Enhancing In-Context Learning for Explainable Translation Evaluation with Generative AI

From Research to Business Integration

Quy Nguyen

image generated with DALL-E 3

AGENDA

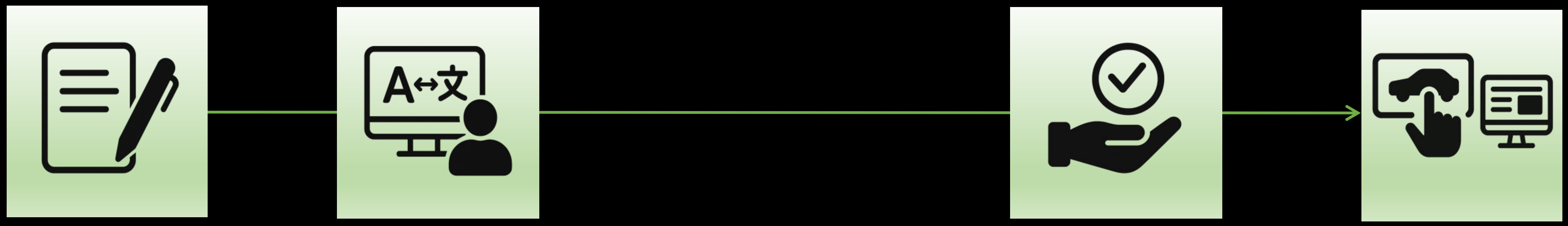
1 Introduction

2 Methodology & Prompting Strategies

3 Evaluation & Results

4 Conclusion & Future Work

Manual Quality Assessment in Translation Workflow



- Expert review process: Translators check TM outputs and suggest corrections
- High-quality judgment: Captures subtle distinctions in translation quality
- Scalability challenges: Costly, time-consuming, and sensitive to terminology updates

Goal: Semi-Automated Quality Assurance

Source Text Creation

Infotainment
OneWeb

Computer-Aided Translation

Human
Translation Memory

Automated Quality Assurance

Human Review

High quality
Not scalable

Flawless Target text

- **Automated pre-check**: Insert QA before human review to improve scalability
- **Human-like output:** Explainable, actionable, and robust error feedback
- **Cross-domain ready:** Works across different content types without retraining

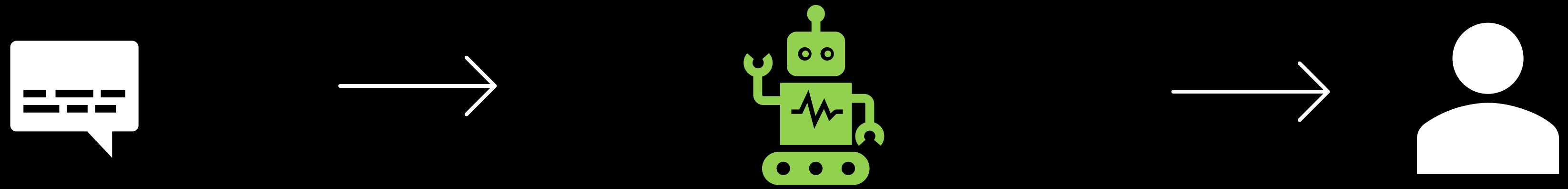
Task Definition: Explainable Translation Evaluation

- **Reference-less setup:** No gold translation needed; context-aware evaluation
- **Three-part output:** rating label, detailed diagnosis, and revised translation
- **Human-like process:** Detects, explains, and corrects flawed translations

Translation Pair

Automated Quality Assurance

Human Review



Translation Pair:

1. Source
2. Target
3. Context

Human Review:

1. Detect
2. Explain
3. Correct

Motivation for a GenAI-Based QA Solution

- **Prompt-driven adaptation**: Handles new tasks without domain-specific training
- **Human-like feedback**: Produces explainable and actionable outputs
- **Scalable & efficient**: Reduces manual effort compared to traditional reviews

Research Questions

- **RQ1 – Prompting Strategies:** How do different prompts affect LLM output quality?
- **RQ2 – Terminology Retrieval**: Can RAG improve error detection and correction?
- **RQ3 – Cross-Domain Robustness:** Does the approach generalize across domains?

AGENDA

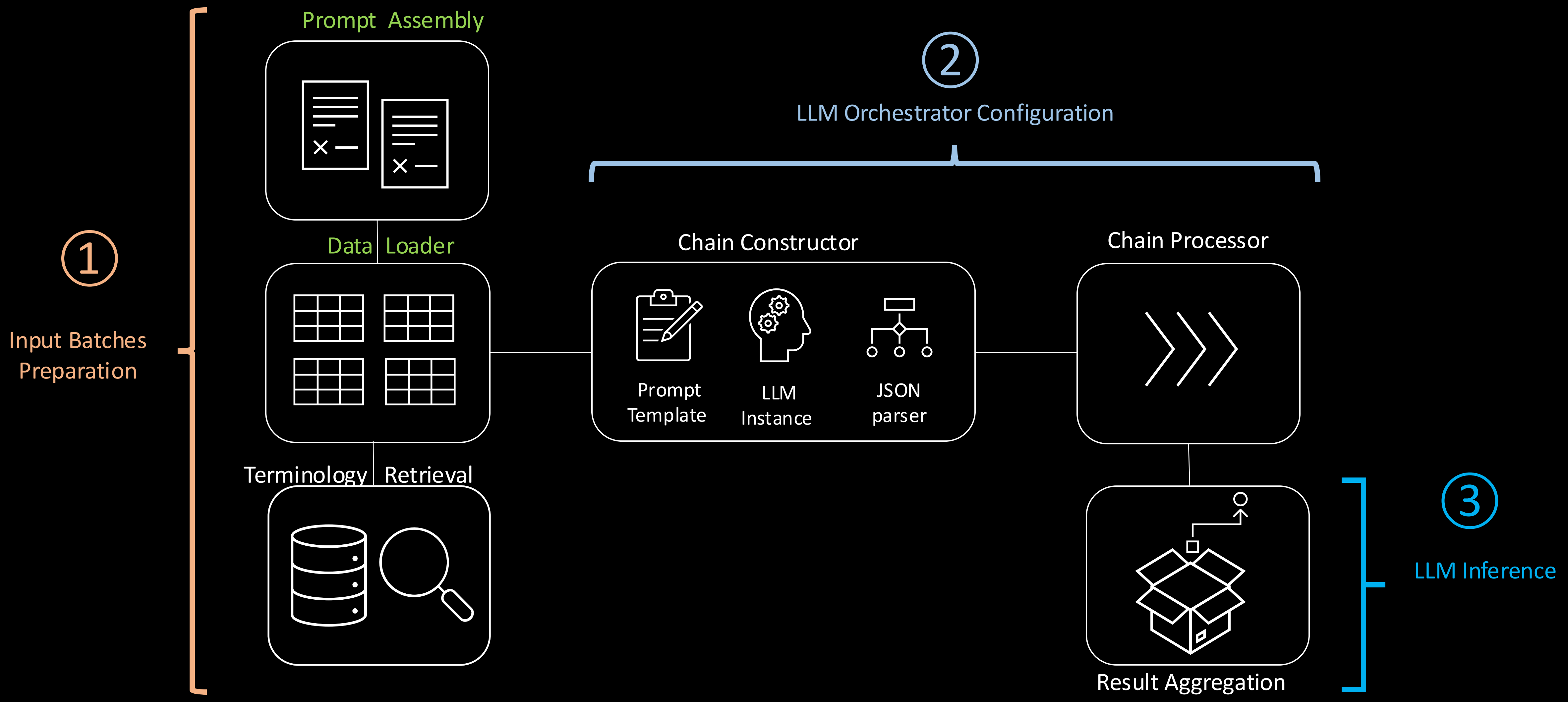
1 **Introduction**

2 **Methodology & Prompting Strategies**

3 **Evaluation & Results**

4 Conclusion & Future Work

System Architecture



Domain Overview

- Text Type & Length
  - Infotainment: Short, context-bound UI strings
  - OneWeb: Longer, narrative-style marketing content
- Linguistic Characteristics
  - Infotainment: High noun density, many UI placeholders, low readability
  - OneWeb: Stylistically rich, no placeholders, higher readability
- Common Error Types
  - Infotainment: Frequent terminology, incorrect, and incomprehensible errors
  - OneWeb: Dominated by style and incompleteness; minimal terminology issues

Prompting Strategies Overview

Core Strategies

- **Baseline:** minimal guidance
- **CLA:** Single-step Chain-of-Thought
- **NEXUS:** Multi-step Least-to-Most

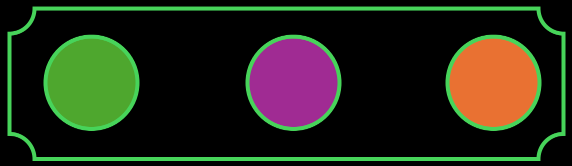
Terminology-Aware Variants

- **RAG:** Adds relevant glossary entries

Domain-Aware Extensions

- **RAG+:** Incorporates domain-specific rules
- **Genre:** Adds information on text type, audience, and context
- **Style:** Introduces a subtask for stylistic appropriateness

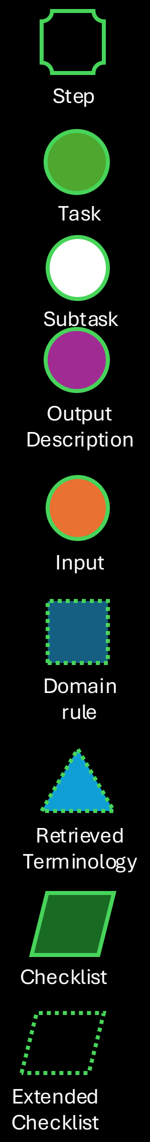
Baseline

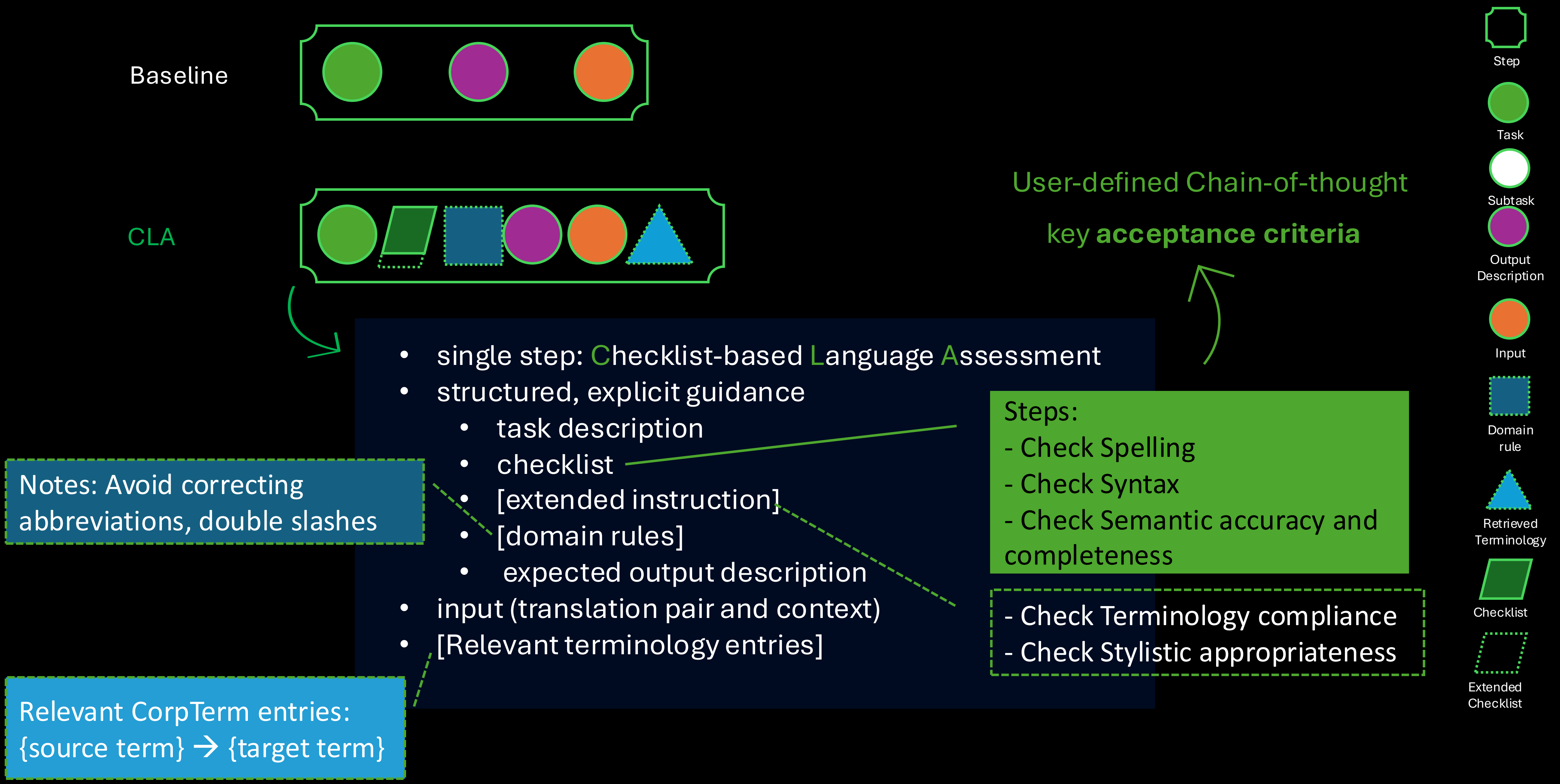


- single prompt
- minimal guidance
  - task description
  - expected output description (3-part outputs: ratings, explanation, correction)
- input (translation pair and context)

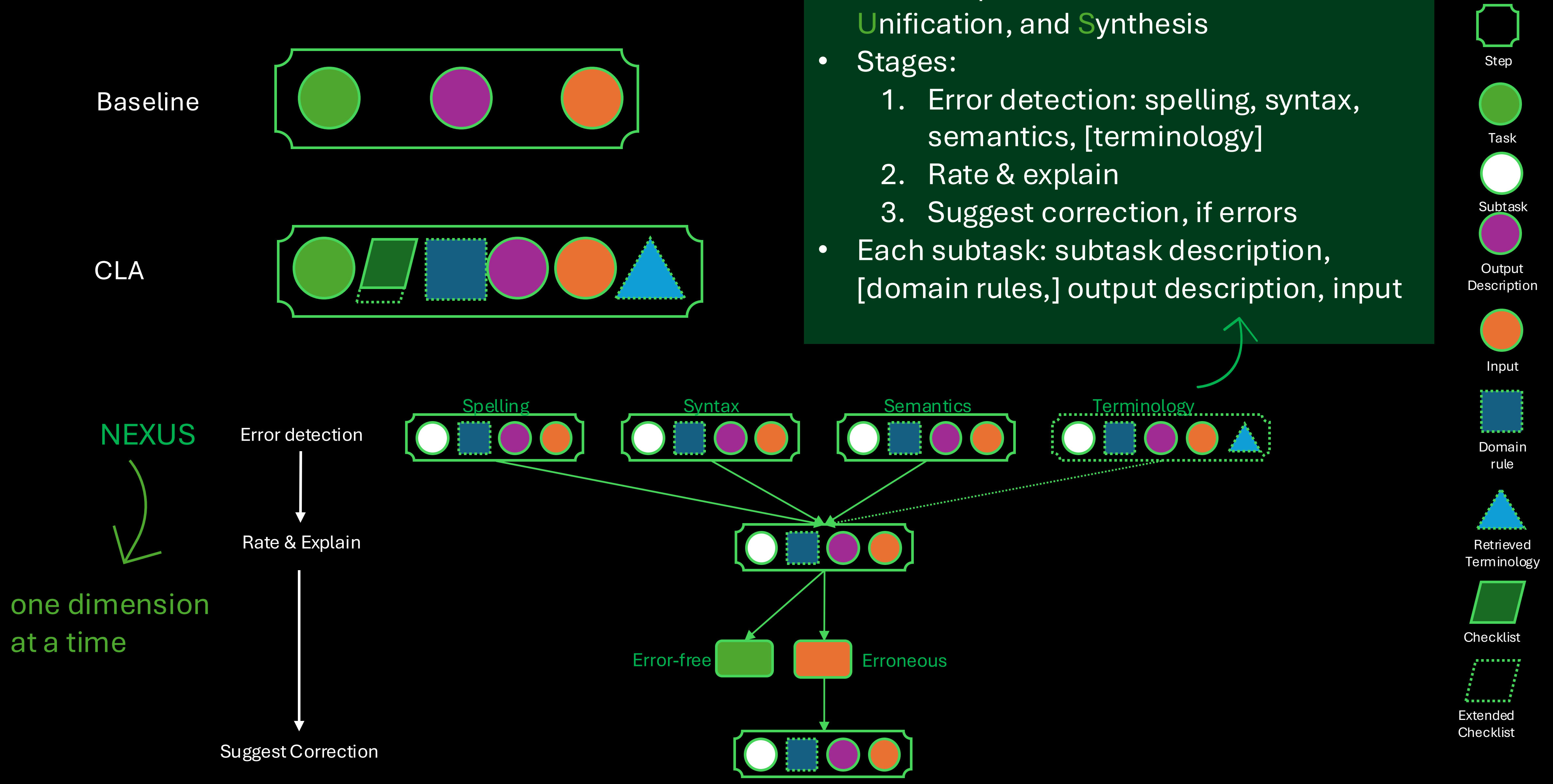
Evaluate translation quality

The output should be in the following schema
`label:` OK or Error,
`explain:` Brief explanation,
`suggest:` Improved translation if Error




Baseline
CLA
single step: Checklist-based Language Assessment
structured, explicit guidance
task description
checklist
[extended instruction]
[domain rules]
expected output description
input (translation pair and context)
[Relevant terminology entries]
Notes: Avoid correcting abbreviations, double slashes
Relevant CorpTerm entries: {source term} → {target term}
Steps:
- Check Spelling
- Check Syntax
- Check Semantic accuracy and completeness
- Check Terminology compliance
- Check Stylistic appropriateness
User-defined Chain-of-thought
key acceptance criteria
Step
Task
Subtask
Output Description
Input
Domain rule
Retrieved Terminology
Checklist
Extended Checklist

- Multi-step: Nodal Evaluation, Xtraction, Unification, and Synthesis
- Stages:
  1. Error detection: spelling, syntax, semantics, [terminology]
  2. Rate & explain
  3. Suggest correction, if errors
- Each subtask: subtask description, [domain rules,] output description, input



AGENDA

1 **Introduction**

2 **Methodology & Prompting Strategies**

3 **Evaluation & Results**

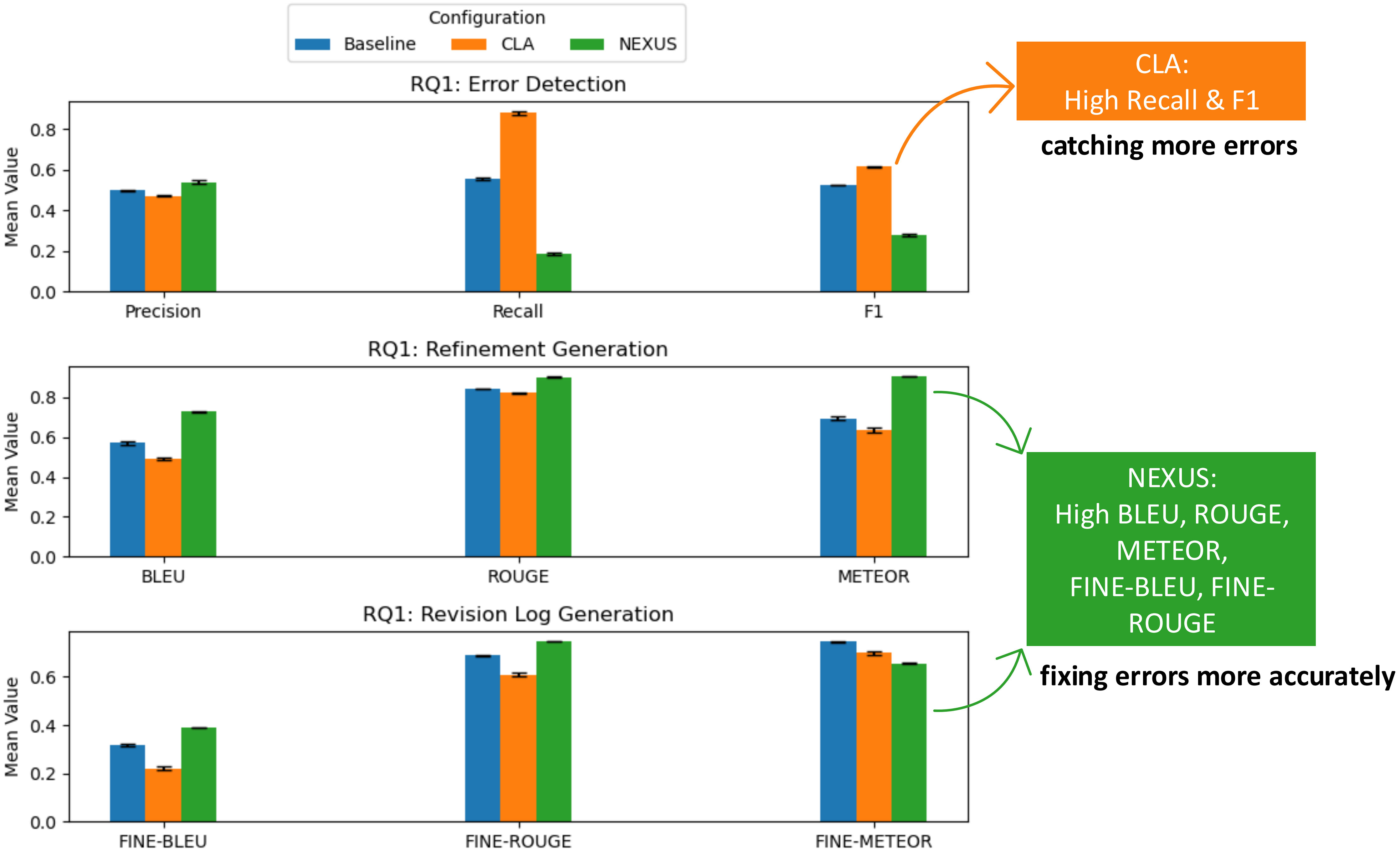
4 Conclusion & Future Work

Evaluation Metrics

- **Error Detection:** Precision, Recall, F1
- **Refinement:** BLEU, ROUGE, METEOR
- **Explanation (Indirect):**
  - Use **FINE Scores** (Fine-grained Identification of Nuanced Edits)
  - No direct evaluation due to ground truth inconsistencies
- **FINE Score Evaluation Steps:**
  1. Derive revision logs from both system and human corrections
  2. Compute BLEU, ROUGE, and METEOR over these edit segments

RQ1 - Prompting Strategies: CLA vs. NEXUS

- **Hypothesis**: Clear guidance (CLA) and structured reasoning (NEXUS) improve performance
- **CLA**
- **NEXUS**


Configuration
Baseline
CLA
NEXUS
RQ1: Error Detection
Mean Value
0.0
0.2
0.4
0.6
0.8
Precision
Recall
F1
RQ1: Refinement Generation
Mean Value
BLEU
ROUGE
METEOR
RQ1: Revision Log Generation
Mean Value
FINE-BLEU
FINE-ROUGE
FINE-METEOR
CLA:
High Recall & F1
catching more errors
NEXUS:
High BLEU, ROUGE, METEOR,
FINE-BLEU, FINE-ROUGE
fixing errors more accurately

Error Detection
CLA: strength in all error types, even subtle ones (style, incomprehensible)

NEXUS: high precision, but not for subtle and subjective errors

RQ1 - Prompting Strategies: CLA vs. NEXUS

- **Hypothesis**:
  - ✓ Clear guidance (CLA) and structured reasoning (NEXUS) improve performance
- **CLA**: Best for **error detection**
- **NEXUS**: Strongest in **correction quality**
- The structure of the prompt directly influences model responses
- **Choosing the right strategy depends on the task goal**

RQ2 - Impact of Terminology and Domain Knowledge Integration

Hypothesis

Integrating enterprise terminology boosts translation quality evaluation.

Adding Terminology (RAG)

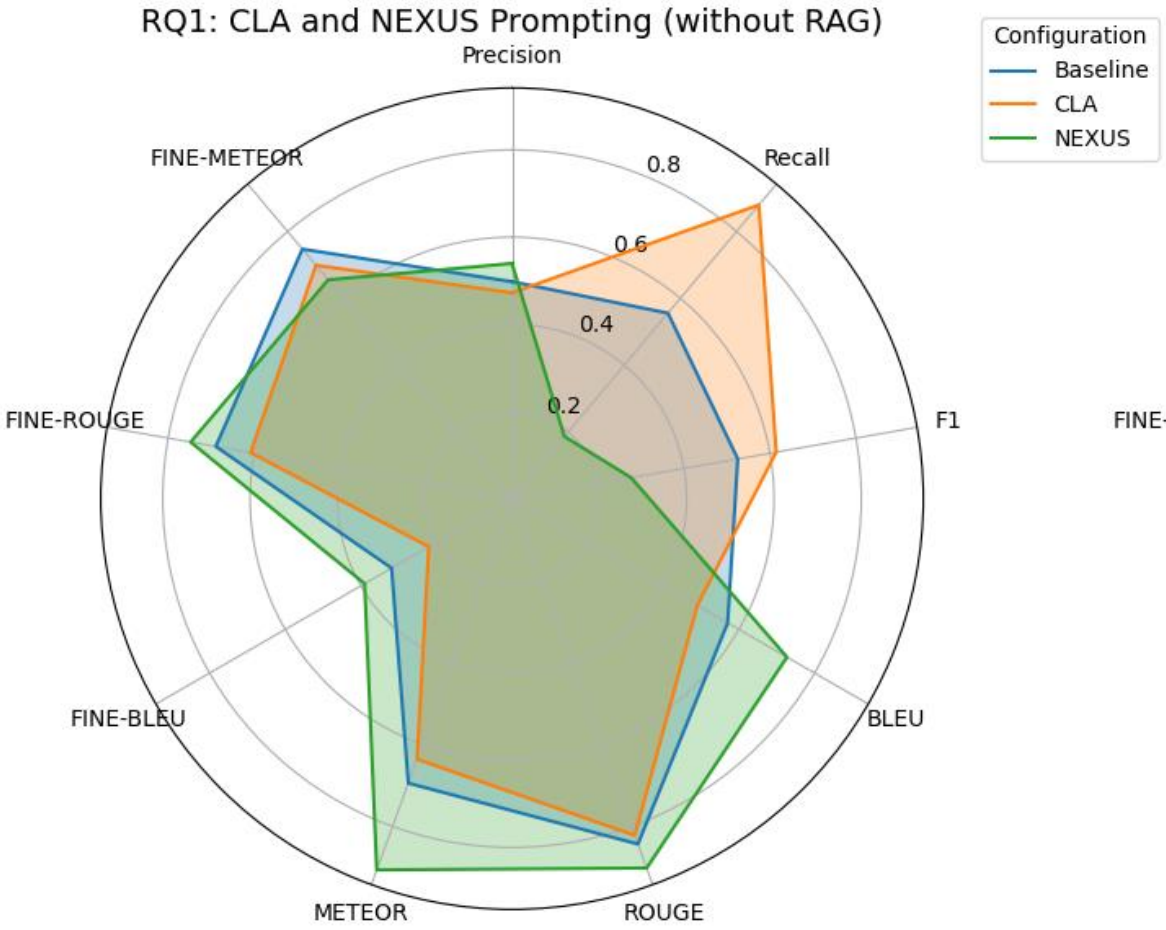
- CLA-RAG
- NEXUS-RAG

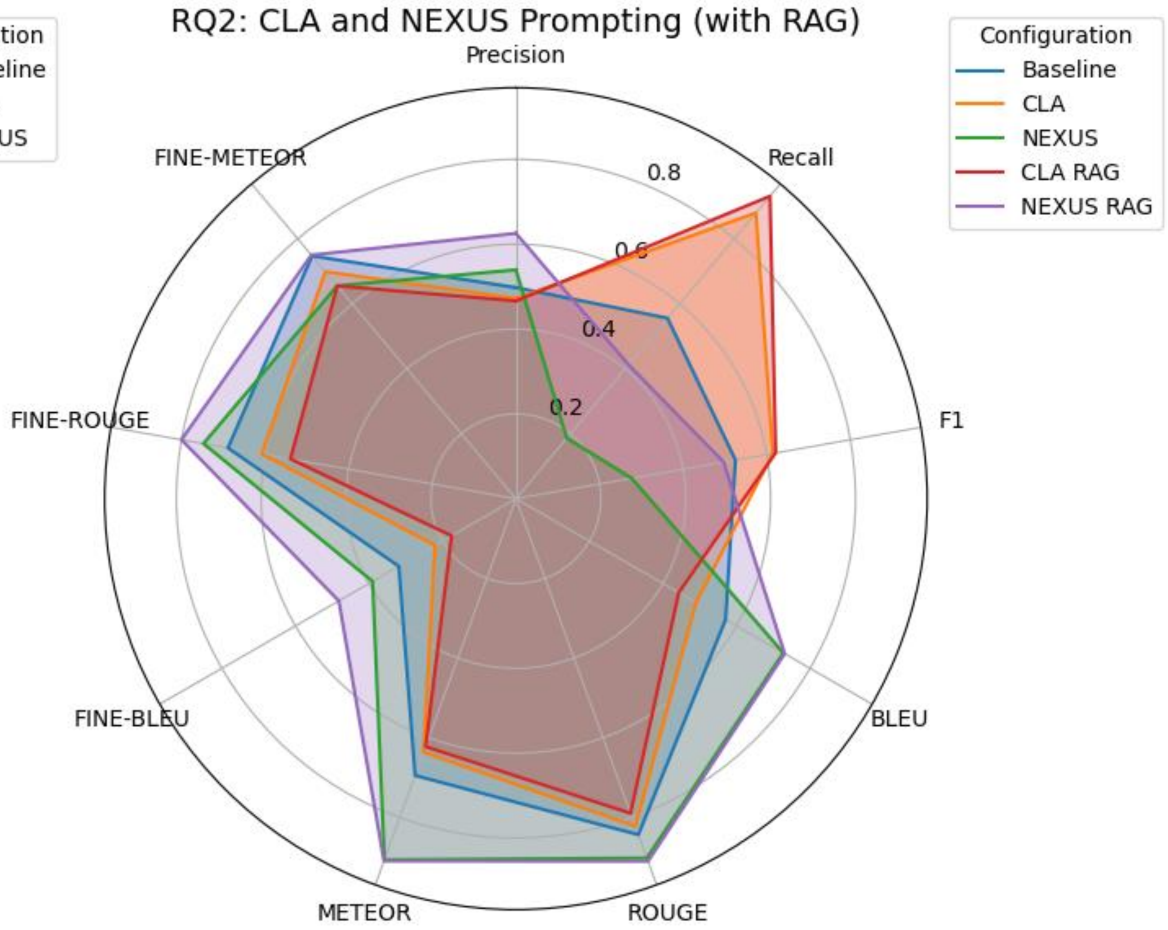
Adding Domain Rules (RAG+)

- CLA-RAG+
- NEXUS-RAG+

CLA vs baseline: gain in Recall & F1

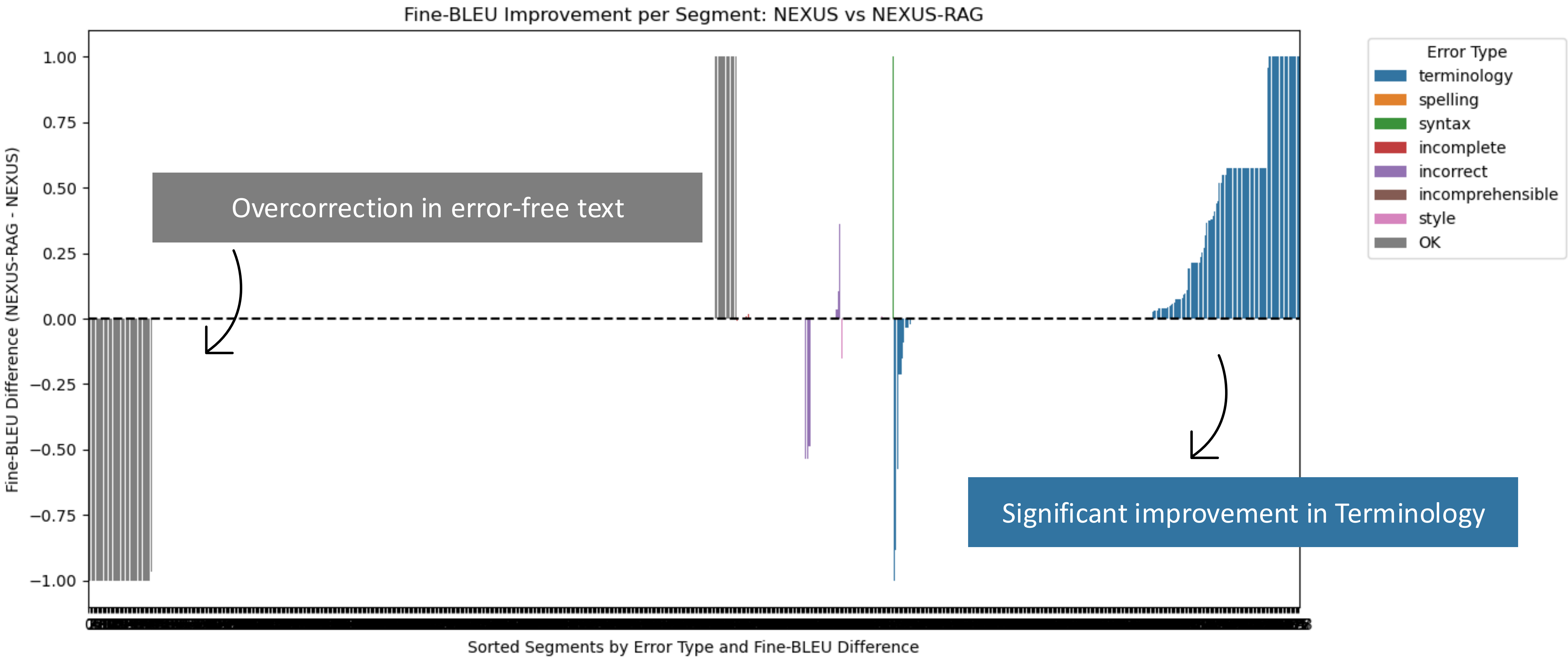
CLA-RAG vs CLA: gain in Recall





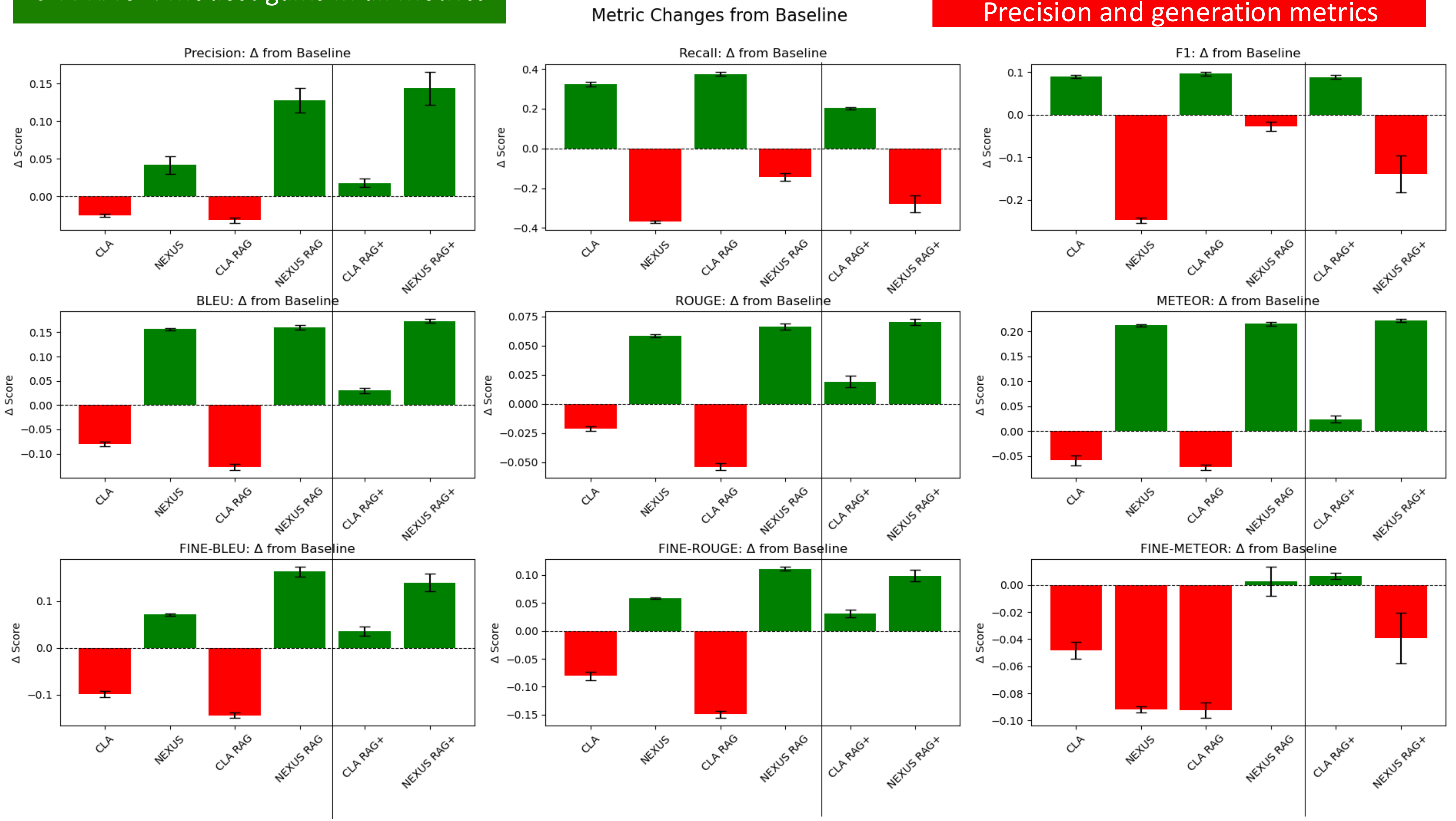
NEXUS vs baseline: gains in Precision and generation metrics

NEXUS-RAG vs NEXUS: gains in all metrics


Fine-BLEU Improvement per Segment: NEXUS vs NEXUS-RAG
Fine-BLEU Difference (NEXUS-RAG - NEXUS)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Sorted Segments by Error Type and Fine-BLEU Difference
Error Type
terminology
spelling
syntax
incomplete
incorrect
incomprehensible
style
OK
Overcorrection in error-free text
Significant improvement in Terminology

CLA-RAG+: modest gains in all metrics

NEXUS-RAG+: significant gains in Precision and generation metrics



RQ2 - Impact of Terminology and Domain Knowledge Integration

Hypothesis

✓ Integrating enterprise terminology boosts translation quality evaluation

Adding Terminology (RAG)

- **CLA-RAG**: Higher **Recall** and **F1** — better at catching terminology errors
- **NEXUS-RAG**: Higher **Precision** and **correction quality** — excels in terminology-heavy segments

Adding Domain Rules (RAG+)

- **CLA-RAG+**: Delivers **modest, consistent gains** across metrics
- **NEXUS-RAG+**: Yields **best overall performance,** but lowers **Recall**

RQ3 - Cross-Domain Robustness

- **Hypothesis**

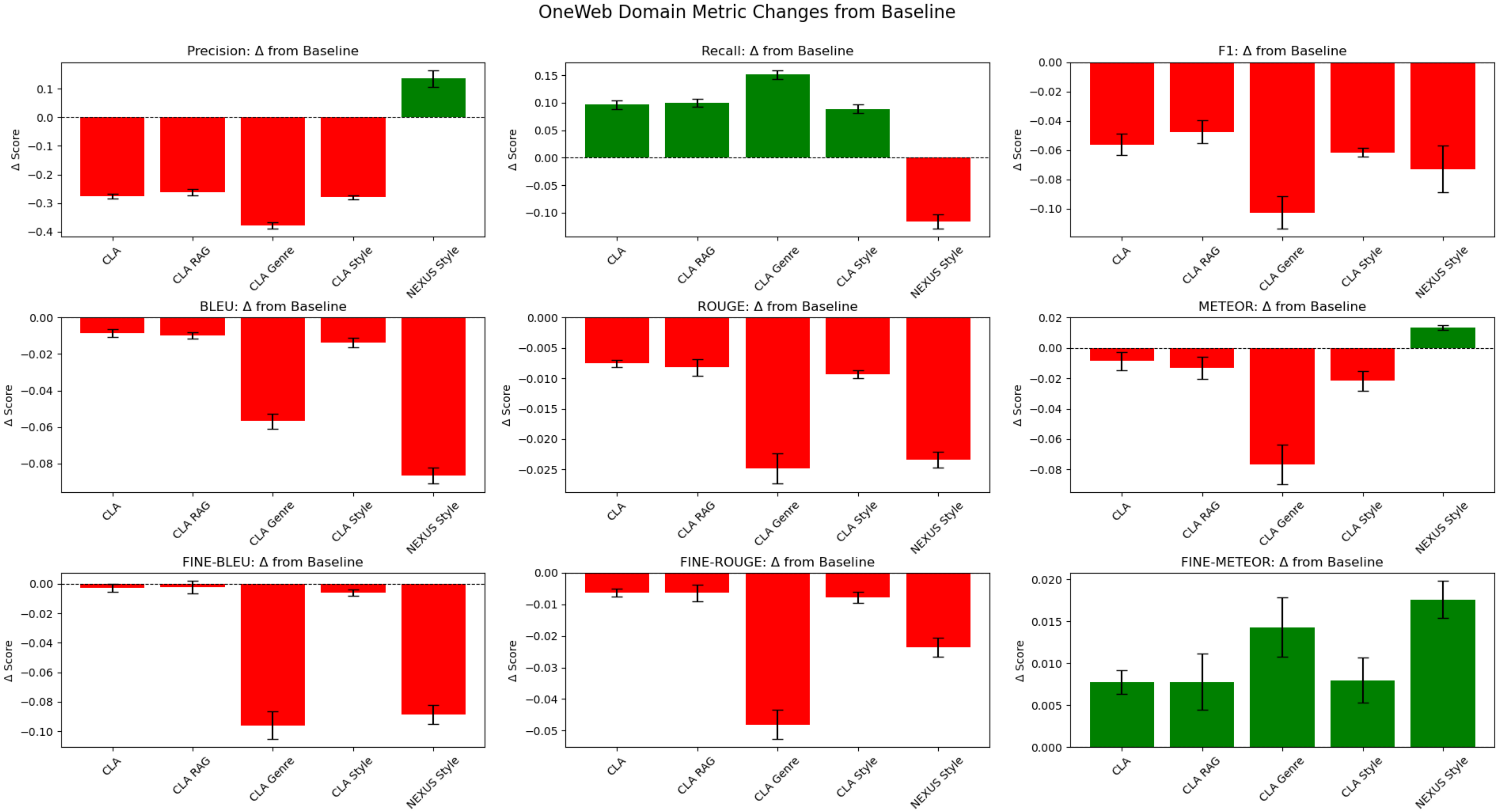
  Domain-agnostic prompts are expected to perform effectively; slight adaptations may further improve performance.

- **CLA-based variants** (CLA, CLA-RAG, Genre, Style)

- NEXUS-Style

CLA-based variants: improve Recall

RAG: minimal gains



NEXUS: improves Precision, METEOR and FINE-METEOR

RQ3 - Cross-Domain Robustness

- **Hypothesis:**

  ✓ Domain-agnostic prompts are expected to perform effectively; slight adaptations may further improve performance.
- **CLA-based variants** (CLA, CLA-RAG, Genre, Style):
  - Improve **Recall**
  - Address **stylistic and omission-related errors**
  - Modest Recall gains, but often at the cost of **Precision**
- NEXUS-Style:
  - Improve **Precision**, **METEOR**, and **FINE-METEOR**
  - Minor gains in revision quality, but lower Recall rate
- Takeaway:

  Baseline works reasonably well; domain-aware adaptations provide measurable benefits, especially when tuned to the genre information

AGENDA

1 **Introduction**

2 **Methodology & Prompting Strategies**

3 **Evaluation & Results**

4 Conclusion & Future Work

Limitations

Annotation Challenges

- **Label inconsistency:** Full rewrites vs. minimal edits, major vs. minor errors
- **Source variation:** Human-annotated (Infotainment) vs. semi-automated (OneWeb)

Technical Constraints

- **Content filtering:** Azure moderation blocked some segments
- **Prompt limitations:** Token and budget constraints restricted complexity

Evaluation Limitations

- **Surface metrics:** BLEU/ROUGE may miss deeper quality differences
- **Direction mismatch:** Infotainment (→ English), OneWeb (← English)
- **Language gaps:** No low-resource or morphologically rich language coverage
- **Lack of feedback:** No human-in-the-loop or user study evaluation

Conclusion

- Practical Contribution
  - Scalable QA system: Enables explainable translation evaluation in enterprise workflows
  - Cross-domain applicability: Works across domains with minimal adaptation
- Theoretical Contribution
  - Structured prompting: Proposes CLA and NEXUS for explainable translation QA
  - Complementary strategies: CLA excels at detection; NEXUS at refinement
  - Terminology integration: RAG enhances performance in term-sensitive domains
  - Prompt tuning: Improves alignment with domain-specific requirements
- A scalable, explainable, and production-ready approach to translation quality assurance.

Future Work

- Prompting Strategies:

  Explore compressed prompts, dynamic few-shot examples from TM, and alternate formats/languages

- RAG Improvements:

  Use richer sources (e.g., style guides, reclamation data), and optimize retrieval quality

- Broader Scope:

  Test on new domains, language pairs, and low-resource settings

- Evaluation Enhancements:

  Incorporate semantic metrics and human feedback

- Deployment:

  Prepare for integration into production QA workflows

Thank You!

- Supervisor: Dr. Agnieszka Faleńska
- Advisors: Nadia Mast, Sevde Ceylan
- Team Translation R&D, Driver Information & Production IPS/142
- Language Technology IPS/14