONL files are used for all six models. For decoder-only models, model-specific chat templates wrap the input-output pair at training time.

Design rationale. Stage 2 is the most impactful stage in our pipeline (Section 5). By training on citation-grounded English dialogue, the model simultaneously learns: (a) dialogue response generation patterns, (b) citation attachment mechanics ([1], [2]), and (c) knowledge grounding - conditioning responses on provided passages. Critically, the citation format acts as an implicit

anti-hallucination mechanism: since every training example contains properly cited responses, the model learns that claims must be supported by numbered references.

3.5 Stage 3: Bilingual Dialogue SFT

Stage 3 extends dialogue capabilities to Hindi while preserving English performance through bilingual fine-tuning.

Data composition. We use a weighted mixture of English and Hindi dialogue examples with citations:

$$\mathcal{L}_{\text{Stage3}} = \alpha \cdot \mathcal{L}_{\text{EN}} + (1 - \alpha) \cdot \mathcal{L}_{\text{HI}} \quad (2)$$

where $\alpha = 0.4$ and $(1 - \alpha) = 0.6$, giving slightly higher weight to Hindi to accelerate Hindi learning while the English inclusion acts as a replay buffer to prevent catastrophic forgetting. The language-specific training dynamics are visualized in Figure 12 (see Appendix).

Cross-lingual transfer. A key finding is that citation formatting learned in Stage 2 (English) transfers effectively to Hindi in Stage 3. The model does not need to re-learn citation mechanics for Hindi - it applies the [1], [2] pattern to Hindi responses automatically. This confirms that citation grounding is a *language-agnostic structural skill* rather than a language-specific one.

3.6 Stage 4: GRPO Alignment

The final stage applies Group Relative Policy Optimization (GRPO) [Shao et al., 2024] to further align the model with citation quality objectives.

GRPO algorithm. For each training prompt x_i , GRPO generates a group of G candidate responses $\{r_i^1, r_i^2, \dots, r_i^G\}$ by sampling from the current policy π_θ . Each response is scored by the reward function $R(r_i^g, \mathcal{K}_i)$, and group-relative advantages are computed:

$$A_i^g = \frac{R(r_i^g) - \mu(\{R(r_i^j)\}_{j=1}^G)}{\sigma(\{R(r_i^j)\}_{j=1}^G) + \epsilon} \quad (3)$$

where μ and σ are the group mean and standard deviation. The policy is updated to maximize:

$$\mathcal{J}(\theta) = \mathbb{E} \left[\sum_{g=1}^G A_i^g \log \pi_\theta(r_i^g | x_i) - \beta \cdot D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right] \quad (4)$$

Table 3: GRPO reward function components with weights.

Sym.	Component	w
r_{fact}	Factual consistency (NLI)	5.0
r_{ent}	Entity overlap	3.0
r_{attr}	Citation attribution	1.5
r_{flu}	Fluency proxy	1.0
r_{len}	Length penalty	-0.1
r_{hal}	Hallucination penalty	-10.0
r_{cite}^+	Correct citation bonus	5.0
r_{cite}^-	Wrong citation penalty	-5.0

where $\beta = 0.04$ is the KL penalty coefficient and π_{ref} is the Stage 3 checkpoint (frozen reference policy).

Composite reward function. We design a citation-aware reward that combines multiple quality signals:

$$R = \sum_j w_j \cdot r_j \quad (5)$$

Table 3 details each component and its weight.

The hallucination penalty ($|w_{\text{hal}}| = 10.0$) is deliberately set as the highest weight to strongly discourage fabricated citations - cases where the model generates a citation marker [N] but N exceeds the number of provided knowledge passages.

Training protocol. We run 500 GRPO steps with group size $G = 4$, temperature $T = 0.7$ for diverse sampling, and a linear warmup schedule. The GRPO reward trajectory and KL divergence dynamics are shown in Figures 13, 14, and 17 (see Appendix).

3.7 Explainability Module

To provide interpretability into the generation process, we implement three complementary analysis methods:

- Cross-Attention Visualization.** For encoder-decoder models, we extract cross-attention weights between decoder output tokens and encoder input tokens, revealing which knowledge passage tokens the model attends to when generating each part of the response.
- Token Attribution via Integrated Gradients.** Following Sundararajan et al. [2017], we

Table 4: Dataset statistics. All splits maintain a balanced English–Hindi distribution.

Split	Total	EN	HI	EN%
Train	135,000	71,957	63,043	53.3
Val	7,500	4,014	3,486	53.5
Test	7,500	4,029	3,471	53.7

compute attribution scores for each input token by integrating gradients along a path from a zero embedding baseline to the actual input.

- Rationale Extraction.** We identify the minimal subset of knowledge passage tokens that are sufficient for the model’s prediction, following the approach of [Lei et al. \[2016\]](#).

The explainability module operates post-hoc on trained models and does not affect the training pipeline. Its primary purpose is to verify that citation markers in the generated response correspond to genuine attention over the cited knowledge passages - i.e., that the model’s citations are *faithful* to its internal reasoning.

4 Experimental Setup

4.1 Datasets

We construct a bilingual English–Hindi dataset by combining three established knowledge-grounded dialogue benchmarks and translating their English portions to Hindi using IndicTrans2 [[Gala et al., 2023](#)]. Table 4 summarizes the dataset statistics, and the data distribution across sources is visualized in Figure 8 (see Appendix).

Source corpora.

- **DSTC9** [[Kim et al., 2020](#)]: Task-oriented dialogues grounded in FAQ and review knowledge.
- **FaithDial** [[Rashkin et al., 2021](#)]: Faithful knowledge-grounded dialogues with hallucination annotations.
- **Wizard of Wikipedia** [[Dinan et al., 2019](#)]: Open-domain knowledge-grounded conversations.

Hindi translation. English examples are translated to Hindi using IndicTrans2 [[Gala et al., 2023](#)], which achieves state-of-the-art EN→HI translation quality. Citation markers ([1], [2]) are preserved during translation through regex-based pre/post-processing.

Table 5: Training hyperparameters. Effective batch size is held constant at 64 via gradient accumulation.

Param.	Base	Large	XL	Mistr.
<i>Stages 1–3: Supervised Fine-Tuning</i>				
LR	5e-5	3e-5	3e-5	1e-5
Batch	8	4	2	2
Grad.Acc.	4	8	16	16
Eff.Batch	64	64	64	64
Epochs	3	3	3	3
Max Seq.	512	512	512	1024
<i>Stage 4: GRPO Alignment</i>				
LR	5e-6	5e-6	5e-6	1e-6
Steps	500	500	500	500
G	4	4	4	4
β	0.04	0.04	0.04	0.04
Temp.	0.7	0.7	0.7	0.7

4.2 Training Configuration

Table 5 presents the hyperparameter configuration across stages and models. All training uses AdamW optimizer, BFloat16 mixed precision, and cosine learning rate scheduling with warmup. The learning rate schedules are visualized in Figure 9 (see Appendix).

Hardware. All experiments are conducted on a single NVIDIA A100 GPU (40GB). Models are trained sequentially in order of parameter count to ensure memory availability.

4.3 Evaluation Metrics

We evaluate along six dimensions, computing metrics separately for English and Hindi:

Lexical overlap.

- **BLEU** [[Papineni et al., 2002](#)]: n -gram precision with brevity penalty.
- **ROUGE-1 / ROUGE-L** [[Lin, 2004](#)]: Uni-gram and longest common subsequence recall.

Semantic similarity.

- **BERTScore** [[Zhang et al., 2020](#)]: Contextual embedding similarity using RoBERTa-Large.

Factual quality.

- **FactScore**: NLI-based factual consistency using DeBERTa [[Honovich et al., 2022](#)].
- **Hallucination Rate**: Fraction of responses with unsupported claims.

Citation quality.

- **Citation F1:** Harmonic mean of citation precision and recall.
- **Has Citation:** Fraction of responses with at least one citation marker.

4.4 Evaluation Protocol

Each of the six models is evaluated at five stages: baseline (pretrained, no fine-tuning), Stage 1, Stage 2, Stage 3, and Stage 4, yielding 30 total evaluation runs. For each run, we generate predictions for all test examples using greedy decoding (`num_beams=1`, `do_sample=False`) with `max_new_tokens=128`.

5 Results and Analysis

We report evaluation results for all six models - Flan-T5 Base, Large, and XL [Chung et al., 2024]; LLaMA-3.2-1B [Meta AI, 2024]; Gemma-2-2B [Google DeepMind, 2024]; and Mistral-7B [Jiang et al., 2023] - across all five training stages. The overall training loss progression is shown in Figure 10 and validation loss comparison in Figure 11 (see Appendix).

In the following tables, superscript arrows indicate the direction of change from the previous stage: \uparrow increased, \downarrow decreased, \sim negligible ($|\Delta| < 0.01$). Baseline rows show absolute values without arrows. For all metrics except Halluc., higher is better; for Halluc., lower is better.

5.1 Overall Progression

Table 6 presents overall metrics (combined English and Hindi) across all stages. The most striking pattern is the *phase transition* at Stage 2: all metrics jump dramatically, with hallucination dropping to exactly zero. The metric progression across all stages is visualized in Figure 1, and the per-model comparison in Figure 2.

5.2 English Results

Table 7 presents English-specific metrics. A remarkable finding is the *convergence* of Base and Large to nearly identical performance after Stage 2.

Key observations.

1. **SFT equalizes model sizes.** After Stage 2, Base and Large achieve identical English BLEU (0.172), Citation F1 (0.980), and BERTScore (0.889). This suggests that for well-defined structured tasks like citation-grounded dialogue, even a 250M model has sufficient capacity.
2. **Stage 1 preserves English.** All English metrics remain stable (± 0.002) after multilingual adaptation, confirming that translation training does not cause catastrophic forgetting of English capabilities.
3. **XL baseline has high Citation F1 (0.955).** Despite having lower BLEU (0.002), XL already understands citation formatting from pretraining. High citation format compliance with low content quality indicates that the 3B model knows the *form* but not the *substance* at baseline.
4. **LLaMA-1B achieves zero hallucination without citations.** From Stage 2 onward, LLaMA-3.2-1B reduces overall hallucination rate from 66.5% (Stage 1) to 0.9% while English Citation-F1 collapses to 0.000 and stays there. This demonstrates that hallucination elimination and citation format learning are separable objectives (see Section 5.5).
5. **Stage 1 can harm small decoder-only models.** LLaMA-1B’s overall hallucination rate increases from 13.5% (baseline) to 66.5% (Stage 1) - a $4.9\times$ increase - in sharp contrast to encoder-decoder models where Stage 1 preserves English stability. This suggests multilingual adaptation training may be ill-suited or requires different hyperparameters for small decoder-only architectures.

5.3 Hindi Results

Table 8 presents Hindi-specific metrics. The progressive pipeline shows clear cumulative benefit for Hindi, with each stage contributing measurably.

Key observations.

1. **Stage 3 is the Hindi game-changer.** For Base, Hindi ROUGE-1 jumps from 0.481 to 0.691 (+0.210) - the largest single-stage improvement in our study. This confirms that bilingual SFT with Hindi examples is essential.

Table 6: Overall evaluation results across all stages. Best result per model is **bolded**. “†” indicates generation collapse (empty outputs). Hallucination rate of 0.0 from Stage 2 onward is highlighted.

Model	Stage	BLEU	ROUGE-1	ROUGE-L	FactScore	Cit-F1	Halluc.	BERTScore
<i>Encoder-Decoder Models</i>								
Flan-T5-Base (250M)	Baseline	0.004	0.201	0.201	0.059	0.673	0.005	0.551
	Stage 1	0.005 ~	0.238 ↑	0.237 ↑	0.056 ~	0.738 ↑	0.028 ↑	0.611 ↑
	Stage 2	0.094 ↑	0.412 ↑	0.388 ↑	0.106 ↑	0.859 ↑	0.000 ↓	0.739 ↑
	Stage 3	0.092 ~	0.507 ↑	0.483 ↑	0.096 ~	0.902 ↑	0.000 ~	0.766 ↑
	Stage 4	0.092 ~	0.507 ~	0.483 ~	0.098 ~	0.902 ~	0.000 ~	0.766 ~
Flan-T5-Large (780M)	Baseline	0.003	0.304	0.303	0.073	0.801	0.078	0.731
	Stage 1	0.005 ~	0.306 ~	0.303 ~	0.106 ↑	0.758 ↓	0.090 ↑	0.734 ~
	Stage 2	0.093 ↑	0.468 ↑	0.445 ↑	0.085 ↓	0.901 ↑	0.000 ↓	0.769 ↑
	Stage 3	0.092 ~	0.499 ↑	0.476 ↑	0.084 ~	0.896 ~	0.000 ~	0.762 ~
	Stage 4	0.092 ~	0.498 ~	0.475 ~	0.084 ~	0.895 ~	0.000 ~	0.762 ~
Flan-T5-XL (3B)	Baseline	0.003	0.105	0.105	0.127	0.610	0.011	0.616
	Stage 1	0.002 ~	0.100 ~	0.099 ~	0.123 ~	0.588 ↓	0.011 ~	0.608 ~
	Stage 2 †				<i>Generation Collapse - empty outputs</i>			
	Stage 3	0.096 ↑	0.489 ↑	0.465 ↑	0.095 ↑	0.898 ↑	0.000 ↓	0.765 ↑
	Stage 4	0.096 ~	0.489 ~	0.466 ~	0.095 ~	0.898 ~	0.000 ~	0.765 ~
<i>Decoder-Only Models</i>								
LLaMA-3.2-1B (1B)	Baseline	0.028	0.193	0.163	0.247	0.424	0.135	0.758
	Stage 1	0.017 ↓	0.109 ↓	0.086 ↓	0.297 ↑	0.254 ↓	0.665 ↑	0.715 ↓
	Stage 2	0.019 ~	0.107 ~	0.090 ~	0.351 ↑	0.041 ↓	0.009 ↓	0.707 ↓
	Stage 3	0.052 ↑	0.376 ↑	0.357 ↑	0.401 ↑	0.362 ↑	0.014 ↑	0.774 ↑
	Stage 4	0.052 ~	0.371 ~	0.351 ~	0.393 ~	0.359 ~	0.014 ~	0.775 ~
Gemma-2-2B (2B)	Baseline	0.031	0.163	0.122	0.237	0.647	0.014	0.745
	Stage 1	0.029 ~	0.129 ↓	0.109 ↓	0.517 ↑	0.186 ↓	0.002 ↓	0.749 ~
	Stage 2	0.123 ↑	0.255 ↑	0.231 ↑	0.101 ↓	0.846 ↑	0.000 ↓	0.789 ↑
	Stage 3	0.187 ↑	0.559 ↑	0.533 ↑	0.226 ↑	0.903 ↑	0.000 ~	0.845 ↑
	Stage 4	0.187 ~	0.558 ~	0.532 ~	0.229 ~	0.903 ~	0.000 ~	0.845 ~
Mistral-7B (7B)	Baseline	0.023	0.122	0.086	0.140	0.544	0.078	0.750
	Stage 1	0.031 ~	0.142 ↑	0.109 ↑	0.289 ↑	0.659 ↑	0.120 ↑	0.743 ~
	Stage 2	0.051 ↑	0.159 ↑	0.136 ↑	0.203 ↓	0.707 ↑	0.010 ↓	0.749 ~
	Stage 3	0.094 ↑	0.344 ↑	0.319 ↑	0.151 ↓	0.768 ↑	0.014 ~	0.786 ↑
	Stage 4	0.095 ~	0.344 ~	0.319 ~	0.155 ~	0.772 ~	0.014 ~	0.787 ~

- Cross-lingual citation transfer.** Stage 2 (English-only SFT) improves Hindi Citation F1 from 0.485 to 0.718 for Base, demonstrating that citation formatting is a language-agnostic structural skill that transfers cross-lingually.
- Stage 1 is critical for small models.** Base Hindi BERTScore jumps from 0.221 to 0.351 (+0.130) after Stage 1, while Large shows minimal change (0.617 to 0.622). This confirms that multilingual adaptation primarily benefits models with limited pretraining Hindi exposure.
- Base overtakes Large after Stage 3.** An interesting crossover: Base achieves Hindi ROUGE-1 of 0.691 vs. Large’s 0.674, and Hindi Citation F1 of 0.812 vs. 0.798. The smaller model benefits more from bilingual SFT because it has more room to improve.
- LLaMA-1B Hindi citation learning surpasses Mistral-7B.** Despite failing entirely on English citations, LLaMA-3.2-1B achieves Hindi Citation-F1 of 0.783 at Stage 3 - exceeding Mistral-7B’s 0.522. The model appears to have allocated its limited citation learning capacity entirely to Hindi, the language that constituted 60% of Stage 3 training examples. This language-selective learning is discussed further in Section 5.5.
- Gemma-2-2B leads Hindi performance.** Gemma-2-2B achieves the strongest Hindi gains among decoder-only models: Hindi ROUGE-1 of 0.719 and Citation-F1 of 0.812 at Stage 4, comparable to the encoder-decoder models. This underscores that the progressive pipeline benefits decoder-only models as well, provided the model has sufficient capacity.

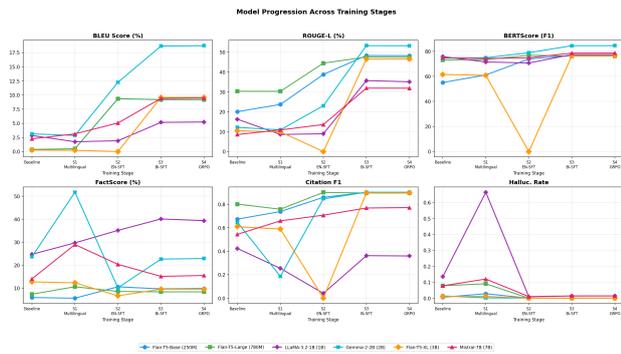


Figure 1: Training progression across all six evaluation metrics (BLEU, ROUGE-L, BERTScore, FactScore, Citation F1, Hallucination Rate) for all models. Note the XL generation collapse at Stage 2 and subsequent recovery at Stage 3.

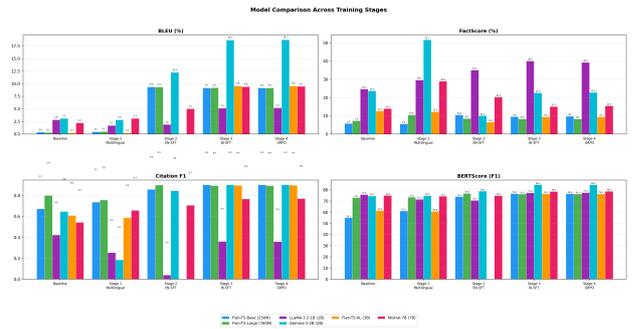


Figure 2: Model comparison across training stages for four key metrics (BLEU, FactScore, Citation F1, BERTScore). The phase transition at Stage 2 is clearly visible across all metrics and models.

Per-Language BLEU Progression

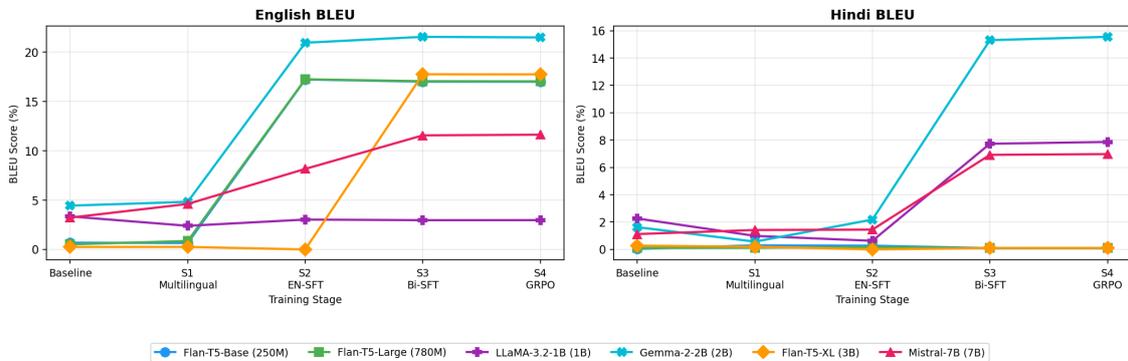


Figure 3: Per-language BLEU progression across training stages. English BLEU (left) shows convergence of all models after Stage 2 SFT, while Hindi BLEU (right) remains near-zero for encoder-decoder models due to morphological variation (Section 6.3). Gemma-2-2B achieves the highest Hindi BLEU (0.155) after Stage 3, followed by Mistral-7B (0.070).

5.4 Flan-T5-XL Generation Collapse

Flan-T5-XL (3B) exhibited *generation collapse* after Stage 2 SFT: the model produced empty strings for all test examples, yielding zero scores across all metrics. However, the model **recovered fully in Stage 3** (Bilingual SFT), achieving performance comparable to Base and Large (overall BLEU 0.096, Citation F1 0.898, BERTScore 0.765).

Diagnosis. The collapse is notable because the Stage 2 checkpoint appeared healthy by standard indicators: (i) teacher-forced evaluation loss was 1.642 (comparable to Base at 1.916 and Large at 1.776), (ii) model weights were intact (5.4 GB across two safetensor shards), and (iii) the generation configuration was correct (standard T5 defaults).

Root cause. We attribute the collapse to the learning rate of 2×10^{-5} , which is too aggressive for a 3B encoder-decoder model. The successful models used proportionally lower rates: Base (250M) at 5×10^{-5} and Large (780M) at 3×10^{-5} . At 3B parameters, the loss landscape is sharper, and the learning rate pushed the model into a degenerate local minimum where emitting the EOS token as the first generated token minimizes the autoregressive loss.

Recovery in Stage 3. The bilingual SFT in Stage 3 effectively rescued the model from collapse. Despite building on the collapsed Stage 2 checkpoint, the continued training with a mixture of English and Hindi dialogue data pushed the model out of the degenerate minimum. After recovery, XL achieves English metrics (BLEU 0.177, Citation F1 0.980, BERTScore 0.891) comparable to Base and Large, and strong Hindi per-

Table 7: English evaluation results. After Stage 2 SFT, Base (250M) and Large (780M) converge to identical performance across all metrics. Citation F1 reaches 0.980 for both models.

Model	Stage	BLEU	ROUGE-1	ROUGE-L	FactScore	Cit-F1	Halluc.	BERTScore	
<i>Encoder-Decoder Models</i>									
Base (250M)	Baseline	0.007	0.100	0.100	0.039	0.960	0.005	0.836	
	Stage 1	0.007~	0.101~	0.101~	0.040~	0.956~	0.004~	0.835~	
	Stage 2	0.172 ↑	0.354 ↑	0.308 ↑	0.121↑	0.980 ↑	0.000 ↓	0.889 ↑	
	Stage 3	0.170~	0.349~	0.304~	0.120~	0.980 ~	0.000 ~	0.889 ~	
	Stage 4	0.170~	0.349~	0.304~	0.124 ~	0.980 ~	0.000 ~	0.889 ~	
Large (780M)	Baseline	0.005	0.100	0.098	0.075	0.866	0.079	0.829	
	Stage 1	0.009~	0.107~	0.102~	0.132 ↑	0.782↓	0.102↑	0.830~	
	Stage 2	0.172 ↑	0.351 ↑	0.307 ↑	0.093↓	0.980 ↑	0.000 ↓	0.889 ↑	
	Stage 3	0.171~	0.348~	0.304~	0.096~	0.980 ~	0.000 ~	0.889 ~	
	Stage 4	0.170~	0.348~	0.304~	0.096~	0.980 ~	0.000 ~	0.889 ~	
XL (3B)	Baseline	0.002	0.097	0.097	0.044	0.955	0.016	0.828	
	Stage 1	0.003~	0.098~	0.097~	0.045~	0.946~	0.019~	0.828~	
	Stage 2†				<i>Generation Collapse - empty outputs</i>				
	Stage 3	0.177 ↑	0.363 ↑	0.318 ↑	0.113 ↑	0.980 ↑	0.000 ↓	0.891 ↑	
	Stage 4	0.177 ~	0.363 ~	0.318 ~	0.113 ~	0.980 ~	0.000 ~	0.891 ~	
<i>Decoder-Only Models</i>									
LLaMA-3.2-1B (1B)	Baseline	0.033	0.189	0.134	0.153	0.693	0.160	0.842	
	Stage 1	0.024↓	0.157↓	0.116↓	0.227↑	0.287↓	0.810↑	0.821↓	
	Stage 2	0.030↑	0.162~	0.130↑	0.288 ↑	0.000↓	0.000 ↓	0.830↑	
	Stage 3	0.030~	0.163~	0.129~	0.268↓	0.000~	0.000 ~	0.831~	
	Stage 4	0.030~	0.165~	0.129~	0.267~	0.000~	0.000 ~	0.831~	
Gemma-2-2B (2B)	Baseline	0.044	0.227	0.153	0.211	0.781	0.023	0.850	
	Stage 1	0.048~	0.223~	0.185↑	0.537 ↑	0.218↓	0.001↓	0.855~	
	Stage 2	0.209↑	0.415↑	0.370↑	0.096↓	0.980 ↑	0.000 ↓	0.900↑	
	Stage 3	0.215 ↑	0.421 ↑	0.374 ↑	0.110↑	0.980 ~	0.000 ~	0.901 ↑	
	Stage 4	0.215 ~	0.420~	0.373~	0.110~	0.980 ~	0.000 ~	0.901 ~	
Mistral-7B (7B)	Baseline	0.032	0.203	0.135	0.124	0.912	0.042	0.848	
	Stage 1	0.046↑	0.227↑	0.165↑	0.247 ↑	0.926↑	0.096↑	0.841~	
	Stage 2	0.082↑	0.258↑	0.215↑	0.139↓	0.980 ↑	0.000 ↓	0.856↑	
	Stage 3	0.116 ↑	0.301 ↑	0.256 ↑	0.070↓	0.980 ~	0.001~	0.861~	
	Stage 4	0.116 ~	0.301 ~	0.255~	0.073~	0.979~	0.002~	0.862 ~	

formance (ROUGE-1 0.636, Citation F1 0.802). This demonstrates the **resilience of continued training**: a collapsed checkpoint is not necessarily unrecoverable.

Implications. This episode highlights a critical lesson: **teacher-forced evaluation loss is not a reliable proxy for generation quality**. A model can achieve reasonable teacher-forced loss while completely failing at autoregressive generation. Training pipelines for generative models should include periodic generation-time validation. This phenomenon is consistent with findings on neural text degeneration [Holtzman et al., 2020] and unlikelihood training [Welleck et al., 2020].

5.5 LLaMA-3.2-1B: Language-Selective Citation Failure

LLaMA-3.2-1B exhibits a qualitatively distinct failure mode from Flan-T5-XL’s generation collapse: *language-selective citation failure*. From Stage 2 onward, the model generates zero citation markers in English (English Citation-F1 = 0.000, 0% of responses contain citations), while simultaneously learning robust Hindi citation behavior by Stage 3 (Hindi Citation-F1 = 0.783, citations present in 94.2% of responses).

Stage 1 hallucination explosion. Unlike all encoder-decoder models, LLaMA-1B experiences a catastrophic hallucination spike after multilingual adaptation: overall hallucination rate rises from 13.5% at baseline to 66.5% after Stage 1 (English: 16.0% → 81.0%). This suggests that

Table 8: Hindi evaluation results. Stage 1 builds the foundation, Stage 2 transfers citation skills, and Stage 3 provides the largest Hindi improvement. Note: Hindi BLEU is near-zero due to morphological variation (see Section 6).

Model	Stage	BLEU	ROUGE-1	ROUGE-L	FactScore	Cit-F1	Halluc.	BERTScore
<i>Encoder-Decoder Models</i>								
Base (250M)	Baseline	0.000	0.318	0.318	0.082	0.340	0.005	0.221
	Stage 1	0.003 ~	0.396 ↑	0.395 ↑	0.075 ~	0.485 ↑	0.056 ↑	0.351 ↑
	Stage 2	0.003 ~	0.481 ↑	0.481 ↑	0.088 ↑	0.718 ↑	0.000 ↓	0.565 ↑
	Stage 3	0.001 ~	0.691 ↑	0.691 ↑	0.069 ↓	0.812 ↑	0.000 ~	0.624 ↑
	Stage 4	0.001 ~	0.690 ~	0.690 ~	0.069 ~	0.811 ~	0.000 ~	0.623 ~
Large (780M)	Baseline	0.001	0.542	0.542	0.071	0.726	0.077	0.617
	Stage 1	0.001 ~	0.536 ~	0.536 ~	0.075 ~	0.729 ~	0.077 ~	0.622 ~
	Stage 2	0.002 ~	0.605 ↑	0.605 ↑	0.077 ~	0.809 ↑	0.000 ↓	0.629 ~
	Stage 3	0.001 ~	0.674 ↑	0.674 ↑	0.070 ~	0.798 ↓	0.000 ~	0.615 ↓
	Stage 4	0.001 ~	0.673 ~	0.673 ~	0.070 ~	0.797 ~	0.000 ~	0.615 ~
XL (3B)	Baseline	0.003	0.114	0.114	0.224	0.209	0.004	0.369
	Stage 1	0.002 ~	0.102 ↓	0.101 ↓	0.213 ↓	0.173 ↓	0.001 ~	0.353 ↓
	Stage 2 [†]			<i>Generation Collapse - empty outputs</i>				
	Stage 3	0.001 ↑	0.636 ↑	0.636 ↑	0.074 ↑	0.802 ↑	0.001 ~	0.619 ↑
	Stage 4	0.001 ~	0.637 ~	0.637 ~	0.074 ~	0.803 ~	0.001 ~	0.619 ~
<i>Decoder-Only Models</i>								
LLaMA-3.2-1B (1B)	Baseline	0.023	0.198	0.197	0.356	0.111	0.107	0.661
	Stage 1	0.010 ↓	0.053 ↓	0.052 ↓	0.377 ↑	0.216 ↑	0.497 ↑	0.592 ↓
	Stage 2	0.006 ↓	0.044 ↓	0.043 ↓	0.424 ↑	0.089 ↓	0.019 ↓	0.565 ↓
	Stage 3	0.077 ↑	0.624 ↑	0.622 ↑	0.555 ↑	0.783 ↑	0.030 ↑	0.708 ↑
	Stage 4	0.079 ~	0.611 ~	0.609 ~	0.540 ~	0.777 ~	0.031 ~	0.709 ~
Gemma-2-2B (2B)	Baseline	0.016	0.088	0.086	0.267	0.490	0.004	0.624
	Stage 1	0.006 ↓	0.021 ↓	0.021 ↓	0.494 ↑	0.148 ↓	0.003 ~	0.626 ~
	Stage 2	0.022 ↑	0.070 ↑	0.070 ↑	0.107 ↓	0.691 ↑	0.001 ~	0.660 ↑
	Stage 3	0.153 ↑	0.718 ↑	0.717 ↑	0.361 ↑	0.812 ↑	0.000 ↓	0.780 ↑
	Stage 4	0.155 ~	0.719 ~	0.717 ~	0.366 ~	0.812 ~	0.000 ~	0.781 ~
Mistral-7B (7B)	Baseline	0.011	0.028	0.028	0.159	0.116	0.119	0.637
	Stage 1	0.014 ~	0.044 ↑	0.044 ↑	0.338 ↑	0.349 ↑	0.148 ↑	0.629 ~
	Stage 2	0.014 ~	0.044 ~	0.044 ~	0.278 ↓	0.390 ↑	0.021 ↓	0.624 ~
	Stage 3	0.069 ↑	0.393 ↑	0.392 ↑	0.245 ↓	0.522 ↑	0.028 ~	0.699 ↑
	Stage 4	0.070 ~	0.393 ~	0.392 ~	0.249 ~	0.531 ~	0.028 ~	0.700 ~

Stage 1 translation-style training is particularly disruptive for small decoder-only models.

Stage 2: Hallucination eliminated without citation format. Stage 2 SFT reduces English hallucination from 81.0% to 0.0% - a dramatic recovery. However, English Citation-F1 collapses simultaneously from 0.287 (Stage 1) to 0.000, where it remains through Stages 3 and 4. The model learned a conservative generation strategy - avoid all specific claims and thus avoid hallucination - rather than grounding claims in cited passages. This demonstrates that **zero hallucination and zero citation are not contradictory**: a model can satisfy the former without the latter by generating only generic, non-committal text.

Stage 3: Hindi citation learning without English recovery. The 60%/40% Hindi-English bilingual SFT in Stage 3 produces a striking asymmetry: Hindi Citation-F1 improves from 0.089 to 0.783 (surpassing Mistral-7B’s 0.522), while English Citation-F1 remains 0.000. The model successfully learned citation format from Hindi training examples but neither generalised this behaviour to English nor recovered the English citation pathway lost in Stage 2. This is consistent with catastrophic forgetting in sequential fine-tuning [McCloskey and Cohen, 1989]: Stage 2 may have irreversibly overwritten English citation associations, and the Hindi-weighted Stage 3 did not re-establish them.

Qualitative evidence. To verify that the zero English Citation-F1 is not a metric format mis-

match (e.g., model generating (1) instead of [1]), we ran direct inference on the Stage 3 checkpoint against the citation test set. English outputs are citation-free and show repetition degeneration:

“No, you cannot bring your own liquor at Pizza Hut Cherry Hinton. Do you have any other questions? Would you like to make a reservation? If so, for what time and day? How many people will be dining? How many people? How many people?”

The expected output was: *“As noted in [1], No, outside beverages are not allowed.”*. No citation marker appears in any English output across tested samples. Hindi outputs, by contrast, correctly produce citation markers embedded in fluent responses (e.g., [2] followed by citation-grounded Hindi text, translating to *“Based on [2], I have booked you a taxi...”* [Li et al., 2024]). This confirms the citation failure is behavioural - the model architecture is capable of producing [N] markers but does not do so for English.

Stage 4 GRPO: No recovery. GRPO training changes all metrics by at most ± 0.008 (Table 9), confirming that RL alignment cannot recover citation behaviour that SFT failed to instil.

Implications. LLaMA-3.2-1B’s behaviour demonstrates two separable properties: (i) hallucination elimination and citation format learning are independent objectives - a model can achieve 0% hallucination without generating any citation markers; and (ii) catastrophic forgetting of citation format can occur even within a carefully staged pipeline. Importantly, Gemma-2-2B (a 2B decoder-only model) achieves Citation-F1 of 0.903 overall and 0.980 on English from Stage 2 onward - demonstrating that citation format learning is achievable in decoder-only models at this parameter scale. This confirms that LLaMA-3.2-1B’s citation failure is a **model-specific anomaly** rather than a general architectural limitation of decoder-only models, plausibly driven by its smaller effective capacity, distinct pretraining corpus, or instruction-tuning protocol.

5.6 Stage 4 (GRPO) Effectiveness

A central question of our study is whether GRPO alignment provides measurable improvement over SFT. Table 9 presents the Stage 3 \rightarrow Stage 4 deltas.

Table 9: Change from Stage 3 to Stage 4 (GRPO). Deltas are negligible for encoder-decoder models. Mistral shows marginal gains in Citation F1 and FactScore.

	Metric	Stage 3	Δ
<i>Encoder-Decoder</i>			
Base	Cit-F1	0.902	0.000
	Halluc.	0.000	0.000
	BERTScore	0.766	0.000
	FactScore	0.096	+0.002
Large	Cit-F1	0.896	-0.001
	Halluc.	0.000	0.000
	BERTScore	0.762	0.000
	FactScore	0.084	0.000
XL	Cit-F1	0.898	0.000
	Halluc.	0.000	0.000
	BERTScore	0.765	0.000
	FactScore	0.095	0.000
<i>Decoder-Only</i>			
LLaMA	Cit-F1	0.362	-0.003
	Halluc.	0.014	0.000
	BERTScore	0.774	+0.001
	FactScore	0.401	-0.008
Gemma	Cit-F1	0.903	0.000
	Halluc.	0.000	0.000
	BERTScore	0.845	0.000
	FactScore	0.226	+0.003
Mistral	Cit-F1	0.768	+0.004
	Halluc.	0.014	0.000
	BERTScore	0.786	+0.001
	FactScore	0.151	+0.004

GRPO adds essentially zero improvement across all metrics for both models. We analyze the GRPO training dynamics in Table 10 to understand why. The full reward trajectory is visualized in Figure 4, with reward distributions in Figure 15 and best-vs-final comparison in Figure 16 (see Appendix).

Analysis. The reward declines from best to final for most models (Figure 16), suggesting that GRPO training is unstable in this setting. The exception is Gemma-2-2B, which achieves the highest absolute reward (3.40) with zero decline - its best and final rewards are identical, indicating improvement continued to the last training step. Among models that show reward regression, Mistral-7B achieves the highest absolute reward (2.889) and the smallest decline (-0.457). We discuss possible causes in Section 6.

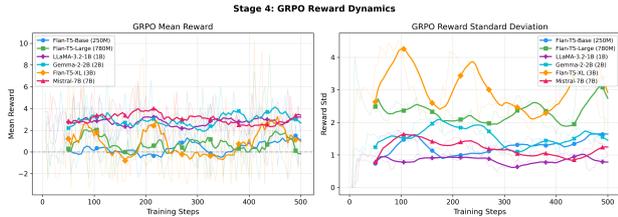


Figure 4: GRPO reward dynamics during Stage 4. Left: mean reward with smoothed curves showing Mistral-7B consistently achieving the highest reward. Right: reward standard deviation, indicating training stability.

Table 10: GRPO reward trajectory during Stage 4 training. Encoder-decoder and decoder-only models (LLaMA, Mistral) show reward decline from best to final, suggesting training instability. Gemma-2-2B is the exception: its best reward equals its final reward, indicating reward improvement continued to the last step.

	Model	Best	Final	Δ
Enc-Dec	Base (250M)	1.040	0.648	-0.391 ↓
	Large (780M)	0.868	0.157	-0.711 ↓
	XL (3B)	1.782	0.482	-1.300 ↓
Dec-Only	LLaMA (1B)	3.08	2.51	-0.57 ↓
	Gemma (2B)	3.40	3.40	0.00
	Mistral (7B)	2.889	2.433	-0.457 ↓

5.7 Training Efficiency

Table 11 summarizes training steps across all stages. The complete training loss curves are presented in Figures 10 and 11 (see Appendix).

Mistral-7B achieves substantially lower loss across all stages (e.g., Stage 3 loss of 0.303 vs. XL’s 0.895), suggesting that the decoder-only 7B model has significantly more capacity for this task. The Stage 2 \rightarrow Stage 3 loss drop is consistent across all models, confirming that bilingual SFT improves the model’s fit to the dialogue distribution.

5.8 Explainability Analysis

We conducted three complementary explainability analyses on representative test examples: (i) attention alignment, measuring cross-attention focus on cited knowledge tokens; (ii) gradient saliency, quantifying input token attribution; and (iii) occlusion sensitivity, testing whether citations are causally grounded in their sources.

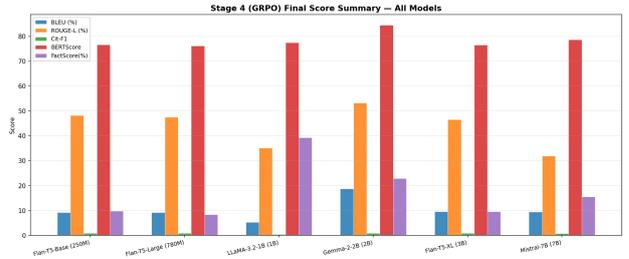


Figure 5: Stage 4 (GRPO) final score summary across all six models, showing five metrics side by side. Gemma-2-2B and Mistral-7B lead in most metrics after GRPO alignment.

Table 11: Training summary showing total steps and best validation loss per stage. Mistral-7B consistently achieves the lowest loss across all stages.

	Model	Stage	Steps	Best Loss
Encoder-Decoder	Base	Stage 1	3,906	1.552
		Stage 2	7,500	1.916
		Stage 3	4,000	0.945
	Large	Stage 1	3,906	1.439
		Stage 2	8,973	1.776
		Stage 3	4,000	0.912
XL	Stage 1	7,812	1.384	
	Stage 2	14,000	1.642	
	Stage 3	6,000	0.895	
Decoder-Only	LLaMA	Stage 1	7,812	1.408
		Stage 2	7,500	1.930
		Stage 3	12,000	0.912
	Gemma	Stage 1	7,812	1.458
		Stage 2	6,000	1.104
		Stage 3	9,500	0.651
	Mistral	Stage 1	7,812	0.742
		Stage 2	7,000	0.927
		Stage 3	9,500	0.303

Attention alignment. For encoder-decoder models (Flan-T5 Base, Large, XL), we measure the mean cross-attention weight assigned to cited knowledge tokens during generation. Decoder-only models (Gemma, LLaMA, Mistral) lack a distinct cross-attention mechanism and yield 0.000 by construction (see Section 6.7). Table 12 summarises per-model, per-stage alignment scores.

All three Flan-T5 variants show peak alignment at Stage 2 (English SFT), the stage with the most dramatic citation metric improvement. Flan-T5-XL’s Stage 2 alignment of 0.000 corroborates the generation collapse detected by evaluation metrics. After recovery at Stage 3, XL’s alignment (0.026) exceeds Base (0.019) and Large (0.022),

causal grounding after fine-tuning. Mistral-7B drops from 0.767 (baseline) to 0.000 at Stage 3–4 despite Citation-F1 of 0.772, and Gemma-2-2B drops from 0.733 (baseline) to 0.000 from Stage 1 onward despite Citation-F1 of 0.903 at Stage 3. In both cases, the model learns to produce citation *format* (marker tokens) without grounding the cited content in the provided knowledge source. This decoder-only format-grounding dissociation contrasts sharply with encoder-decoder models, where cross-attention provides a structural mechanism for source conditioning that persists through fine-tuning.

6 Discussion

We organize our discussion around seven key findings and their broader implications for knowledge-grounded dialogue research.

6.1 Citation-Grounded SFT as Anti-Hallucination

Under automatic NLI-based evaluation, hallucination rate drops to 0.0% after Stage 2 for encoder-decoder models (Flan-T5 Base and Large) and remains there through Stages 3 and 4. For Mistral-7B (decoder-only), hallucination drops from 0.078 to 0.010 after Stage 2, stabilizing at 0.014 after Stage 3. This architecture-dependent pattern suggests that encoder-decoder models, with their explicit cross-attention over knowledge passages, may be more amenable to hallucination reduction through citation-grounded SFT. We note that these results are based on automatic metrics; human evaluation may reveal a different picture and remains an important avenue for future work.

The likely mechanism is that when every training example contains properly cited responses of the form “According to [1], ...”, the model learns to associate claims with numbered references. At inference time, this encourages the model to ground claims in provided passages rather than generating unsupported content. Whether this transfers fully to real-world deployment conditions warrants further study.

LLaMA-1B counterexample: hallucination without citation grounding. LLaMA-3.2-1B’s results challenge the assumption that citation-grounded SFT is the *mechanism* by which hallucination is reduced. From Stage 2 onward, LLaMA-1B achieves 0.0% English hal-

lucination - yet its English Citation-F1 is simultaneously 0.000 (no citation markers generated). The model suppresses hallucination not by citing passages but by adopting a conservative, non-committal generation strategy that avoids specific factual claims altogether. This demonstrates that **zero hallucination and zero citation grounding are not contradictory**, and that the causal path from SFT to hallucination reduction may be model-specific.

These observations suggest that citation-grounded SFT may be a promising direction for reducing hallucination, potentially complementing post-hoc detection approaches. We encourage future work to validate these findings with human evaluation across diverse domains.

6.2 SFT Sufficiency for Structured Tasks

The marginal gains observed after Stage 4 GRPO (Section 5.6) invite a broader question: *under what conditions does RL alignment provide meaningful benefit over well-tuned SFT?*

We offer one possible interpretation: citation-grounded dialogue has a well-defined output format with relatively clear quality criteria. SFT directly optimises for this format by maximising likelihood of reference outputs. When the reference outputs already satisfy quality criteria, the RL signal may have limited additional structure to exploit. We note, however, that our experimental configuration represents one point in a large hyperparameter space, and we do not claim this as a general conclusion.

We identify four factors that may have limited GRPO’s contribution in our specific setting:

1. **Possible reward saturation.** Citation quality and hallucination metrics were near-optimal after SFT, which may have reduced the reward signal available for RL to exploit.
2. **KL penalty strength.** The coefficient $\beta = 0.04$ constrains the policy to stay close to the SFT checkpoint; a lower β may allow more beneficial exploration.
3. **Reward signal granularity.** The citation component is largely binary, which may provide limited gradient signal when the model already cites consistently.
4. **Training budget.** At 500 steps with group size 4, only 2,000 candidate responses are ex-

plored; longer training may yield different outcomes.

Notably, both Gemma-2-2B (3.40) and Mistral-7B (2.889) achieve substantially higher absolute rewards than encoder-decoder models (0.48–1.78), suggesting that RL behaviour may be scale- and architecture-dependent. Gemma-2-2B’s zero reward decline is particularly noteworthy: it may indicate that a longer GRPO training budget would yield further improvement for this model. Systematic hyperparameter exploration - varying β , training steps, group size, and reward design - remains an important direction for future work.

6.3 The Hindi BLEU Problem

Hindi BLEU remains near-zero for encoder-decoder models (≤ 0.003) across all stages, even when ROUGE-1 (up to 0.691) and BERTScore (up to 0.624) indicate strong Hindi generation quality. Decoder-only models achieve somewhat higher but still modest BLEU values (LLaMA-3.2-1B: 0.079, Gemma-2-2B: 0.155), yet these too substantially underestimate quality relative to ROUGE-1 (0.624, 0.719) and BERTScore (0.709, 0.781) respectively. This is not a model failure but a **metric limitation** (see Figure 3).

BLEU [Papineni et al., 2002] measures exact n -gram precision, which is inherently biased against morphologically rich languages. Hindi exhibits: (a) productive morphological inflection, (b) relatively free word order, and (c) multiple valid surface realizations. This observation is consistent with Post [2018]. For Hindi evaluation, we recommend **BERTScore as the primary metric**, with ROUGE-1 as a secondary signal.

6.4 Contextualising Absolute Performance

To place our results in context, we compare against FaithDial [Rashkin et al., 2021], the most closely related benchmark for faithful knowledge-grounded dialogue evaluation. The best-performing FaithDial model (T5-CTRL trained on FaithDial+WoW) reports BLEU of 0.115 and BERTScore of 0.835. Our Stage 3 encoder-decoder models achieve BERTScore of 0.762–0.891, matching or exceeding FaithDial on semantic quality, while Mistral-7B reaches 0.787. Our BLEU (0.092–0.095) is slightly below FaithDial’s 0.115, which we attribute partly to the structured citation constraint: our responses contain explicit

citation markers ([1], [2]) that are absent from free-form FaithDial references, reducing exact n -gram overlap by design. Importantly, FaithDial does not evaluate citation accuracy or attribution - tasks that our pipeline additionally addresses. Direct cross-paper comparison should be interpreted cautiously given differences in test sets, knowledge sources, and evaluation protocols. We treat BERTScore as the primary semantic quality metric and Citation-F1 with hallucination rate as the primary grounding metrics, with BLEU serving as a secondary indicator.

6.5 Scaling Behavior and Architecture Effects

Our results reveal nuanced interactions between model size, architecture, and training stage:

English performance saturates early. After Stage 2, both Base (250M) and Large (780M) achieve identical English metrics (BLEU 0.172, BERTScore 0.889, Citation F1 0.980). This suggests that for constrained, well-defined tasks, model capacity beyond 250M provides no English benefit. This has significant deployment implications: a 250M model is $3\times$ cheaper to serve than 780M with no quality loss.

Hindi benefits from scale at baseline. Large starts with Hindi BERTScore of 0.617 vs. Base’s 0.221 - a $2.8\times$ gap attributable to more Hindi data in the larger pretraining corpus. However, this gap narrows through training: after Stage 3, Base reaches 0.624 vs. Large’s 0.615, effectively closing the gap. This demonstrates that our progressive pipeline can compensate for limited pretraining Hindi exposure.

Generation collapse is recoverable. Flan-T5-XL (3B, encoder-decoder) collapsed at Stage 2 under a learning rate of 2×10^{-5} , but recovered fully at Stage 3 (Bilingual SFT), achieving performance comparable to Base and Large. This demonstrates that continued training can rescue a collapsed checkpoint, and that encoder-decoder models at scale require more careful learning rate tuning - Mistral-7B trains with 1×10^{-5} without collapse.

Decoder-only vs. encoder-decoder trade-offs. Mistral-7B achieves higher absolute BLEU than encoder-decoder models at baseline (0.023

vs. 0.003–0.004) and maintains a lead in FactScore throughout training (0.155 at Stage 4 vs. 0.095–0.098 for Flan-T5). However, Mistral does not eliminate hallucination entirely (0.014 at Stage 3/4 vs. 0.000 for Flan-T5 Base/Large), and its Hindi Citation F1 (0.531) lags behind the encoder-decoder models (0.797–0.803). This suggests that encoder-decoder architectures are more effective at learning the citation grounding constraint.

Heterogeneous decoder-only behavior.

The decoder-only models in our study show strikingly different outcomes. Mistral-7B (7B) successfully learns English citation format (Citation-F1 0.979 at Stage 4), while LLaMA-3.2-1B (1B) fails entirely (Citation-F1 0.000 from Stage 2 onward). Gemma-2-2B (2B), however, achieves Citation-F1 of 0.980 on English and 0.812 on Hindi by Stage 3 - demonstrating that the failure is LLaMA-specific rather than a scale threshold effect: a 2B decoder-only model *can* learn bilingual citation format under our pipeline. This indicates that **citation learning failure is not an architectural property of decoder-only models** and is not simply a function of parameter count; it is plausibly driven by LLaMA-3.2-1B’s specific pretraining corpus composition, instruction-tuning protocol, or optimizer state.

Larger models hallucinate more before grounding. Large (780M) has a baseline hallucination rate of 0.078, Mistral-7B 0.078, vs. Base’s 0.005. Larger models generate more fluent, confident text, which manifests as more plausible-sounding but unsupported claims. This makes knowledge grounding training *more important* for larger models, not less.

6.6 Citation Phrase Salience and Language Module Strength

Output inspection of LLaMA-3.2-1B Stage 3 (Section 5.5) enables a hypothesis about why citation learning succeeds in Hindi but fails in English. We propose that citation learnability is jointly determined by two factors:

Citation phrase salience. Hindi citation phrases such as “[1] में कहा गया है” (it is said in [1]) or “[2] के आधार पर” (based on [2]) are formulaic, high-salience constructions - the citation marker and

its framing phrase form an inseparable grammatical unit that stands out as a distinctive pattern during training. English citation phrases (“According to [1]”, “As noted in [1]”) are semantically richer but also more varied and closer to non-cited paraphrase phrasing, making them lower-salience patterns.

Pretrained language module strength.

LLaMA-3.2-1B’s English baseline BERTScore (0.842) substantially exceeds its Hindi BERTScore (0.661), reflecting stronger English pretraining. We hypothesize that strong pretrained English fluency creates a resistance to the citation pattern: during Stage 2 SFT (training loss 1.930 > Stage 1 loss 1.408), the gradient signal from citation learning was insufficient to overcome the inertia of English fluency representations. Hindi, where the pretrained module is weaker, offered less resistance to learning the new citation pattern during Stage 3.

Implication. If this hypothesis holds, it suggests a design principle for citation format selection: citation markers should be maximally distinctive from natural-language paraphrase patterns, especially when fine-tuning strong pretrained models. Alternatives such as <cite:1> or [SOURCE:1] may be more learnable for high-resource languages than the conventional [1] notation. We acknowledge this remains a hypothesis requiring controlled ablation studies (e.g., varying citation format style while holding all else constant) to verify.

6.7 Interpretability of Citation Behaviour

The explainability analyses (Section 5.8) reveal qualitatively distinct mechanisms of citation grounding across architectures.

Encoder-decoder cross-attention as citation grounding.

Flan-T5 models possess a cross-attention sublayer that directly attends to encoder representations of the knowledge passage during decoding. Our attention alignment scores (Table 12) confirm that cited tokens receive disproportionate cross-attention weight: mean alignment increases from 0.017 at baseline to a peak of 0.035–0.037 at Stage 2, the stage where Citation-F1 shows its most dramatic jump. The Stage 2 SFT teaches the decoder to route attention toward cited passage segments at the moment of

generating citation markers - a direct, mechanistically interpretable grounding signal.

Decoder-only models: absence of cross-attention. Decoder-only architectures (Gemma-2-2B, LLaMA-3.2-1B, Mistral-7B) process the full prompt (query + knowledge passages) as a single causal sequence. There is no dedicated cross-attention sublayer; knowledge is accessed through the same self-attention mechanism as all other context. Cross-attention alignment is therefore architecturally 0.000 for these models - a measurement constraint, not a failure. These models must learn to retrieve and cite knowledge using self-attention alone, a fundamentally different mechanism that requires the model to internally route long-range attention from the response token to the relevant source token positions.

Saliency as a training quality signal. The monotonic decrease in saliency concentration across training stages (Table 13) suggests that well-trained models distribute credit more evenly across input tokens, consistent with broader contextual integration. Importantly, Flan-T5-XL’s Stage 2 NaN saliency provides an *independent* confirmation of generation collapse that complements the zero-score evaluation result: no gradient signal flows when the model generates empty outputs, making saliency undefined.

Decoder-only format-grounding dissociation. A consistent architecture-level pattern emerges from occlusion analysis: *both* decoder-only models exhibit 0.000 causal grounding after fine-tuning. Mistral-7B drops from 0.767 to 0.000 at Stage 3–4 despite Citation-F1 of 0.772, and Gemma-2-2B drops from 0.733 to 0.000 from Stage 1 onward despite Citation-F1 of 0.903 at Stage 3 (Table 14). This dissociation - citation format present but causal grounding absent - indicates that decoder-only models learn to produce the syntactic pattern of citation (inserting [N] markers at contextually appropriate positions) without the semantic grounding of citation content (claims genuinely conditioned on the removed source). The model may generate correct citation *numbers* based on positional or conversational context rather than genuinely consulting source content. That this pattern holds across two architecturally distinct decoder-only models (Mistral’s sliding-window attention vs. Gemma’s

grouped-query attention) suggests it is a general property of causal language models fine-tuned on citation tasks, not an artefact of a single architecture. This finding strengthens the argument that Citation-F1 alone is an insufficient grounding metric and must be complemented by occlusion-based causal grounding tests.

6.8 Limitations

We acknowledge several limitations:

1. **Hindi data quality.** Hindi examples are machine-translated, which may not fully reflect natural conversational patterns.
2. **Single-reference evaluation.** All automatic metrics compare against a single reference response.
3. **GRPO exploration.** Comprehensive hyperparameter search might reveal configurations where GRPO provides measurable improvement.
4. **Human evaluation.** This study relies entirely on automatic metrics.
5. **Explainability scope.** Attention alignment analysis is limited to encoder-decoder models with cross-attention; comparable interpretability tools for decoder-only models (e.g., causal attention attribution) were not included.

6.9 Dataset-Specific vs. Merged Training

Our current study merges three dialogue datasets (DSTC9, Wizard of Wikipedia, FaithDial) to create a unified training corpus. While this approach demonstrates that progressive training is effective across dialogue types, it may obscure important task-specific differences in citation learning dynamics:

- **Task-oriented dialogue (DSTC9)** focuses on FAQ-style knowledge with specific, goal-directed citations.
- **Open-domain dialogue (WoW)** uses Wikipedia articles for exploratory, conversational citations.
- **Hallucination-aware dialogue (FaithDial)** emphasises cautious, verifiable citations.

Future work should investigate whether citation learning strategies differ across these dialogue types by training models separately on each

dataset. Additionally, separate English and Hindi training tracks (rather than bilingual mixed training) would isolate language-specific effects from cross-lingual transfer effects.

7 Conclusion and Future Work

We presented XKD-DIAL, a progressive four-stage training pipeline for explainable, knowledge-grounded dialogue generation in a bilingual English–Hindi setting. Through a comprehensive empirical study across six models (250M–7B parameters, encoder-decoder and decoder-only architectures), we demonstrated several key findings:

1. **Citation-grounded SFT substantially reduces hallucination.** Training with explicit citation format reduces hallucination to 0.0% under automatic NLI-based evaluation for encoder-decoder models, and to 0.010–0.014 for Mistral-7B. Notably, LLaMA-3.2-1B also achieves 0.0% English hallucination after Stage 2 - but without generating any citation markers - demonstrating that hallucination elimination and citation format learning are separable outcomes. Human evaluation remains future work.
2. **Progressive training prevents catastrophic forgetting.** English metrics remain stable while Hindi capabilities improve substantially through Stages 1–3, validating the incremental skill-composition approach.
3. **SFT as a strong baseline for structured grounded tasks.** In our experimental configuration, GRPO alignment provided marginal improvement over SFT. This suggests that well-designed SFT may be a competitive baseline for citation-grounded dialogue, though the role of RL under different configurations warrants further investigation.
4. **Small models match large models after SFT.** A 250M-parameter model achieves identical English performance to a 780M model after Stage 2, with implications for cost-effective deployment.
5. **Generation collapse is recoverable.** Flan-T5-XL (3B) exhibited generation collapse at Stage 2 but recovered fully at Stage 3, demonstrating the resilience of continued training and highlighting the need for generation-time validation.

7.1 Future Work

Several directions emerge from this study:

Improved GRPO training. Reducing the KL penalty, increasing training steps, and designing graded reward signals may unlock RL benefits.

Human evaluation. Complementing automatic metrics with human judgments would provide a more complete evaluation picture.

Additional languages. Extending the pipeline to other Indian languages (Bengali, Tamil, Marathi) would test generalizability.

Explainability-guided training. Using attention visualization insights to design training objectives that encourage faithful attention over cited knowledge passages.

This work establishes a progressive training methodology for citation-grounded dialogue and provides baseline results across six language models. Future work will examine dataset-specific effects by training separately on DSTC9, Wizard of Wikipedia, and FaithDial, as well as language-specific effects through separate English and Hindi training tracks.

References

- BigScience Workshop. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2023.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. In *Journal of Machine Learning Research*, 2024.
- Raj Dabre, Ratish Puduppully, Anoop Kumar, Anoop Kunchukuttan, et al. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

- Jay Gala, Pranjal A. Chitale, Raghavan AK Raghavan, et al. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research (TMLR)*, 2023.
- Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kuber, Leshem Choshen, Raphael Sznajder, and Noam Slonim. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.
- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023.
- Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, et al. MuRIL: Multilingual representations for indian languages. In *arXiv preprint arXiv:2103.10730*, 2021.
- Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2020.
- Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Miaoran Li, Baolin Peng, Jianfeng Gao, and Zhu Zhang. OPERA: Harmonizing task-oriented dialogs and information seeking experience. *ACM Transactions on the Web*, 18(4):45, 2024. doi: 10.1145/3623381.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL Workshop*, 2004.
- Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

- Michael McCloskey and Norman J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.
- Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.
- Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018.
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Ankur Parikh. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arber, Angela Fan, et al. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. In *arXiv preprint arXiv:2208.03188*, 2022.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- Xin Tang, Dingcheng Feng, Adam Jatowt, et al. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- Sarah Wiegrefe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Appendix A: LLaMA-3.2-1B Generated Output Examples

Table 15 presents representative Stage 3 outputs from LLaMA-3.2-1B alongside the expected citation-grounded reference. English outputs contain factually correct content but no citation markers; Hindi outputs correctly produce [N] markers integrated into natural-language phrasing.

Table 15: Representative LLaMA-3.2-1B Stage 3 outputs vs. expected references. EN outputs contain zero citation markers despite the model having access to the numbered knowledge passage. HI outputs correctly produce [N] markers.

L	Query (excerpt)	Expected Output	LLaMA-3.2-1B Output	Cit.
EN	Can you bring your own liquor to Pizza Hut Cherry Hinton?	As noted in [1], No, outside beverages are not allowed.	No, you cannot bring your own liquor at Pizza Hut Cherry Hinton. Do you have any other questions? Would you like to make a reservation?...	0
EN	How long did California Love stay on Billboard Hot 100?	Based on [1], yeah, stayed on the billboard hot 100's for two weeks straight!	It was one of 2Pac's most popular singles, reaching number one on the Billboard Hot 100 for two weeks. It was a double A-side single...	0
EN	Restaurant booking confirmation	According to [1], booking was successful. Ref: ENOALFAV.	Your reservation is all set. The reference number is 1ZQDOZ1A...	0
HI	[Query about wheelchair accessibility at Prezzo]	[2] में कहा गया है कि हां, प्रेज़ो में व्हीलचेयर उपलब्ध है।	[2] के आधार पर, हां, प्रेज़ो में व्हीलचेयर उपलब्ध है इसलिए उन्हें आपकी सहायता करने में सक्षम होना चाहिए।	[2] ✓
HI	[Query about the color black]	[1] में उल्लेख है कि काले रंग के बारे में, यह पूरी तरह से दिखाई देने वाले प्रकाश को अवशोषित करता है।	[1] में कहा गया है कि मुझे नहीं पता था कि यह था, काले रंग के बारे में...	[1] ✓

Appendix B: Figures

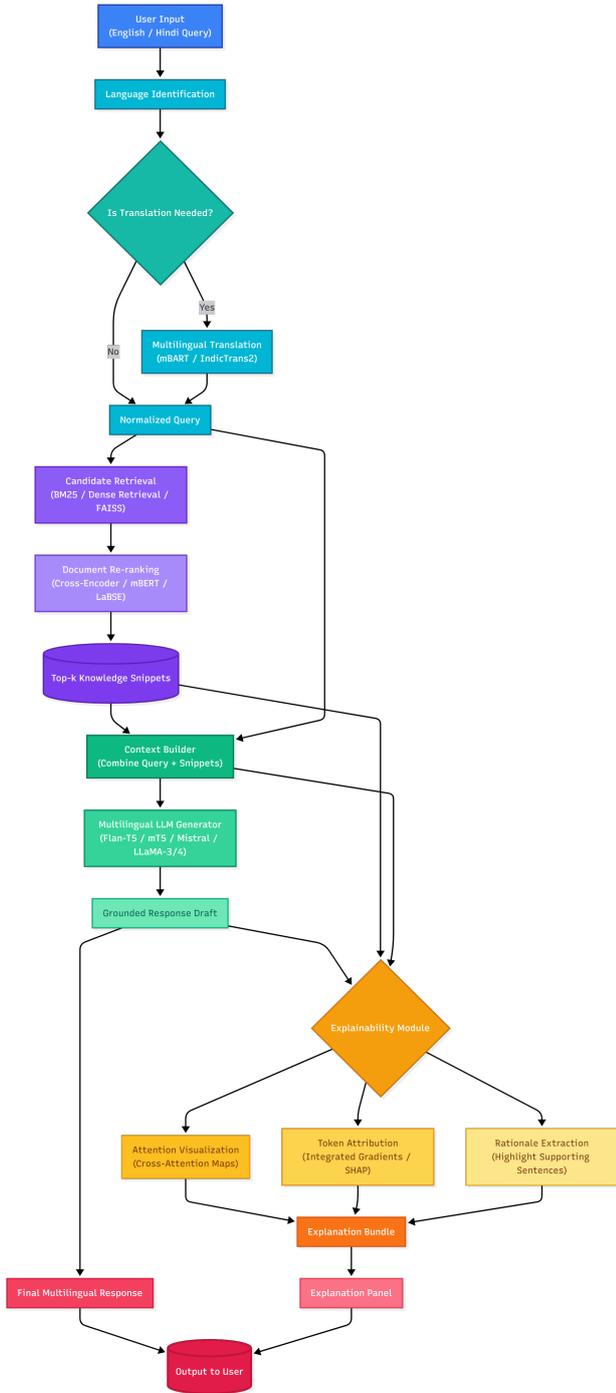


Figure 7: System architecture of XKD-DIAL. User queries undergo language identification and optional translation before retrieval. The context builder combines the query with top- k knowledge snippets. The multilingual LLM generator produces a citation-grounded response, which is analyzed by the explainability module.

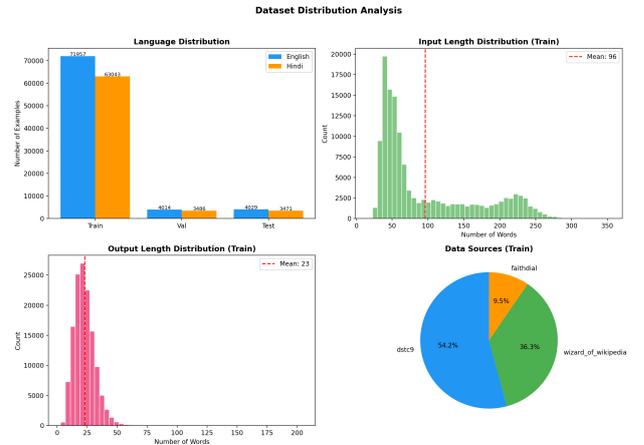


Figure 8: Data distribution across source corpora (DSTC9, FaithDial, Wizard of Wikipedia) and language splits (English, Hindi) across the train, validation, and test partitions.

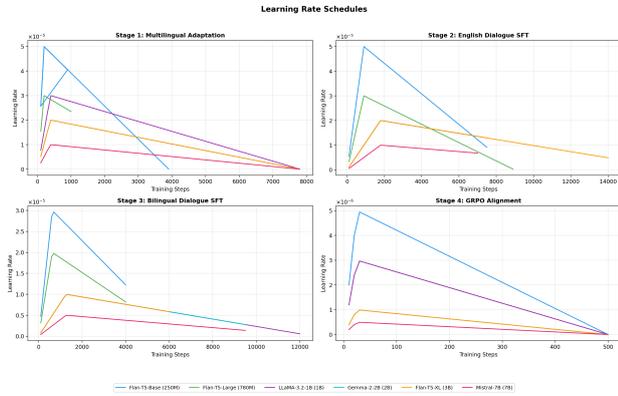


Figure 9: Learning rate schedules across all training stages and models. Cosine decay with warmup is used for SFT stages; a lower learning rate with linear warmup is used for GRPO.

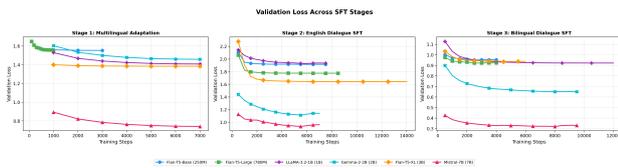


Figure 11: Validation loss across SFT stages for all six models. Mistral-7B achieves the lowest loss throughout; Gemma-2-2B reaches 0.651 at Stage 3. Flan-T5-Large (dashed line, Stage 1) uses training loss as fallback since eval loss was not logged for that stage.

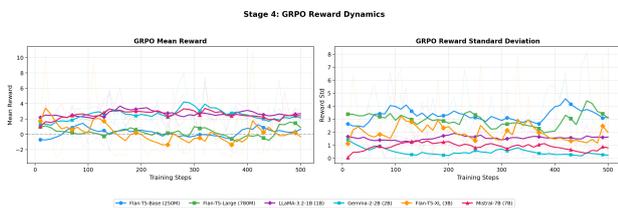


Figure 13: GRPO reward trajectory during Stage 4 training for all six models. Gemma-2-2B achieves the highest absolute reward (3.40) with no decline; most other models show reward regression from best to final.

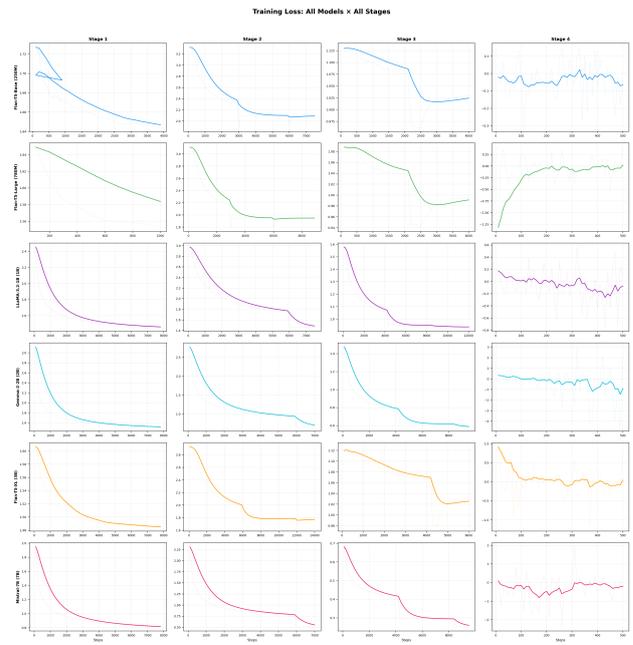


Figure 10: Training loss progression across all stages for all six models. Stage 2 (English SFT) and Stage 3 (Bilingual SFT) show consistent convergence across architectures.

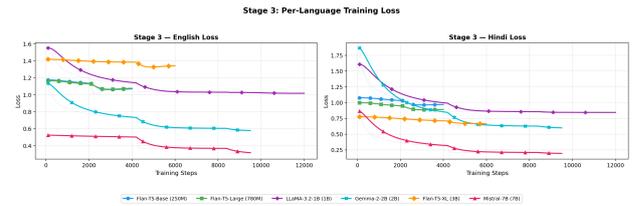


Figure 12: Language-specific training loss during Stage 3 (Bilingual SFT). Hindi loss starts higher but decreases rapidly, converging with English.

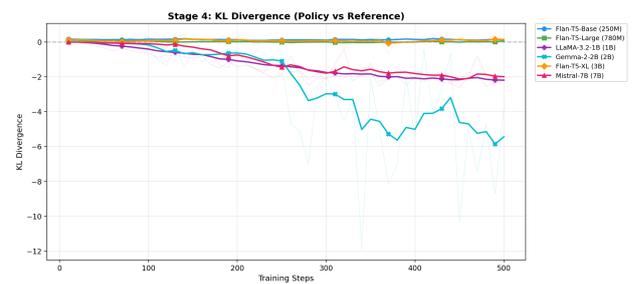


Figure 14: KL divergence between the policy and reference model during GRPO training. The KL penalty ($\beta = 0.04$) constrains policy drift.

Appendix C: Additional Evaluation Figures

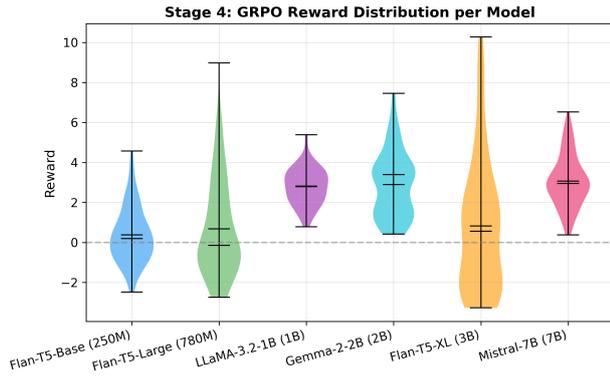


Figure 15: Violin plot of GRPO reward distributions per model. Mistral-7B (7B) shows a compact, high-reward distribution centered around 3.0, while encoder-decoder models show wider spread with lower medians.

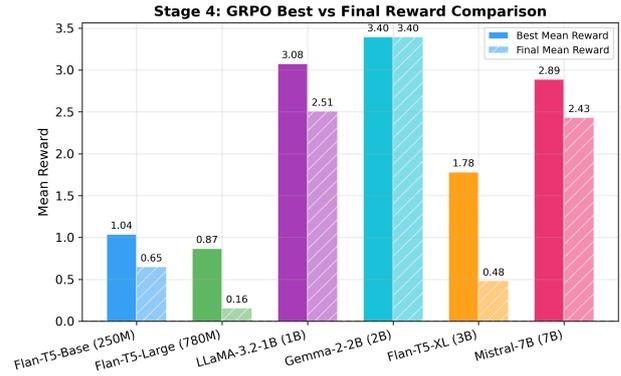


Figure 16: GRPO best vs. final mean reward comparison. All models show reward decline from best to final, with Flan-T5-XL exhibiting the largest drop (-1.30) and Mistral-7B the smallest (-0.46).

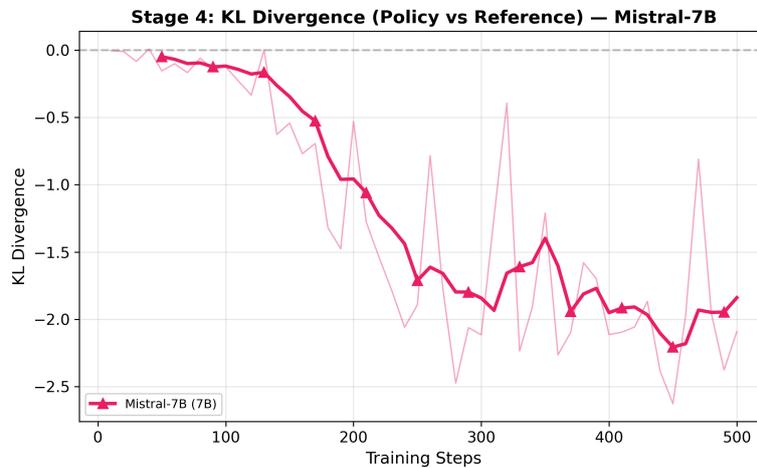


Figure 17: KL divergence (policy vs. reference) for Mistral-7B during GRPO. The negative KL indicates the policy diverges from the reference, stabilizing around -2.0 after 300 steps despite the KL penalty ($\beta = 0.04$).