My initial Evaluation Process (I did not use the evaluation script the organiser provided since there are some bugs that I cannot fixed)

Objective:

- the **retain set** achieves **high avg. ROUGE-**L score.
- the **forget set** achieves **low avg. ROUGE-**L score (successful "unlearning").

Testing Approach:
Model performance is evaluated using **50 random samples** from:

- **Forget set** (data intended to be "unlearned").
- **Retain set** (data meant to be preserved).

Training Process & Adjustments

Initial Setup:

- **Loss weights:** Retain Loss = 1, Forget Loss = 1
- **Batch size:** 16
- **Learning rate:** 0.005
- **Result:** Garbled outputs for both sets.

Batch Size Adjustment:

- **Increased batch size** to 32 (maximum before memory errors).
- **Result:** No improvement; outputs remained garbled.

Loss Weight Tuning:

- **First adjustment:** Retain Loss = 1, Forget Loss = 0.5
- **Result:** Retain set still garbled; minimal improvement.

Learning Rate Reduction:

- **Lowered learning rate** to 0.00005.
- **Adjusted weights:** Retain Loss = 1, Forget Loss = 0.3.
- **Improvement:** Retain set started producing correct answers. (I did not mark down the score)

Final Configuration:

- **Optimal loss weights:** Retain Loss = 1, Forget Loss = 0.2.

- **Batch size:** 32.

Key Result:

- **Retain set avg ROUGE-**L score significantly improved (~ 0.64)
- **Forget set get a low avg. ROUGE-**L score (~0.07)

P.S. I forgot whether **Retain Loss = 1 and Forget Loss = 0.1 or lower** was tested. (Maybe I was too tired to test at that moment since I had to wait in a long queue to train the model each time. T~T)

Validation & Training Dynamics

Validation Strategy:

- Used only the **retain validation set** (I am not using the forget validation set since I don't know how to forget something that the model has not been trained to forget)
- Tried using validation loss and **ROUGE-**L score separately as a metric
- Both have no validation improvement when the patience = 4 or more. Applied **early stopping** (patience = 4) to prevent overfitting.

Training Outcome:

- Training stopped at **500 steps** due to no validation improvement.
- **Potential limitation:** The **1B model** may have restricted baseline performance.