



Framing the Language: Fine-Tuning Gemma 3 for Manipulation Detection

*M. Khandoga, Y. Kostiuk, A. Polishko,
K. Kozlov, Y. Filipchuk and A. Kiulian*



OpenBabylon

ARG-tech
Centre for Argument Technology

Manipulation Detection

(in Ukrainian Social Media)



Span identification
(Binary Span Detection)

Technique classification
(Multilabel classification)

Manipulation Detection

(in Ukrainian Social Media)



Span identification

(Binary Span Detection)

- Loaded language
- Glittering generalities
- Euphoria
- Appeal to fear
- FUD

Technique classification

(Multilabel classification)

- Bandwagon/Appeal to people
- Thought-terminating cliché
- Whataboutism
- Cherry Picking
- Straw Man

Dataset

Новий огляд мапи DeepState від російського військового експерта, кухара путіна 2 розряду, спеціаліста по снарядному голоду та ректора музичної академії міноборони рф Євгенія Пригожина.

Пригожин прогнозує, що невдовзі настане день звільнення Криму і день розпаду росії. Каже, що передумови цього вже створені.

*Відео взяли з каналу

ФД

.

@informnapalm

Новий огляд мапи DeepState від російського військового експерта, кухара путіна 2 розряду, спеціаліста по снарядному голоду та ректора музичної академії міноборони рф Євгенія Пригожина.

Пригожин прогнозує, що невдовзі настане день звільнення Криму і день розпаду росії. Каже, що передумови цього вже створені.

*Відео взяли з каналу

ФД

.

@informnapalm

Technique classification

Euphoria

Loaded Language

Новий огляд мапи DeepState від російського військового експерта, кухара путіна 2 розряду, спеціаліста по снарядному голоду та ректора музичної академії міноборони рф Євгенія Пригожина.

Пригожин прогнозує, що невдовзі настане день звільнення Криму і день розпаду росії. Каже, що передумови цього вже створені.

*Відео взяли з каналу

ФД

.

@informnapalm

Span identification

Пригожин прогнозує, що невдовзі настане день звільнення Криму і день розпаду росії. Каже, що передумови цього вже створені

від російського військового експерта

кухара путіна 2 розряду

спеціаліста по снарядному голоду та ректора музичної академії міноборони рф Євгенія Пригожина

Technique classification

Euphoria

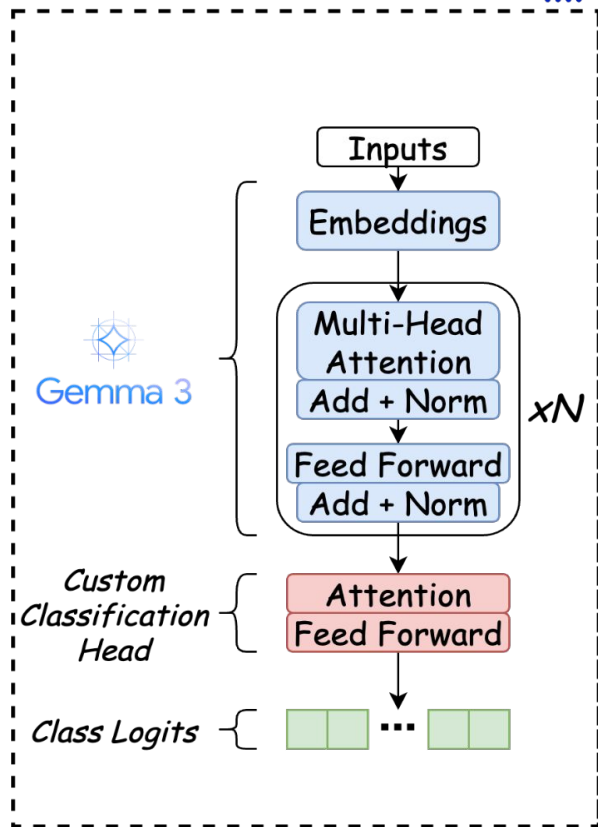
Loaded Language

Train	3,822
<i>Language distribution</i>	UA: 2,147 RU: 1,675
<i>Contain at least one manipulation</i>	2,589 (67,7%)
Test (Private)	5,735
<i>Estimated language distribution</i>	UA: 3,230 RU: 2,505

Methodology

Technique Classification

LLM Base Classifier



Train/dev split: 80/20 %

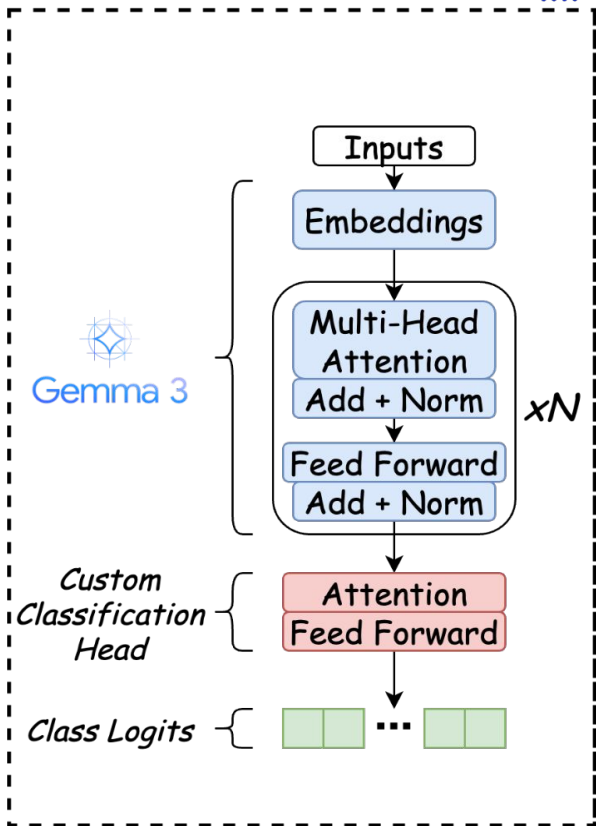
10 epochs

Best on dev.

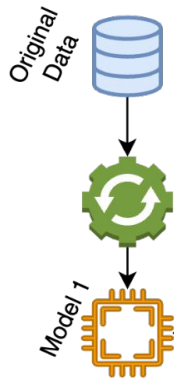
LR: $7e-5$

LR: $2e-6$

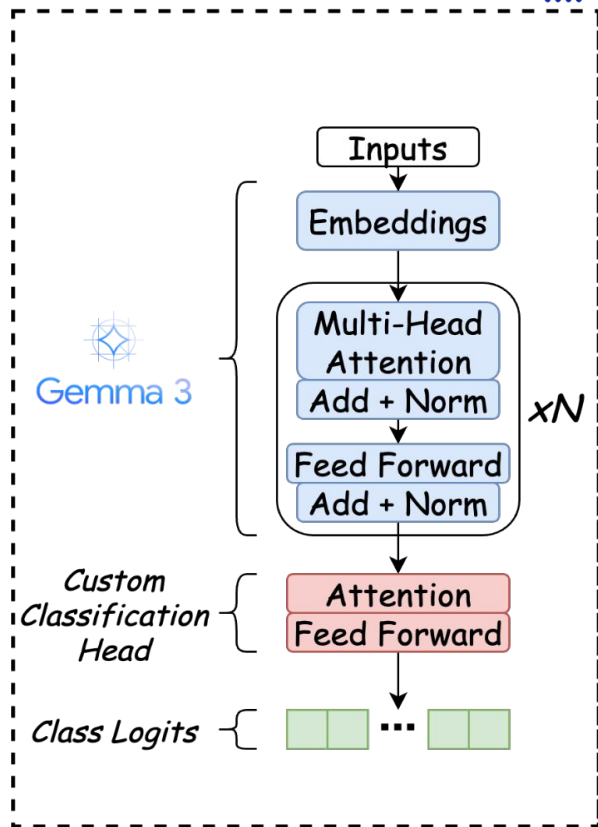
LLM Base Classifier



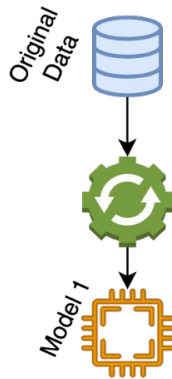
1. Train Classifier on Original Data



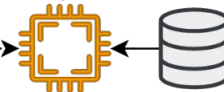
LLM Base Classifier



1. Train Classifier on Original Data



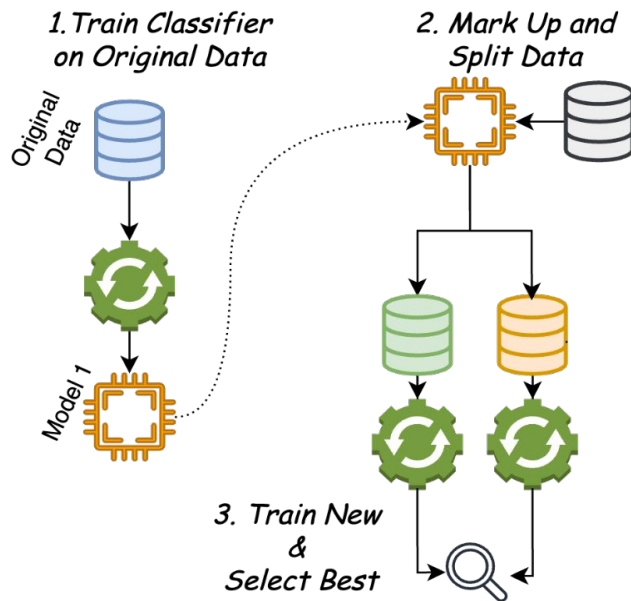
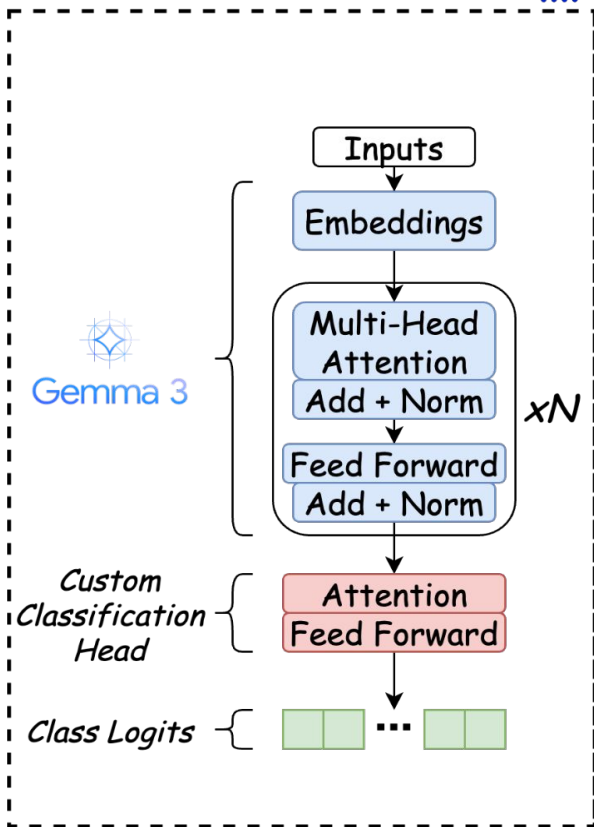
2. Mark Up and Split Data



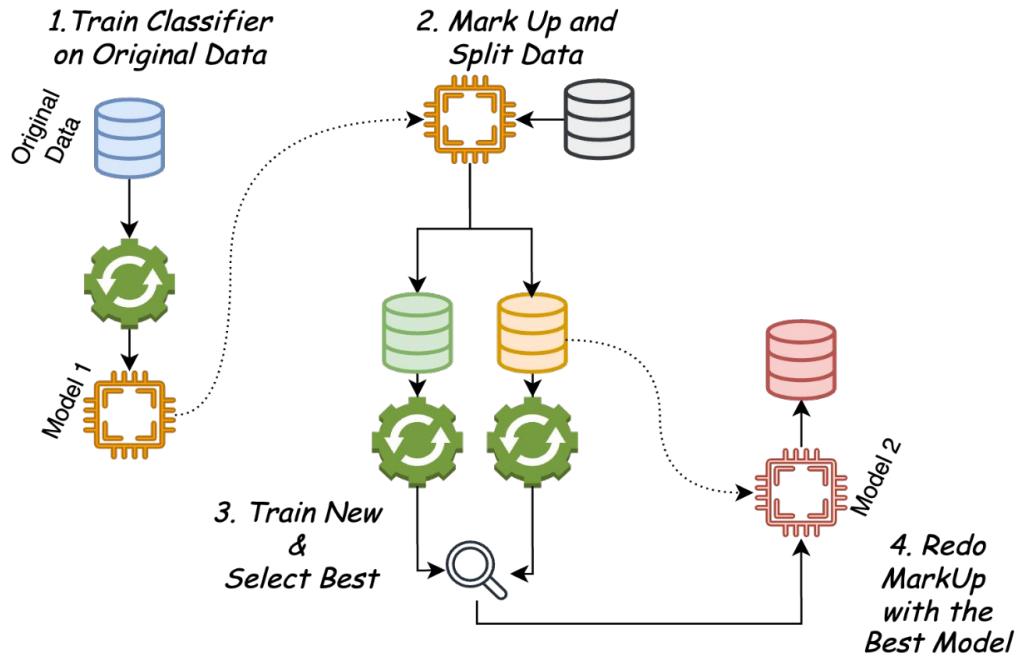
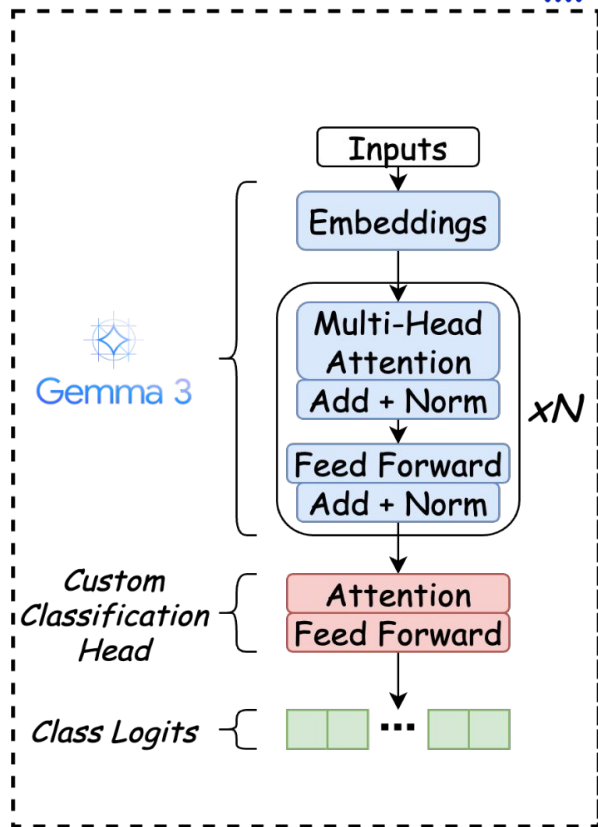
400k UKR news dataset



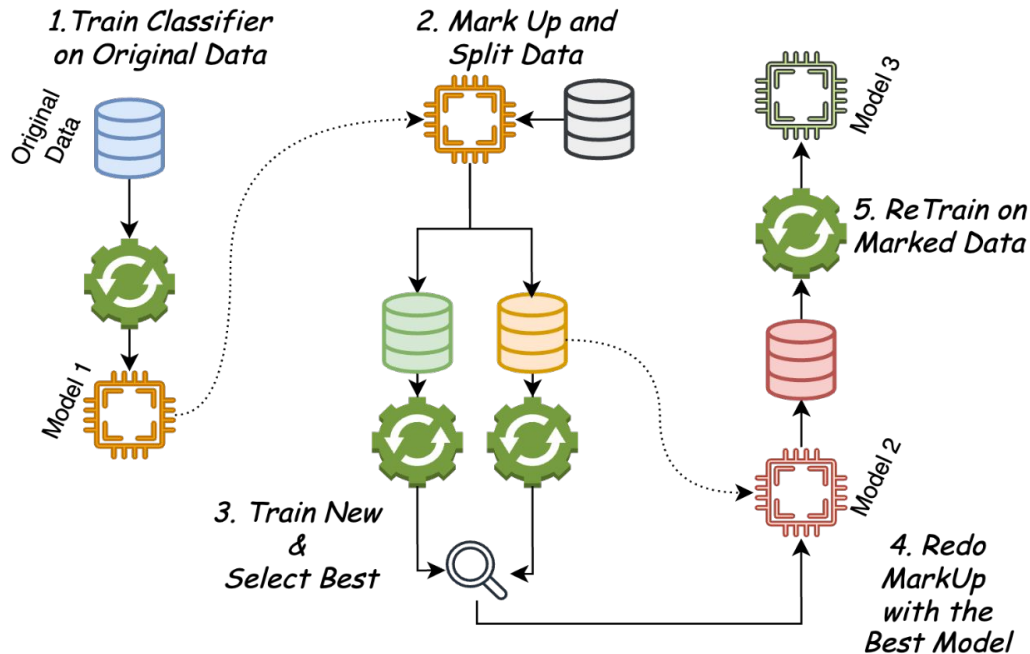
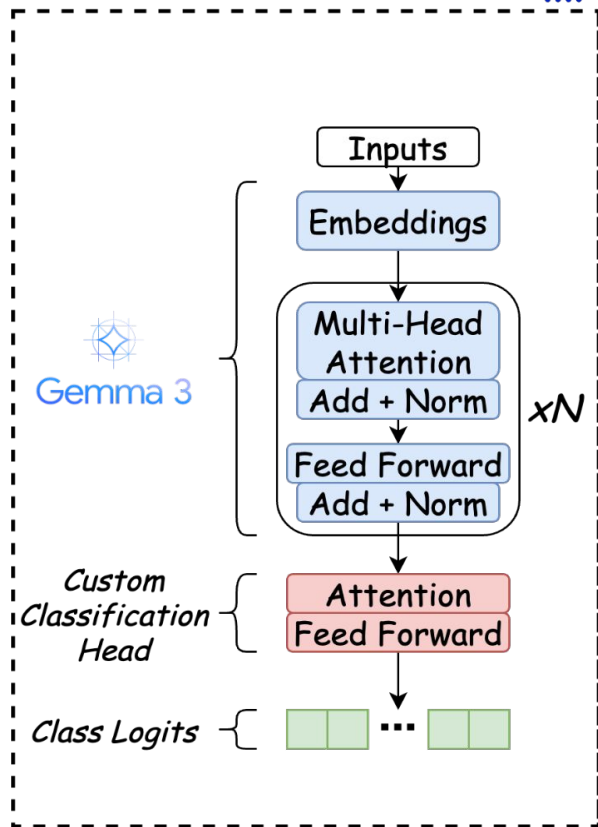
LLM Base Classifier



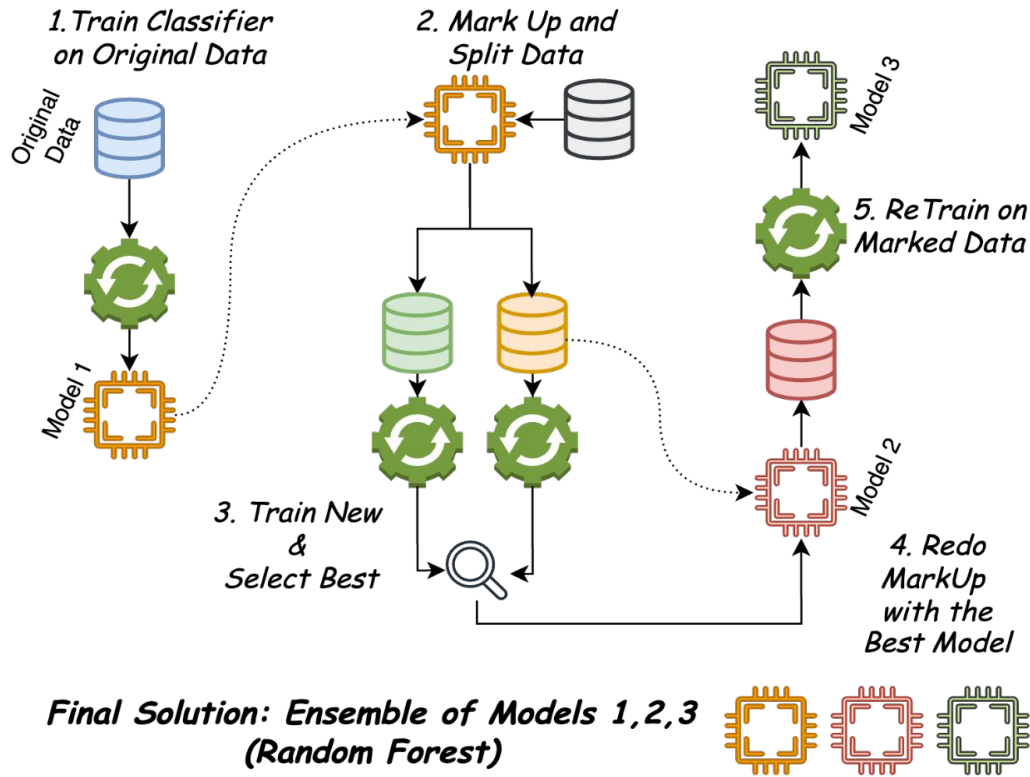
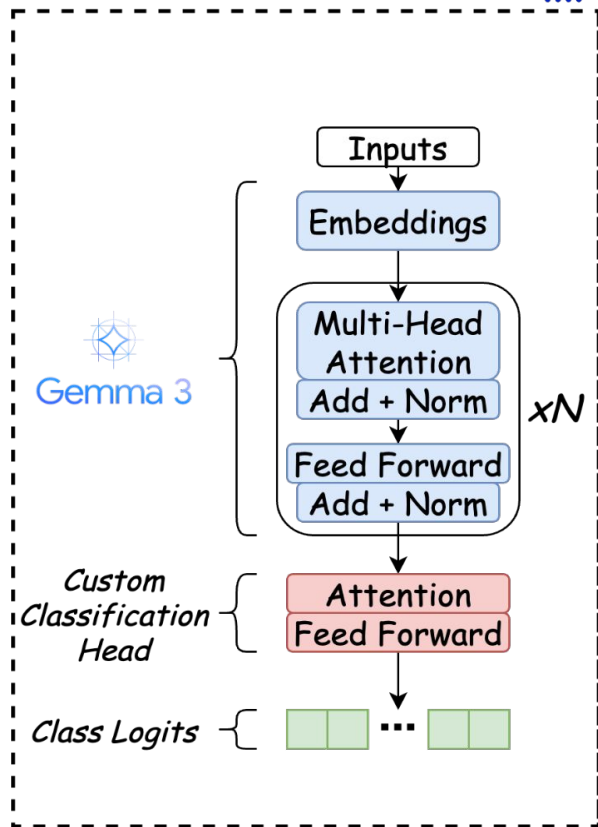
LLM Base Classifier



LLM Base Classifier



LLM Base Classifier



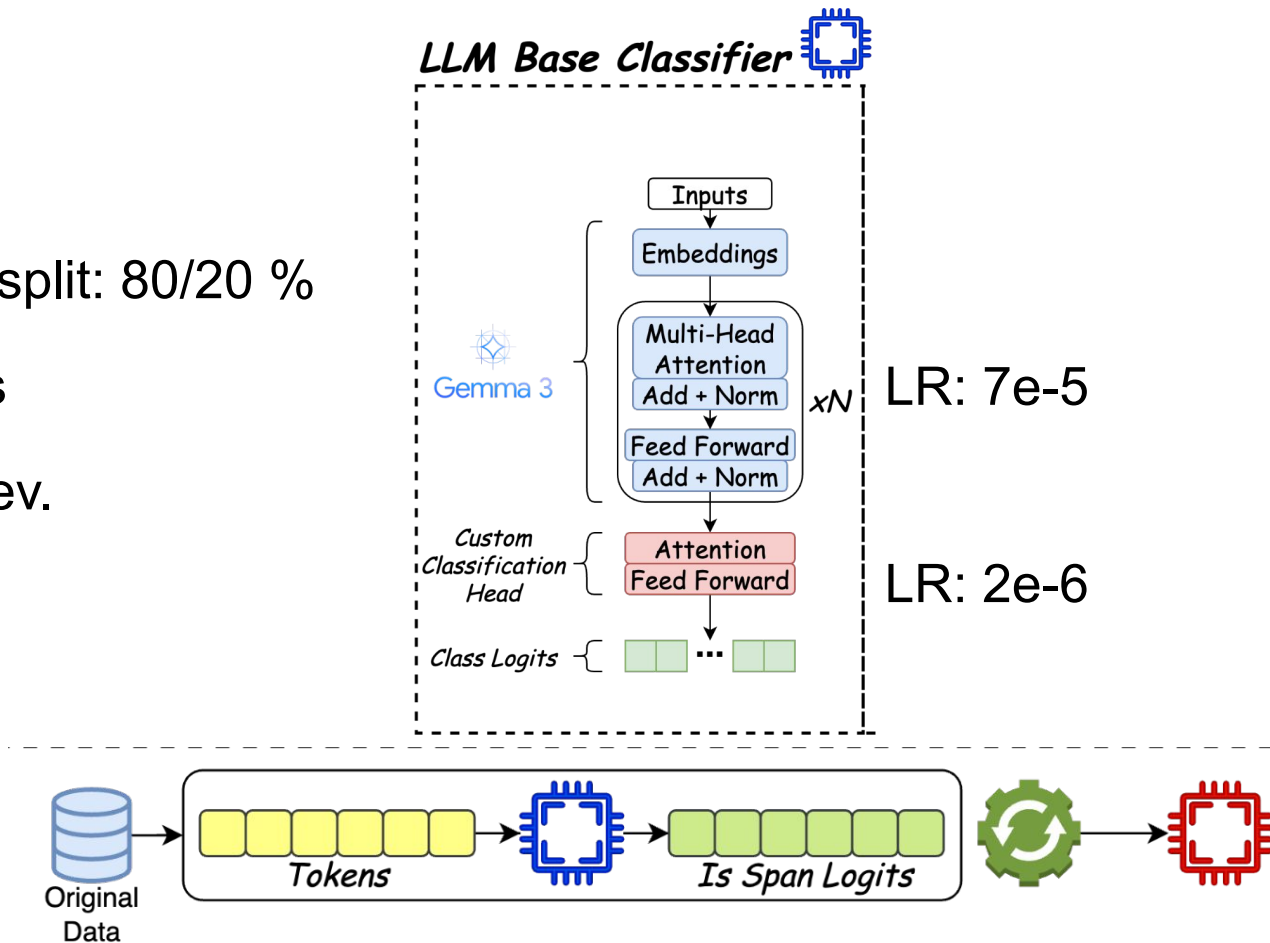
Methodology

Span Identification

Train/dev split: 80/20 %

10 epochs

Best on dev.



Technique classification

<i>Leaderboard</i>	<i>F1 Macro</i>
Public (3rd place)	0.43916
Private (4th place)	0.45265

Span Identification

<i>Leaderboard</i>	<i>Span-level F1</i>
Public (4th place)	0.59142
Private (4th place)	0.59096

Technique Classification
Private Leaderboard Results of other Experiments

Experiment	Macro F1 Score
LLaMa 3 8b	0.38870
LlaMa 3 Guard 8b	0.35896
MaxSent-BERT	0.37094
Mistral-UA 7b	0.38255
Gemma 3 12b + paraphrased	0.35228
Gemma 3 12b + generated	0.35982
Gemma 3 12b (Model 0)	0.42232
Gemma 3 12b (Model 1)	0.44754
Gemma 3 12b (Model 2)	0.44934
Gemma 3 12b (Model 3)	0.45134
Model 1 & 2 ensemble	0.45100
Model 1, 2 & 3 ensemble	0.45265

Thank you for your attention!



Code: <https://github.com/OpenBabylon/unlp2025-pub>

Contacts: *kosteugeneo@gmail.com*