

NLP Semester Assignment

"Over-Tokenized Transformer: Vocabulary is Generally Worth Scaling"

Συγγραφείς paper:

**Hongzhi Huang, Defa Zhu, Banggu Wu, Yutao Zeng,
Ya Wang, Qiyang Min, Xun Zhou**

Ομάδα Φοιτηών:

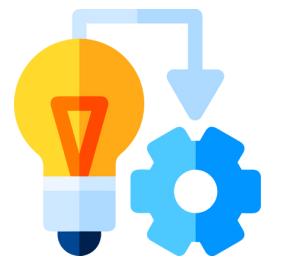
Κατσαϊδώνης Νικόλαος 03121868
Τζαμουράνης Γεώργιος 03121141
Κατσιαδράμης Κυριάκος 03121132
Φωτάκης Ανδρέας 03121100



Overview



Paper Analysis



Our Experiments



Project review

Overview



Paper Analysis



Our Experiments



Project review

Σύντομη σύνοψη του paper (1)

Κύρια ιδέα: Σκοπός του paper είναι να διερευνήσει την επίδραση του **μεγέθους του vocabulary** σε ένα γλωσσικό μοντέλο, τόσο ως προς την απόδοση όσο και ως προς το κόστος εκπαίδευσης. Επιπλέον, εξετάζεται ο **διαχωρισμός του λεξιλογίου εισόδου και εξόδου** και η επίδρασή του στους παραπάνω παράγοντες.

Σύντομη σύνοψη του paper (1)

Κύρια ιδέα: Σκοπός του paper είναι να διερευνήσει την επίδραση του **μεγέθους του vocabulary** σε ένα γλωσσικό μοντέλο, τόσο ως προς την απόδοση όσο και ως προς το κόστος εκπαίδευσης. Επιπλέον, εξετάζεται ο **διαχωρισμός του λεξιλογίου εισόδου και εξόδου** και η επίδρασή του στους παραπάνω παράγοντες.

Over encoding (ΟΕ):

1 Χρήση n-gram tokens στην είσοδο (π.χ. "cat", "cat is", "cat is sitting"). Αντί να έχουμε ένα token για κάθε λέξη, χρησιμοποιούμε ιεραρχικά n-gram tokenization και τελικά προσθέτουμε τις αναπαραστάσεις.

Σύντομη σύνοψη του paper (1)

Κύρια ιδέα: Σκοπός του paper είναι να διερευνήσει την επίδραση του **μεγέθους του vocabulary** σε ένα γλωσσικό μοντέλο, τόσο ως προς την απόδοση όσο και ως προς το κόστος εκπαίδευσης. Επιπλέον, εξετάζεται ο **διαχωρισμός του λεξιλογίου εισόδου και εξόδου** και η επίδρασή του στους παραπάνω παράγοντες.

Over encoding (OE):

1 Χρήση n-gram tokens στην είσοδο (π.χ. "cat", "cat is", "cat is sitting"). Αντί να έχουμε ένα token για κάθε λέξη, χρησιμοποιούμε ιεραρχικά n-gram tokenization και τελικά προσθέτουμε τις αναπαραστάσεις.

Over decoding (OD):

2 Μεγαλύτερο output vocabulary. Παραγωγή multi-token εξόδου (π.χ "on the sofa" αντί για 3 λέξεις ξεχωριστά)

Σύντομη σύνοψη του paper (1)

Κύρια ιδέα: Σκοπός του paper είναι να διερευνήσει την επίδραση του **μεγέθους του vocabulary** σε ένα γλωσσικό μοντέλο, τόσο ως προς την απόδοση όσο και ως προς το κόστος εκπαίδευσης. Επιπλέον, εξετάζεται ο **διαχωρισμός του λεξιλογίου εισόδου και εξόδου** και η επίδρασή του στους παραπάνω παράγοντες.

Over encoding (OE):

1 Χρήση n-gram tokens στην είσοδο (π.χ. "cat", "cat is", "cat is sitting"). Αντί να έχουμε ένα token για κάθε λέξη, χρησιμοποιούμε ιεραρχικά n-gram tokenization και τελικά προσθέτουμε τις αναπαραστάσεις.

Over decoding (OD):

2 Μεγαλύτερο output vocabulary. Παραγωγή multi-token εξόδου (π.χ "on the sofa" αντί για 3 λέξεις ξεχωριστά)

3 **Over-Tokenized Transformer(OTT):**
Συνδιασμός OE & OD για εκμάθηση πιο σύνθετων εξαρτήσεων.

Σύντομη σύνοψη του paper (2)

Σύντομη σύνοψη του paper (2)

Πείραμα 1 – Συνθετικό CFG Dataset (GPT-2):

- Training με χρήση ΟΕ (3-gram input tokens).
- Βελτιωμένο **accuracy** & μικρότερο **loss** με μεγαλύτερο input vocabulary.
- ΟD βλάπτει μικρά μοντέλα.

Σύντομη σύνοψη του paper (2)

Πείραμα 1 – Συνθετικό CFG Dataset (GPT-2):

- Training με χρήση ΟΕ (3-gram input tokens).
- Βελτιωμένο accuracy & μικρότερο loss με μεγαλύτερο input vocabulary.
- OD βλάπτει μικρά μοντέλα.

Πείραμα 2 – Downstream Tasks (PIQA κ.ά.):

- Δοκιμή σε μοντέλα με 151M – 1B παραμέτρους.
- ΟΕ-400M → Ίδιο training cost με baseline 1B, καλύτερο performance.
- Έως και 3.9x ταχύτερη σύγκλιση, +1.3% accuracy με ΟΕ+OD

Σύντομη σύνοψη του paper (2)

Πείραμα 1 – Συνθετικό CFG Dataset (GPT-2):

- Training με χρήση ΟΕ (3-gram input tokens).
- Βελτιωμένο accuracy & μικρότερο loss με μεγαλύτερο input vocabulary.
- OD βλάπτει μικρά μοντέλα.

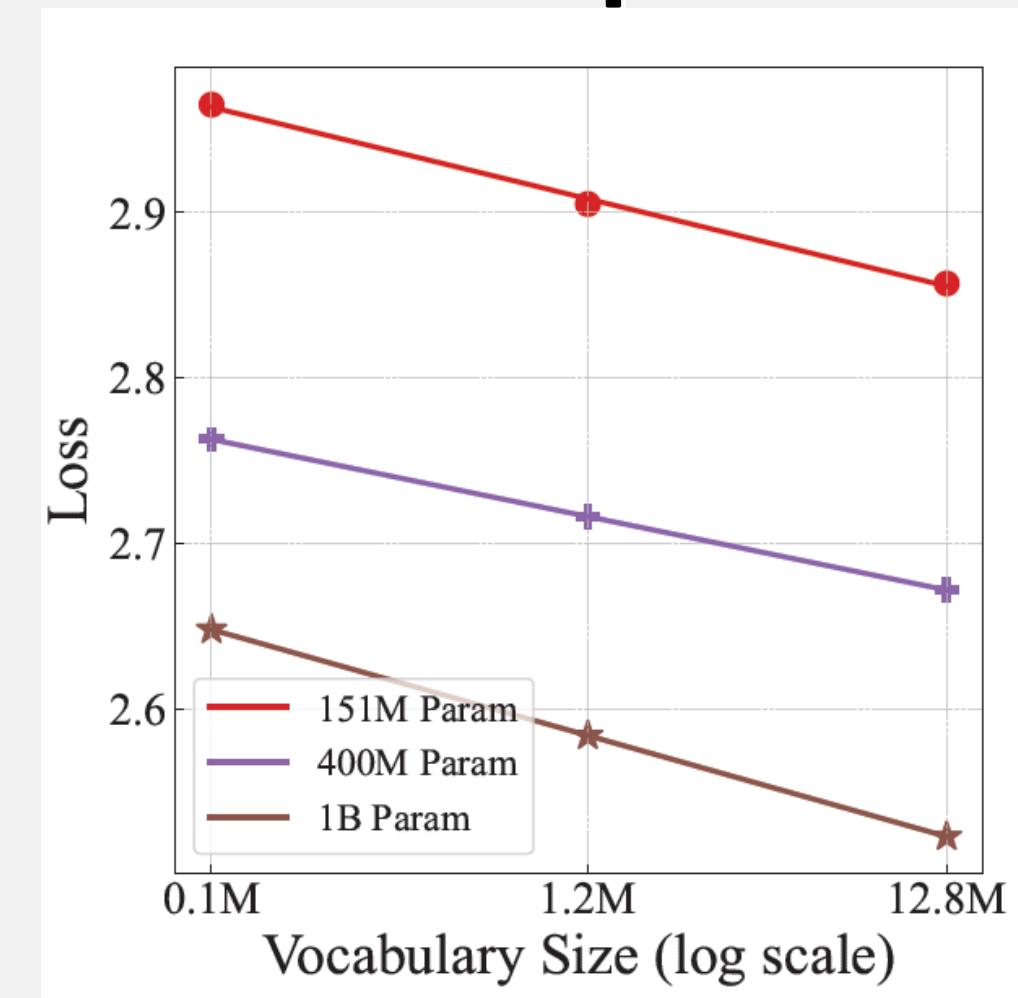
Πείραμα 2 – Downstream Tasks (PIQA κ.ά.):

- Δοκιμή σε μοντέλα με 151M – 1B παραμέτρους.
- ΟΕ-400M → Ίδιο training cost με baseline 1B, καλύτερο performance.
- Έως και 3.9x ταχύτερη σύγκλιση, +1.3% accuracy με ΟΕ+OD

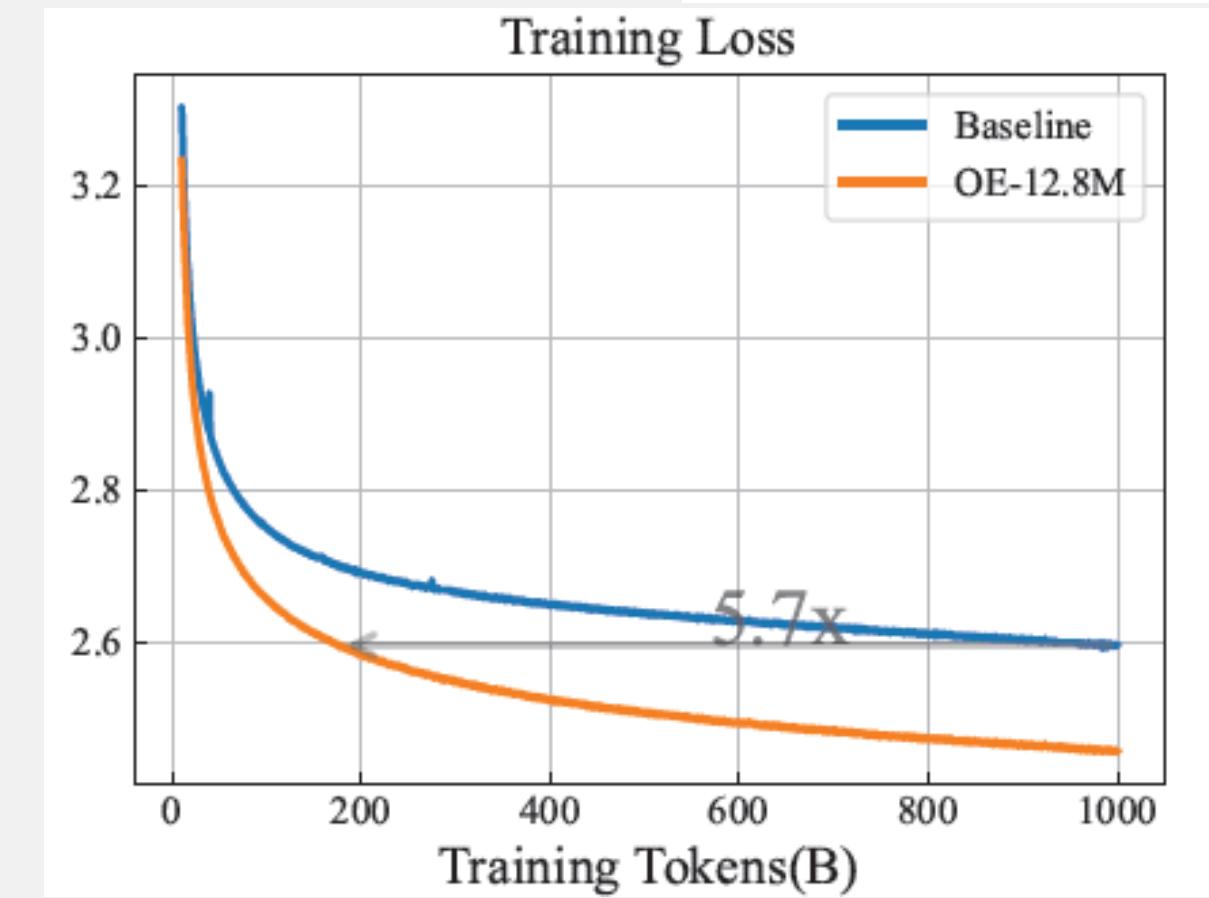
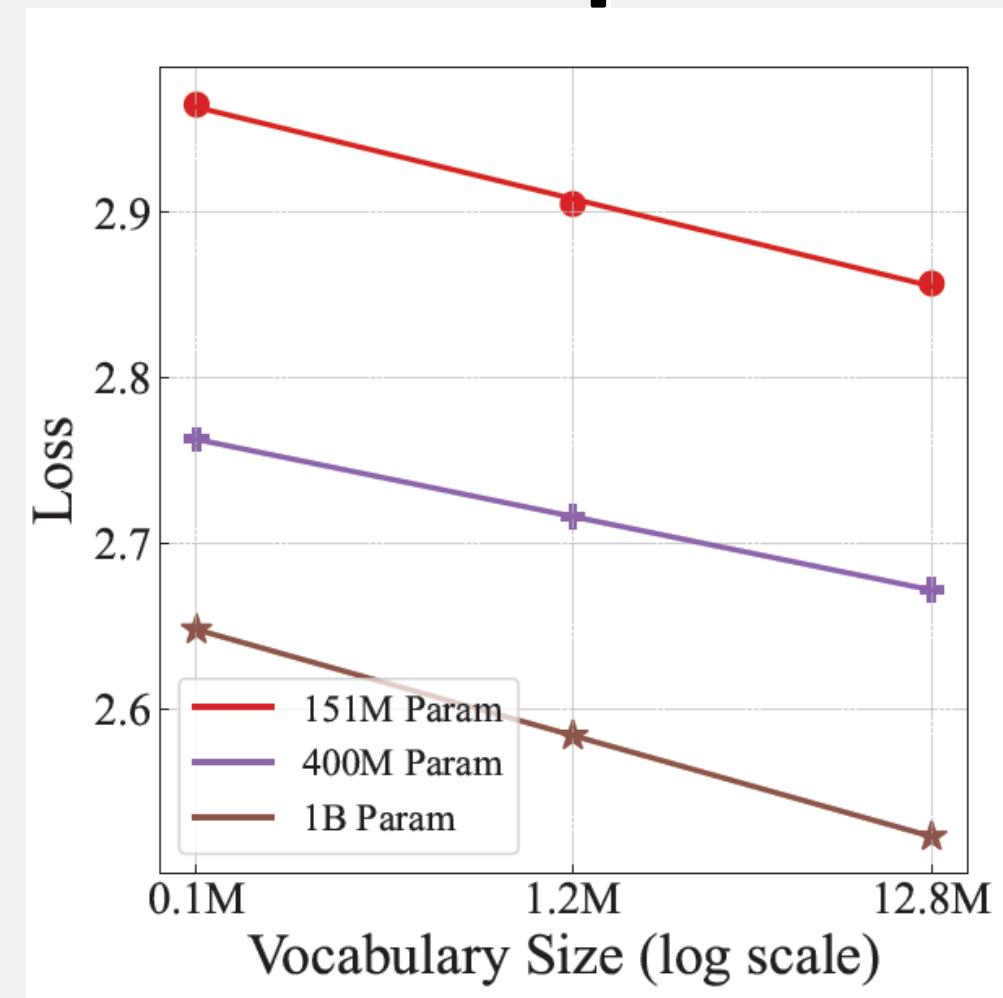
Ευρήματα:

- Μεγαλύτερο input λεξιλόγιο (μέσω multi-gram embeddings) βελτιώνει σταθερά την απόδοση, και μειώνει το loss ανεξαρτήτως μεγέθους μοντέλου (log-linear σχέση μεταξύ του vocabulary size και του loss).
- Μεγαλύτερο output λεξιλόγιο μπορεί να βλάψει μικρότερα μοντέλα.

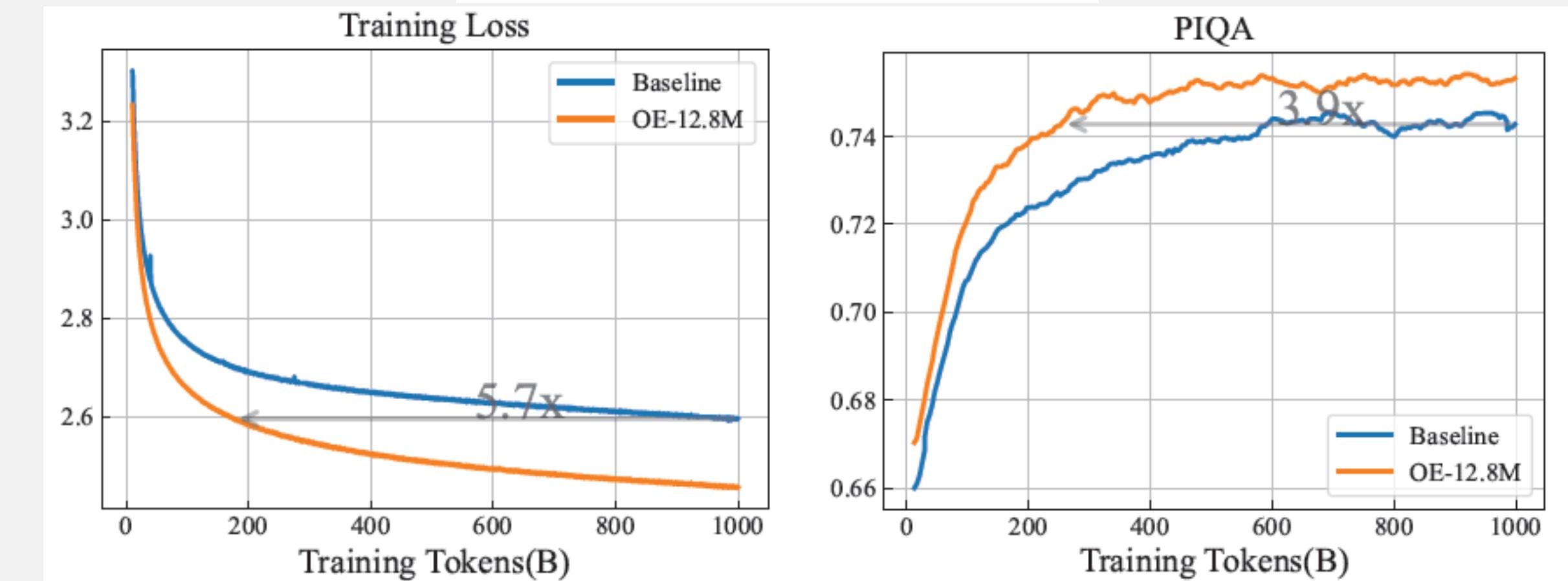
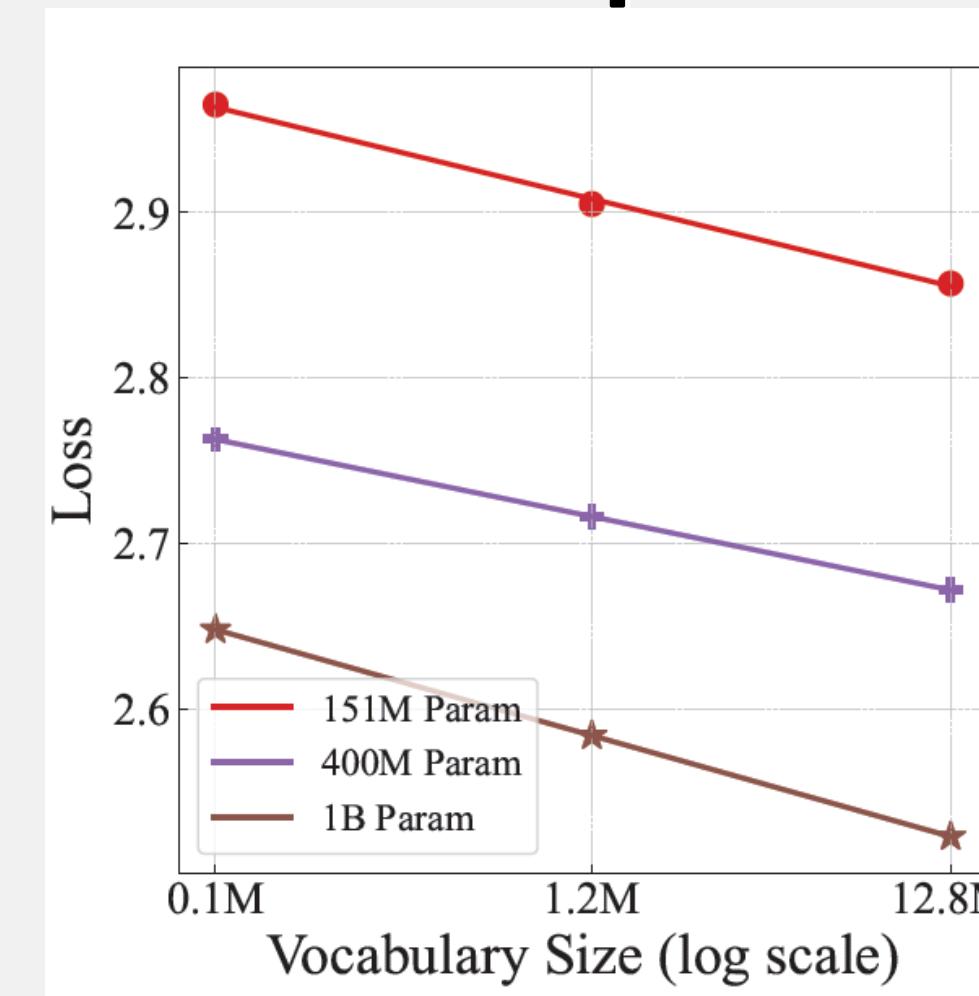
Μερικά αποτελέσματα του papaer



Μερικά αποτελέσματα του papaer



Μερικά αποτελέσματα του papaer



Overview



Paper Analysis



Our Experiments



Project review

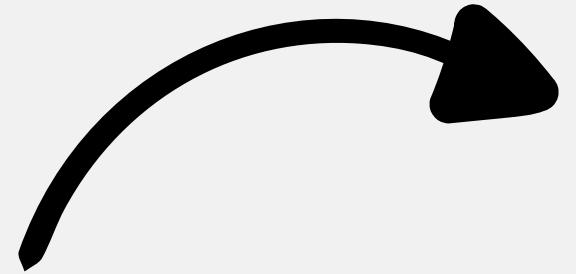
Πείραμα 1ο: CFG dataset (1)

Πείραμα 1ο: CFG dataset (1)

Training με **GPT-2** (2.4M params) ως
Baseline μοντέλο χωρίς over encoding και
καταμέτρηση **Loss** και **Generation**
Accuracy ανά εποχή.

Πείραμα 1ο: CFG dataset (1)

Training με **GPT-2** (2.4M params) ως **Baseline** μοντέλο χωρίς over encoding και καταμέτρηση **Loss** και **Generation Accuracy** ανά εποχή.



Δοκιμή **OE-m** μοντέλου και σύγκριση Loss και Generation Accuracy με τα αντίστοιχα μεγέθη του Baseline.

Πείραμα 1ο: CFG dataset (1)

Training με **GPT-2** (2.4M params) ως **Baseline** μοντέλο χωρίς over encoding και καταμέτρηση **Loss** και **Generation Accuracy** ανά εποχή.

Δοκιμή **OE-m** μοντέλου και σύγκριση Loss και Generation Accuracy με τα αντίστοιχα μεγέθη του Baseline.

Αύξηση του m και σύγκριση των παραπάνω μεγεθών για διάφορα OE-m μοντέλα.

Πείραμα 1ο: CFG dataset (1)

Training με **GPT-2** (2.4M params) ως **Baseline** μοντέλο χωρίς over encoding και καταμέτρηση **Loss** και **Generation Accuracy** ανά εποχή.

Δοκιμή **OE-m** μοντέλου και σύγκριση Loss και Generation Accuracy με τα αντίστοιχα μεγέθη του Baseline.

Εξαγωγή αντίστοιχων **διαγραμμάτων**.

Αύξηση του m και σύγκριση των παραπάνω μεγεθών για διάφορα OE-m μοντέλα.

Πείραμα 1ο: CFG dataset (1)

Training με **GPT-2** (2.4M params) ως **Baseline** μοντέλο χωρίς over encoding και καταμέτρηση **Loss** και **Generation Accuracy** ανά εποχή.

δοκιμή **OE-m** μοντέλου και σύγκριση Loss και Generation Accuracy με τα αντίστοιχα μεγέθη του Baseline.

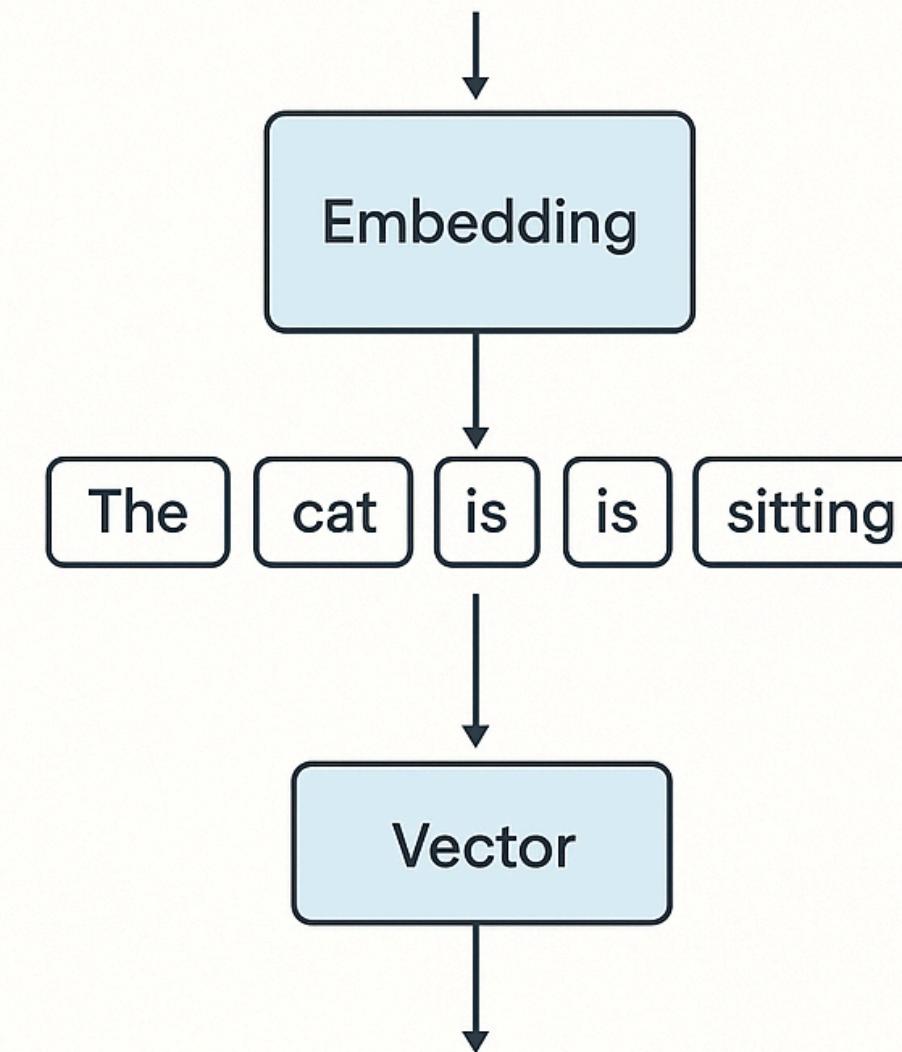
Σημείωση: Το vocabulary μας είναι **{1,2,3,<PAD>}** οπότε έχουμε **4 1-grams, 4² 2-grams και 4³ 3-grams** και συνεπώς δεν μπορούμε να κάνουμε λογαριθμική αύξηση του m αλλά απλώς αύξηση για να δούμε πως αυτό επηρεάζει την επίδοση και το loss.

Εξαγωγή αντίστοιχων **διαγραμμάτων**.

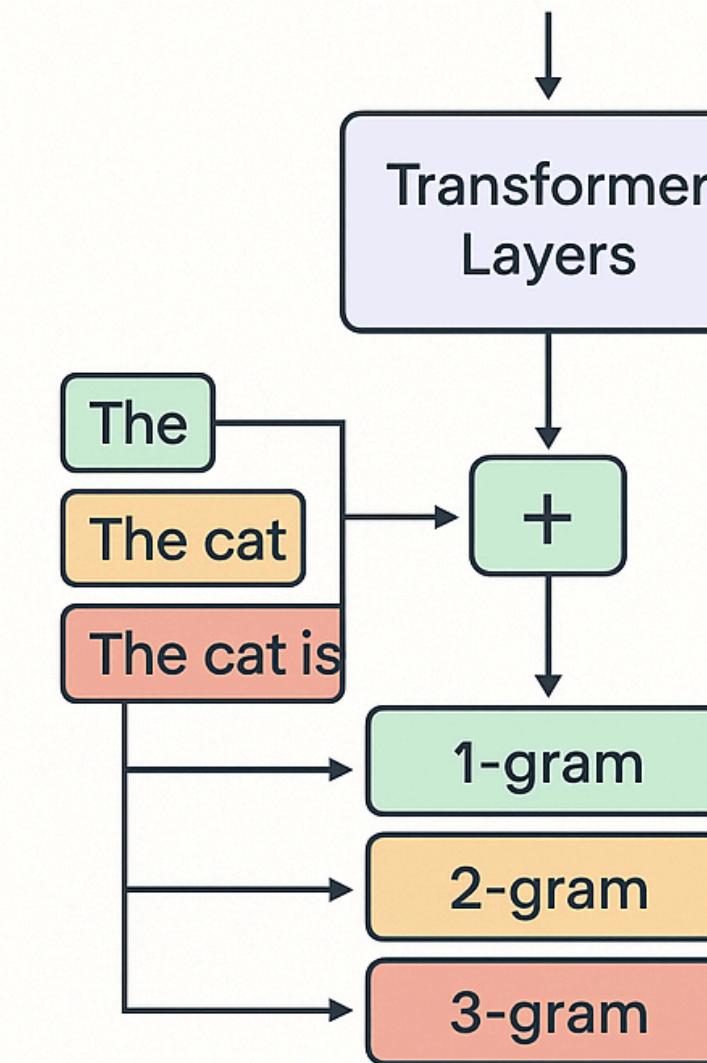
Αύξηση του m και σύγκριση των παραπάνω μεγεθών για διάφορα OE-m μοντέλα.

Σχηματική αναπαράσταση

Κλασικό Embedding Layer (Simple Lookup)



Embedding Layer με Over-Encoding (OE) (Context-Aware)



Πείραμα 1ο: Datasets

Κατασκευή ενός **συνθετικού dataset με vocabulary = {1,2,3}** χρήση της context-free grammar με τους παρακάτω κανόνες:

Πείραμα 1ο: Datasets

Κατασκευή ενός **συνθετικού dataset με vocabulary = {1,2,3}** χρήση της context-free grammar με τους παρακάτω κανόνες:

root -> 20 21	19 -> 18 16 18	16 -> 15 15	13 -> 11 12	10 -> 8 9 9	7 -> 2 2 1	<i>an example sentence</i>	332213123312113123211322312312111213211322311311
root -> 20 19 21	19 -> 17 18	16 -> 13 15 13	13 -> 12 11 12	10 -> 9 7 9	7 -> 3 2 2		32233312312111213113311213212133331232212131232
root -> 21 19 19	19 -> 18 18	16 -> 14 13	13 -> 10 12 11	10 -> 7 9 9	7 -> 3 1 2		22111121332213113113111113231233133133311331
root -> 20 20	20 -> 16 16	16 -> 14 14	14 -> 10 12	11 -> 8 8	7 -> 3 2		33333223121131112122111211233312331121113313333
	20 -> 16 17	17 -> 15 14 13	14 -> 12 10 12	11 -> 9 7	8 -> 3 1 1		331123333131111333121132113121211333321211121
	20 -> 17 16 18	17 -> 14 15	14 -> 12 11	11 -> 9 7 7	8 -> 1 2		213223223322133221113221132323313111213223223221
	21 -> 18 17	17 -> 15 14	14 -> 10 12 12	12 -> 7 9 7	8 -> 3 3 1		211133331121322221332211212133121331332212213221
	21 -> 17 16	18 -> 14 15 13	15 -> 10 11 11	12 -> 9 8	9 -> 1 2 1		211213331232233312
	21 -> 16 17 18	18 -> 15 13 13	15 -> 11 11 10	12 -> 8 8 9	9 -> 3 3		
	21 -> 16 18	18 -> 13 15	15 -> 10 10		9 -> 1 1		
			15 -> 12 12 11				

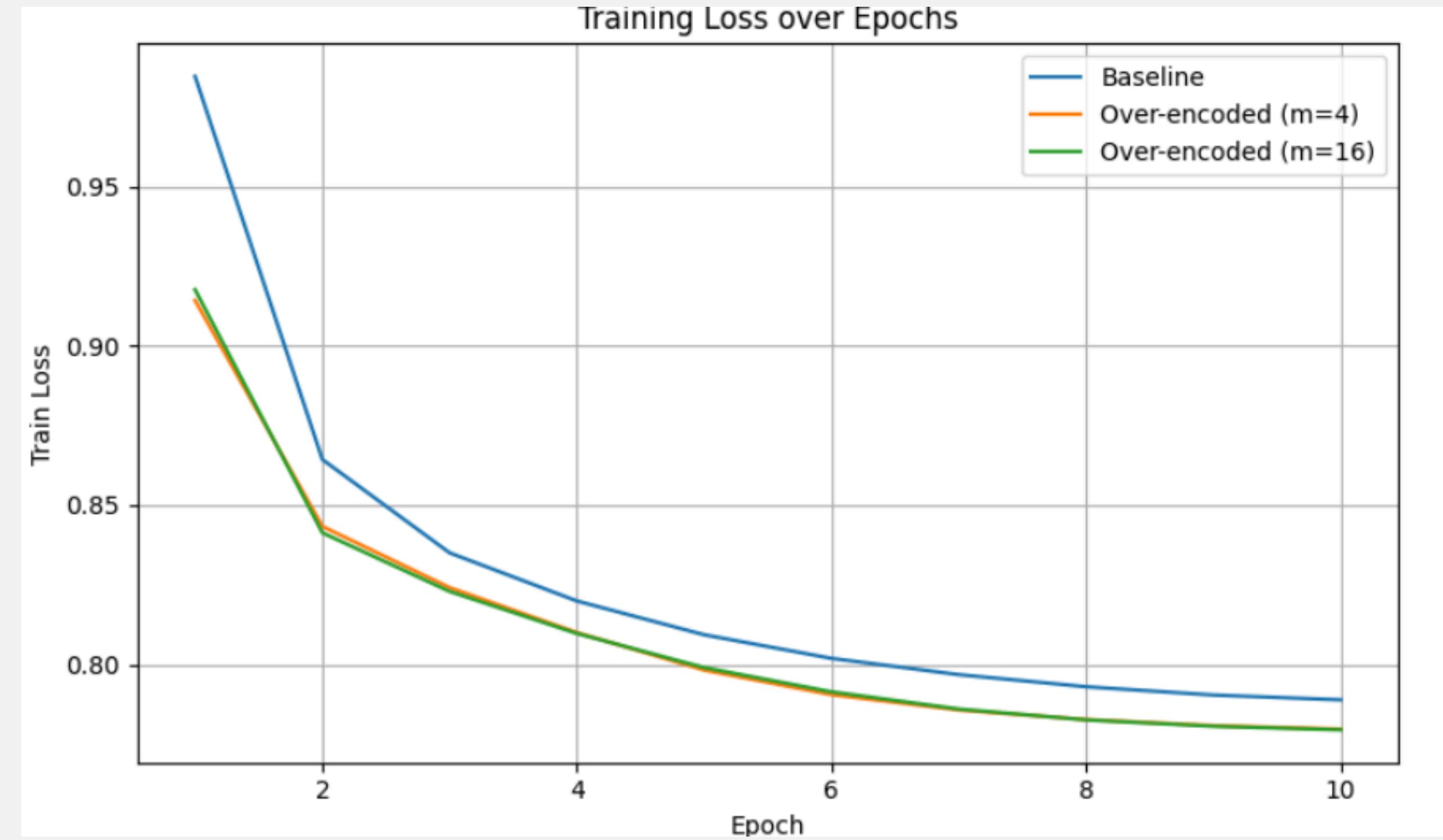
Πείραμα 1ο: Datasets

Κατασκευή ενός **συνθετικού dataset με vocabulary = {1,2,3}** χρήση της context-free grammar με τους παρακάτω κανόνες:

root -> 20 21	19 -> 18 16 18	16 -> 15 15	13 -> 11 12	10 -> 8 9 9	7 -> 2 2 1	<i>an example sentence</i>	332213123312113123211322312312111213211322311311
root -> 20 19 21	19 -> 17 18	16 -> 13 15 13	13 -> 12 11 12	10 -> 9 7 9	7 -> 3 2 2		32233312312111213113311213212133331232212131232
root -> 21 19 19	19 -> 18 18	16 -> 14 13	13 -> 10 12 11	10 -> 7 9 9	7 -> 3 1 2		22111121332213113113111113231233133133311331
root -> 20 20	20 -> 16 16	16 -> 14 14	14 -> 10 12	11 -> 8 8	7 -> 3 2		33333223121131112122111211233312331121113313333
	20 -> 16 17	17 -> 15 14 13	14 -> 12 10 12	11 -> 9 7	8 -> 3 1 1		3311233331311113333121132113121211333321211121
20 -> 17 16 18	17 -> 14 15	14 -> 12 11	11 -> 9 7 7	8 -> 1 2			21322322332213322113221132323313111213223223221
	21 -> 18 17	17 -> 15 14	14 -> 10 12 12	12 -> 7 9 7	8 -> 3 3 1		211133331121322221332211212133121331332212213221
21 -> 17 16	18 -> 14 15 13	15 -> 10 11 11	12 -> 9 8	9 -> 1 2 1			211213331232233312
	21 -> 16 17 18	18 -> 15 13 13	15 -> 11 11 10	12 -> 8 8 9	9 -> 3 3		
21 -> 16 18	18 -> 13 15	15 -> 10 10			9 -> 1 1		
			15 -> 12 12 11				

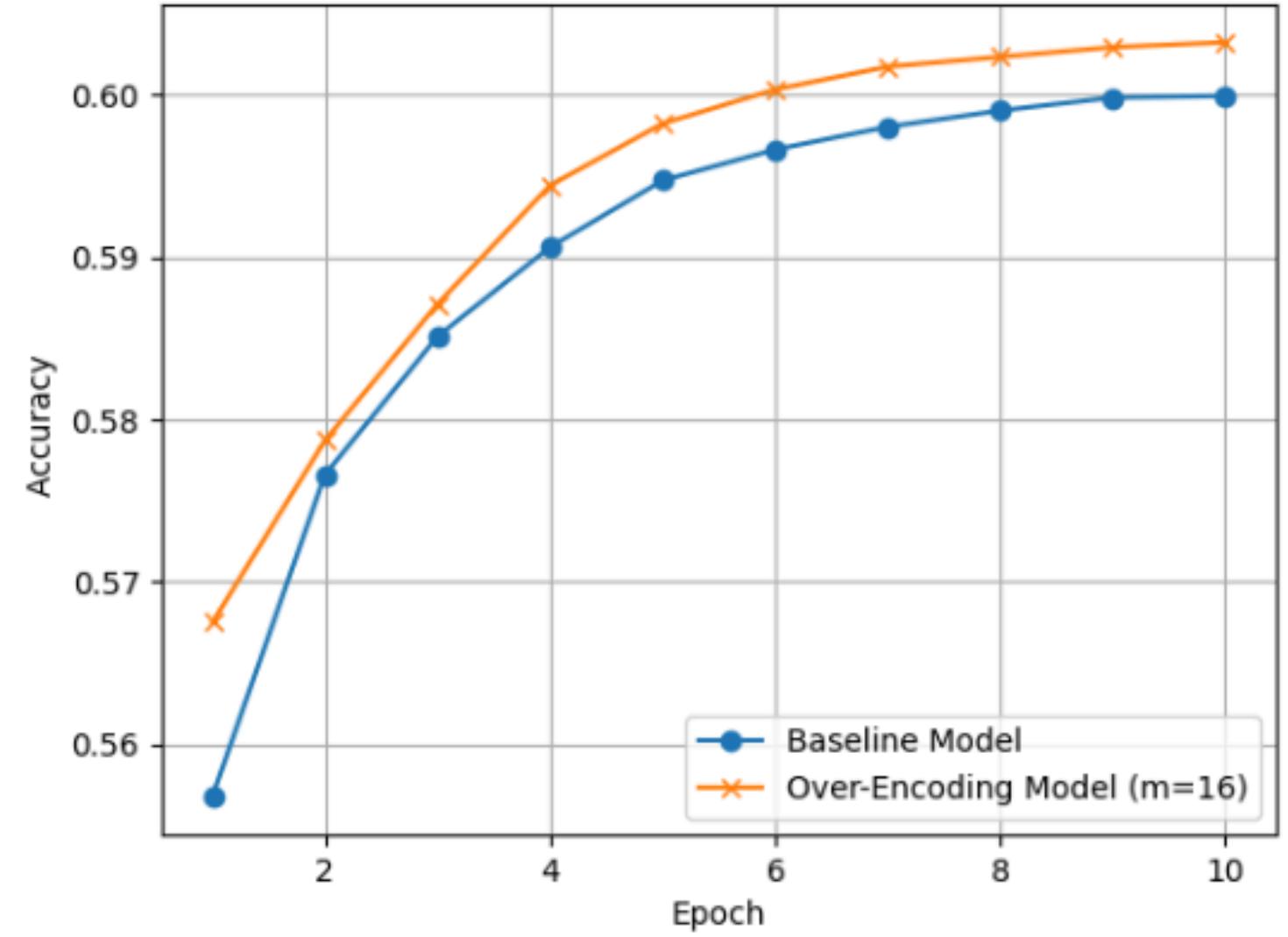
- **Συνολικό dataset:** 200.000 ακολουθίες μήκους έως 729 χαρακτήρες.
- **Train/val split:** 160.000 training sequences, 40.000 validation sequences.

Αποτελέσματα (1)

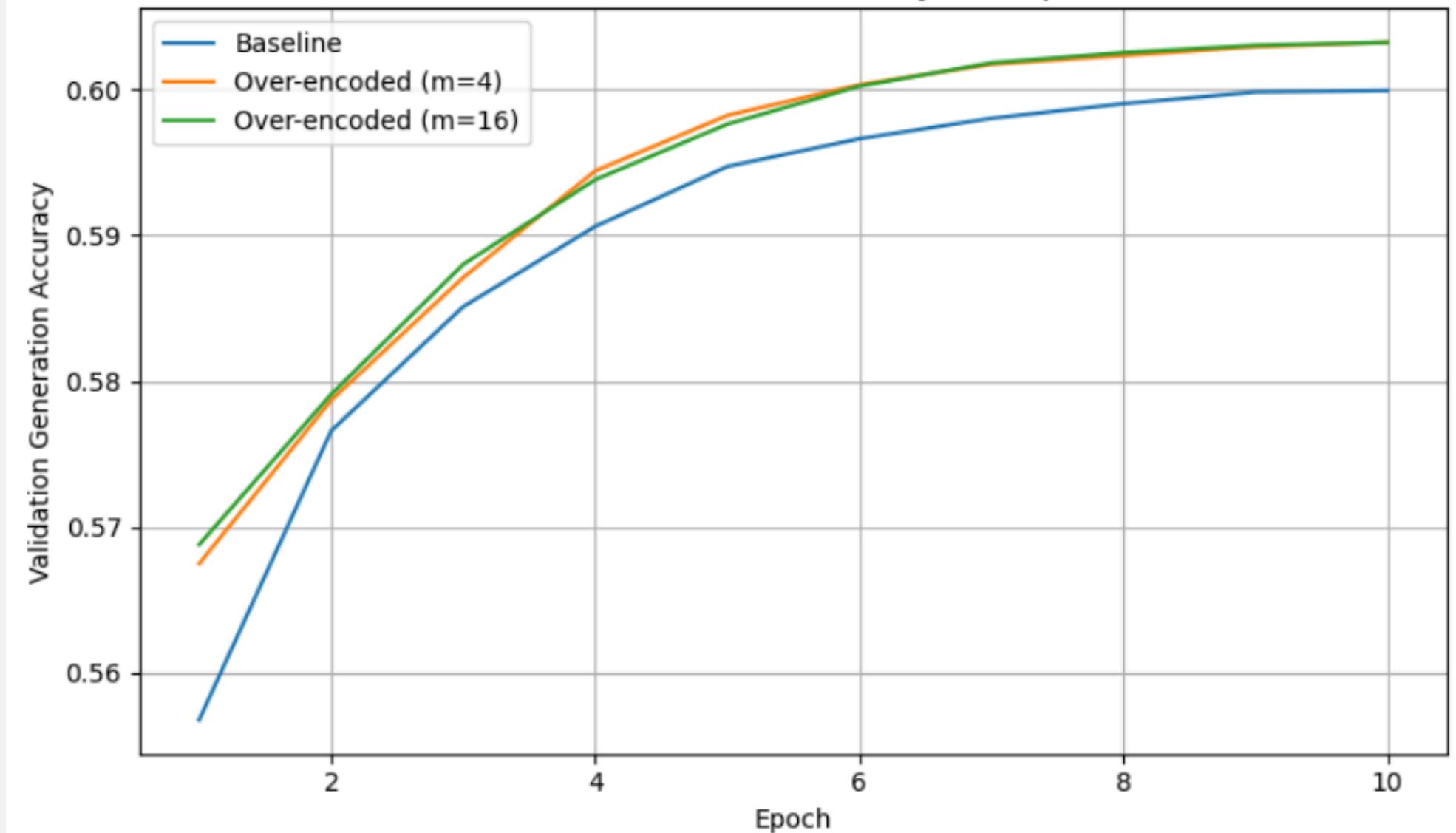


Αποτελέσματα (2)

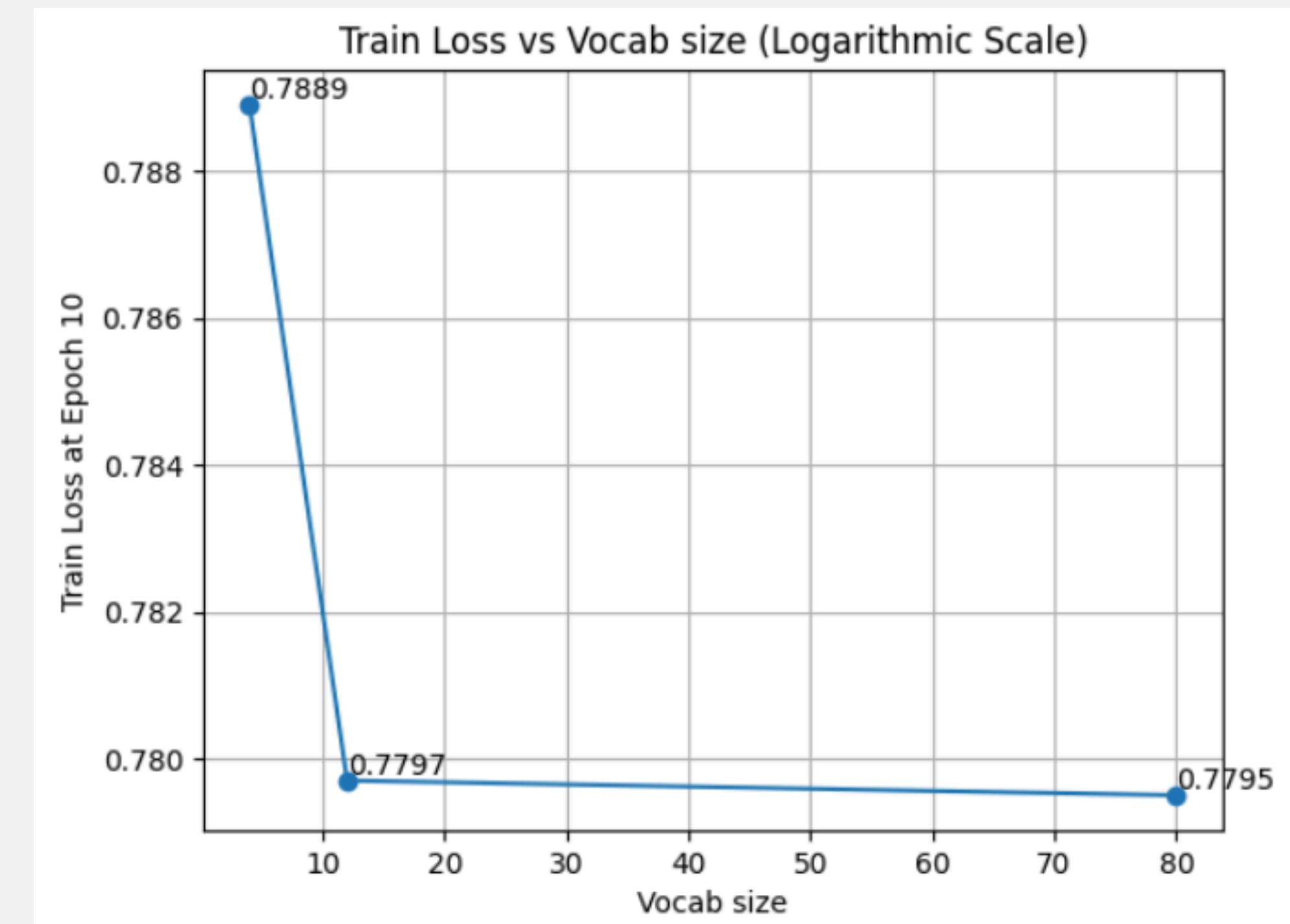
Accuracy: Baseline VS Over encoding model



Validation Generation Accuracy over Epochs



Αποτελέσματα (3)



Είναι προφανές πως για το τόσο μικρό μέγεθος λεξιλογίου η μειωση είναι πολύ μικρής τάξης καθώς δεν είναι εφικτή η λογαριθμική αύξηση του m σε τόσο μικρό λεξιλόγιο.

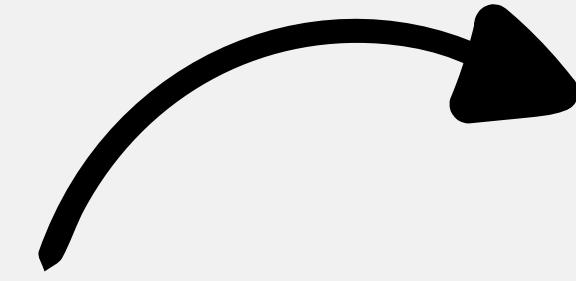
Πείραμα 2ο: Downstream tasks (Training)

Πείραμα 2ο: Downstream tasks (Training)

Εύρεση συνηθέστερων n-gram
συνδιασμών σε γενικό dataset
κειμένου (**Wikitext**).

Πείραμα 2ο: Downstream tasks (Training)

Εύρεση συνηθέστερων n-gram
συνδιασμών σε γενικό dataset
κειμένου (**Wikitext**).



Φόρτωση pretrained GPT-2 και
freeze των παραμέτρων του ώστε να
μη χάσει την γενική γνώση που έχει
αποκτήσει.

Πείραμα 2ο: Downstream tasks (Training)

Εύρεση συνηθέστερων n-gram συνδιασμών σε γενικό dataset κειμένου (**Wikitext**).

Φόρτωση pretrained GPT-2 και freeze των παραμέτρων του ώστε να μη χάσει την γενική γνώση που έχει αποκτήσει.

Επέκταση του embedding table με τους n-gram συνδιασμούς οι οποίοι αρχικοποιούνται στον μέσο όρο των επιμέρους embeddings.

Πείραμα 2ο: Downstream tasks (Training)

Εύρεση συνηθέστερων n-gram συνδιασμών σε γενικό dataset κειμένου (**Wikitext**).

Φόρτωση pretrained GPT-2 και freeze των παραμέτρων του ώστε να μη χάσει την γενική γνώση που έχει αποκτήσει.

Λογαριθμική αύξηση λεξιλογίου-ngrams και υπολογισμός training loss.

Training πάνω στα n-gram tokens διατηρώντας τα 1-gram tokens ίδια με αυτά του pretrained GPT-2.

Επέκταση του embedding table με τους n-gram συνδιασμούς οι οποίοι αρχικοποιούνται στον μέσο όρο των επιμέρους embeddings.

Πείραμα 2ο: Downstream tasks (Testing)

Πείραμα 2ο: Downstream tasks (Testing)

>HellaSwag dataset

- Testing του μοντέλου στο task του HellasSwag: **Ένα κείμενο ως input** (context) και **4 πιθανές συνέχειες**, μία από της οποίες είναι λογική.

Πείραμα 2ο: Downstream tasks (Testing)

>HellaSwag dataset

- Testing του μοντέλου στο task του HellasSwag: **Ένα κείμενο ως input** (context) και **4 πιθανές συνέχειες**, μία από της οποίες είναι λογική.

>PIQA dataset

- Testing του μοντέλου στο task του PIQA: **Μια ερώτηση** ως input και **2 πιθανές απαντήσεις**, μία σωστή.

Πείραμα 2ο: Downstream tasks (Testing)

> HellaSwag dataset

- Testing του μοντέλου στο task του HellasSwag: **Ένα κείμενο ως input** (context) και **4 πιθανές συνέχειες**, μία από της οποίες είναι λογική.

> PIQA dataset

- Testing του μοντέλου στο task του PIQA: **Μια ερώτηση** ως input και **2 πιθανές απαντήσεις**, μία σωστή.

> Story cloze dataset

- Testing του μοντέλου στο task του story cloze: **Input 4 προτάσεις context** που σχηματίζουν μία ιστορία και 2 για ending της ιστορίας.

Πείραμα 2ο: Downstream tasks (Testing)

>HellaSwag dataset

- Testing του μοντέλου στο task του HellasSwag: **Ένα κείμενο ως input** (context) και **4 πιθανές συνέχειες**, μία από της οποίες είναι λογική.

>PIQA dataset

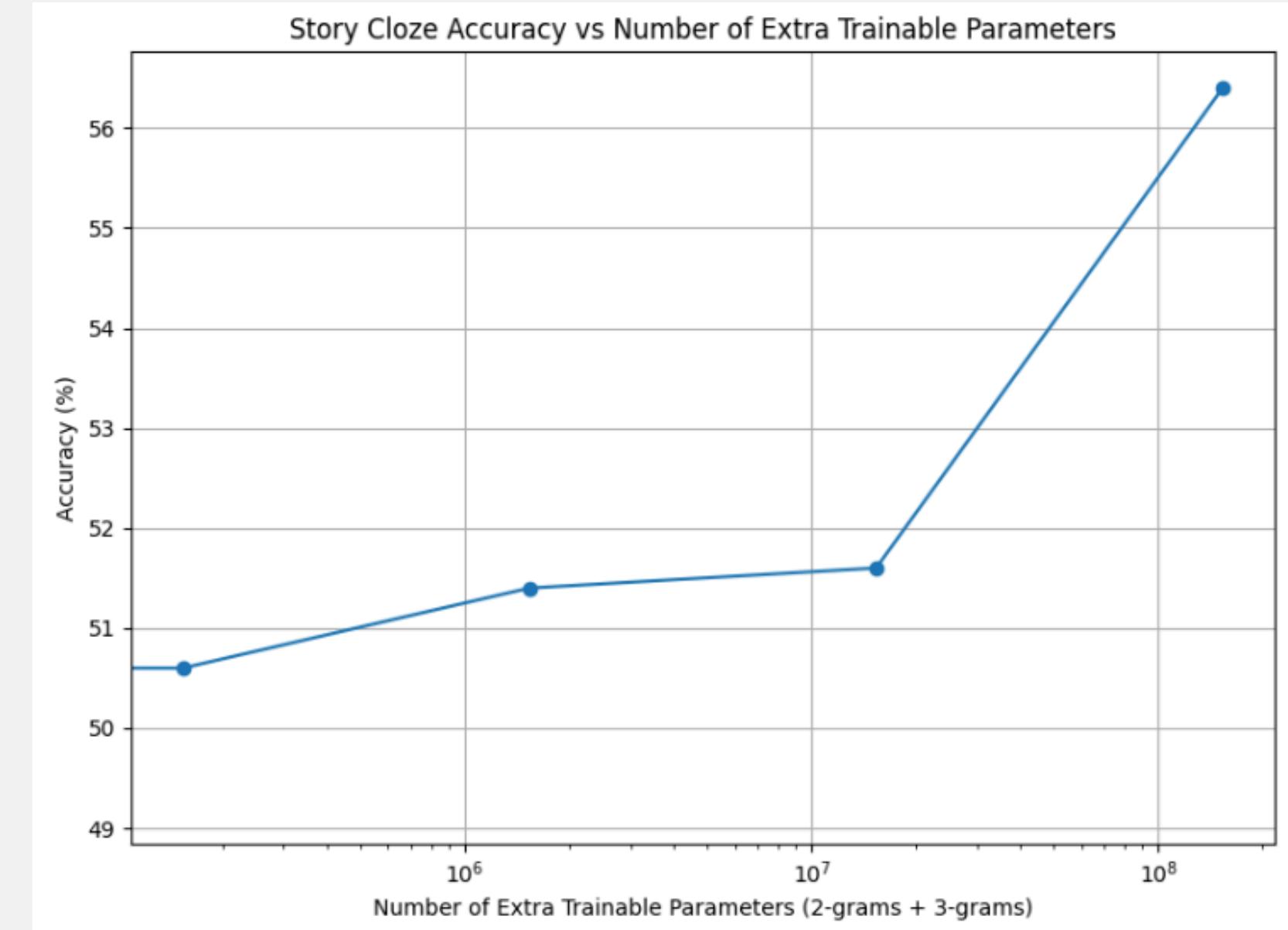
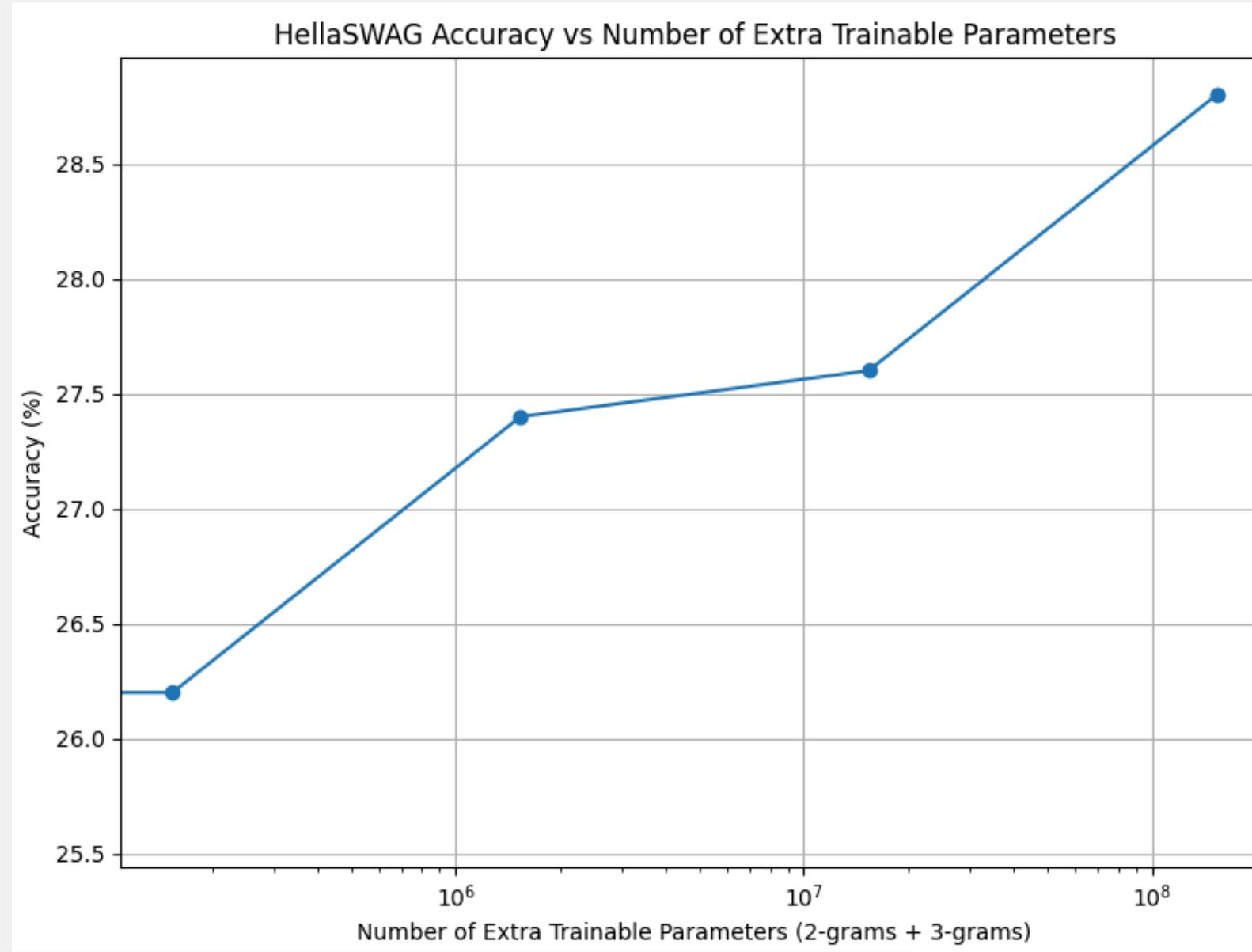
- Testing του μοντέλου στο task του PIQA: **Μια ερώτηση** ως input και **2 πιθανές απαντήσεις**, μία σωστή.

>Story cloze dataset

- Testing του μοντέλου στο task του story cloze: **Input 4 προτάσεις context** που σχηματίζουν μία ιστορία και 2 για ending της ιστορίας.

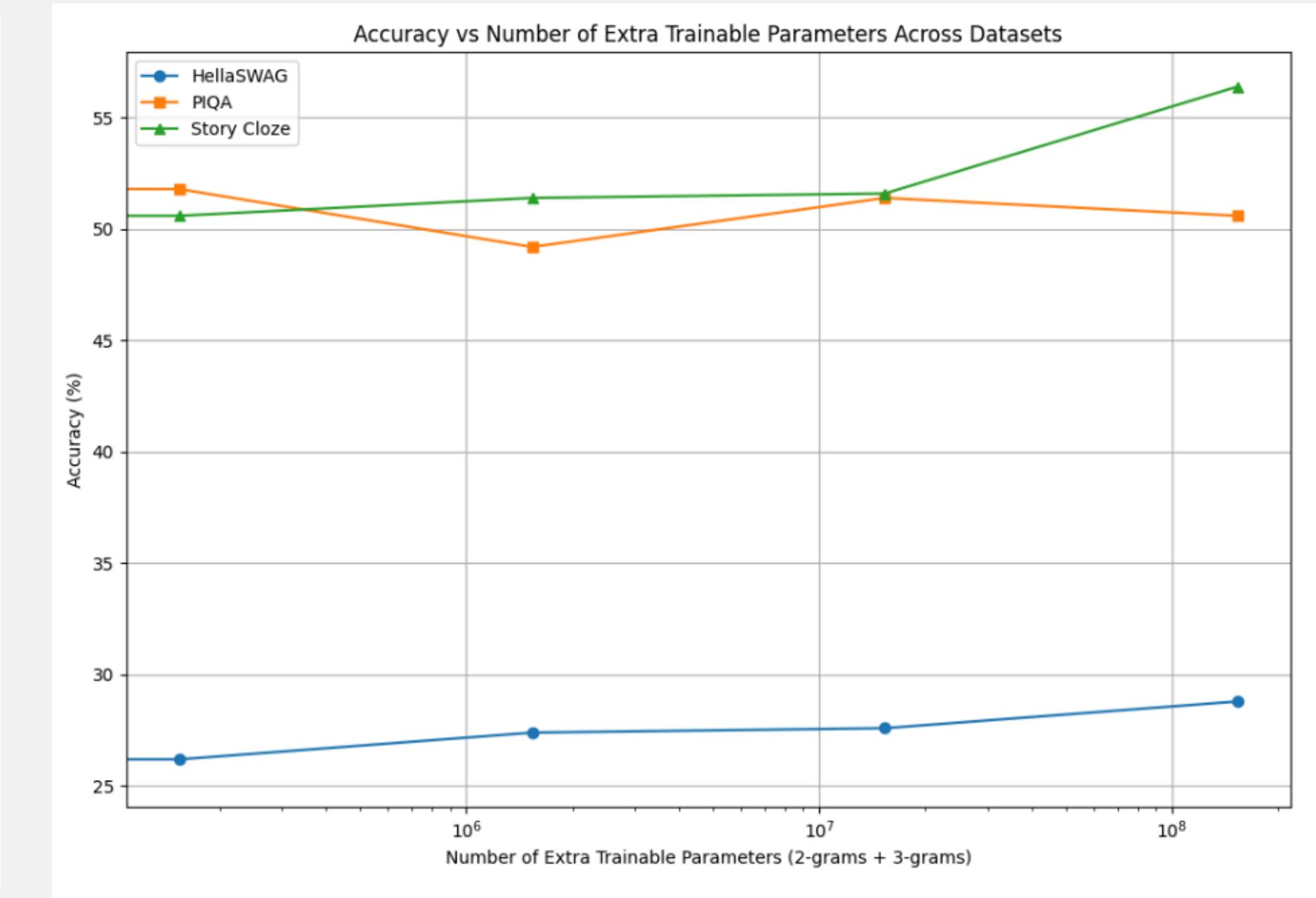
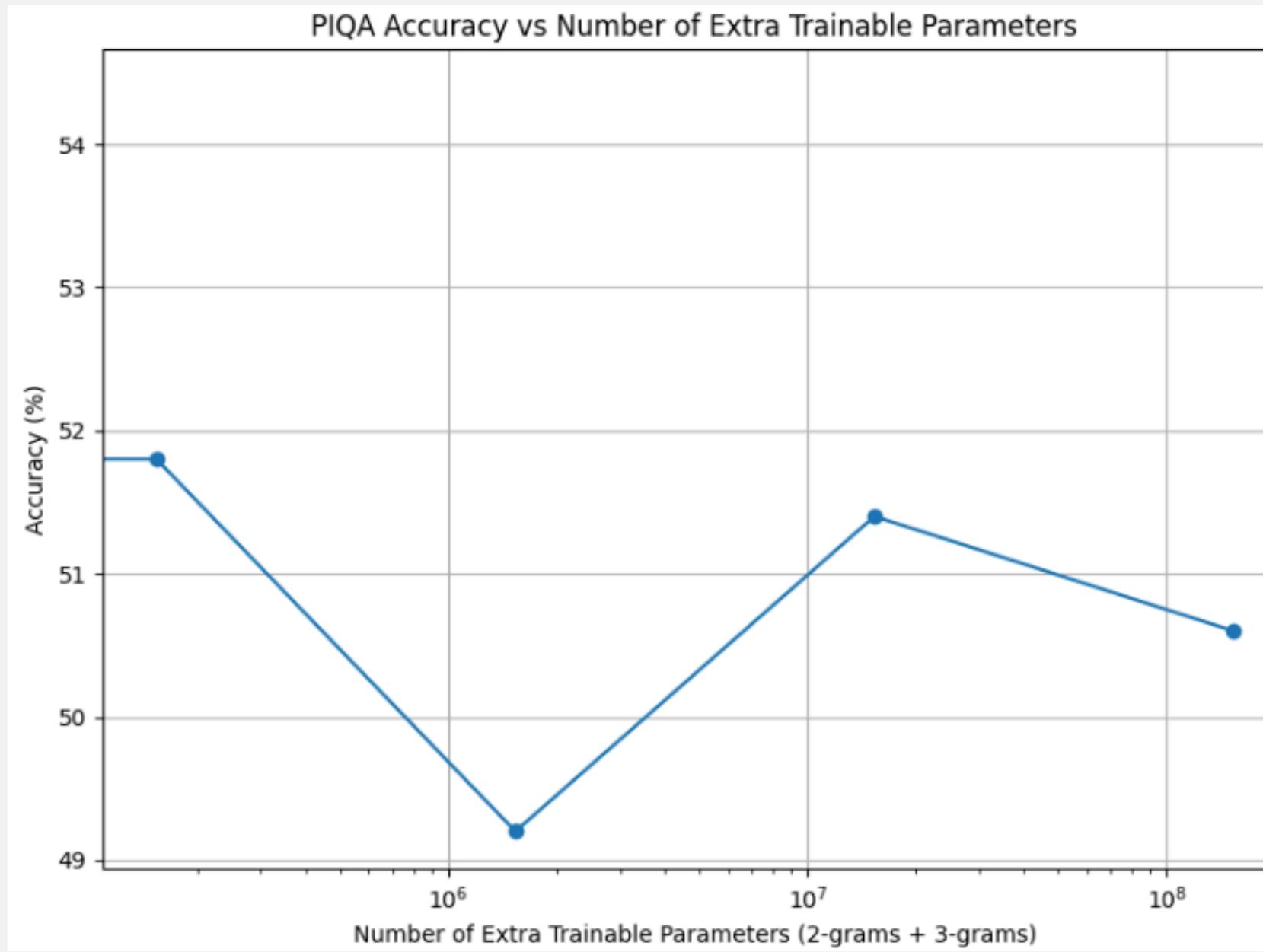
Χρησιμοποιήσαμε για το testing 500 samples για κάθε task και μετρήσαμε το accuracy των διάφορων μοντέλων συγκρίνοντάς το με το baseline GPT-2.

Αποτελέσματα (1)



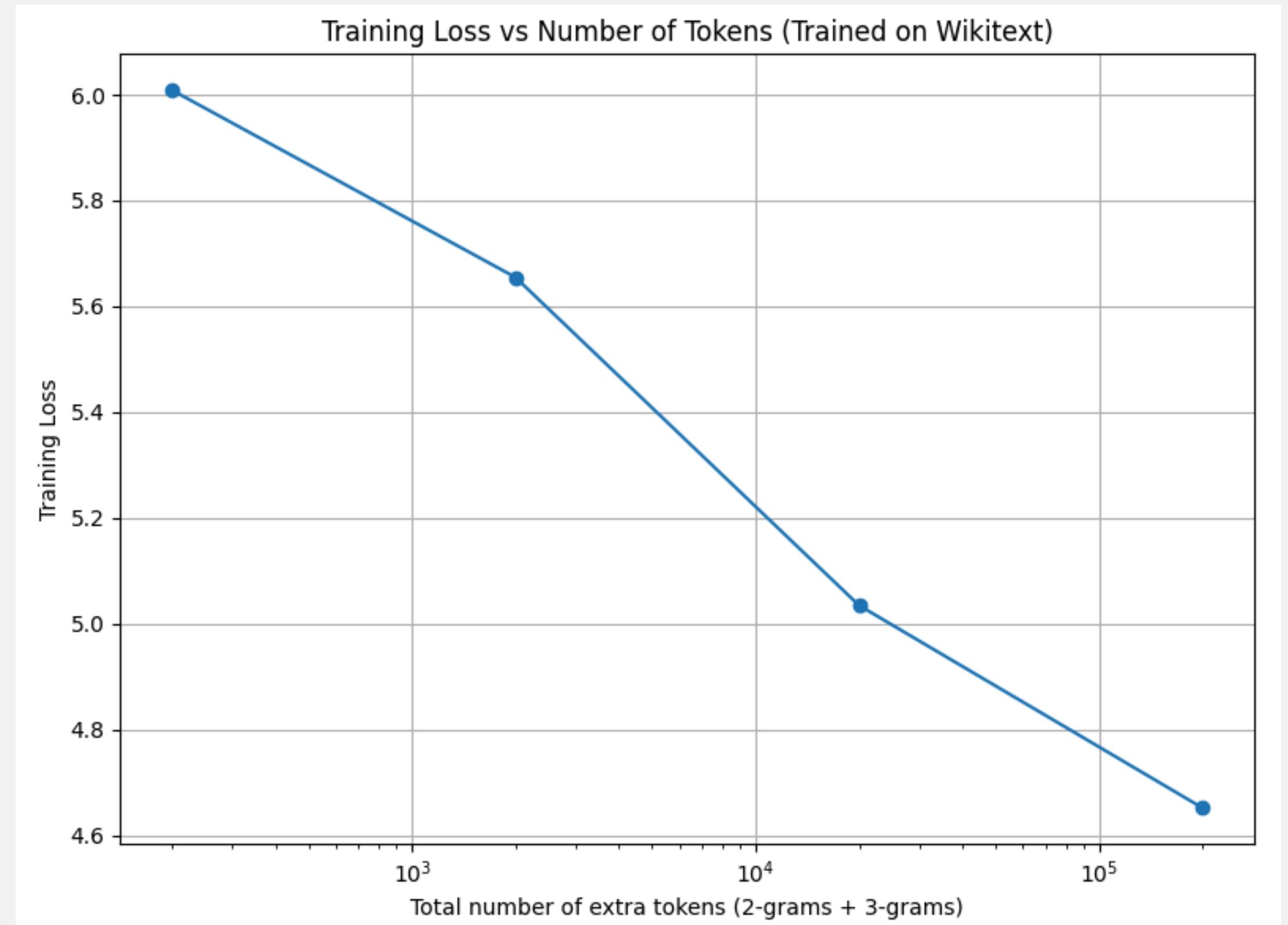
Παρατηρούμε πως σε **context creation tasks** οπως το **HellaSWAG** ή **το Story cloze**, το μοντέλο αυξάνει την επίδοσή του όσο αυξάνουμε το πλήθος των n-gram tokens.

Αποτελέσματα (1)



Από την άλλη, σε datasets όπως το PIQA τα n-gram tokens δεν βοηθάνε στην αύξηση του accuracy. Τα HellaSWAG και Story cloze ταιριάζουν με τον autoregressive τρόπο εκπαίδευσης του GPT-2, γι' αυτό αποδίδει καλά. Αντίθετα, το PIQA απαιτεί reasoning και συγκριτική κρίση, κάτι που το GPT-2 δυσκολεύεται να χειριστεί χωρίς επιπλέον fine-tuning πάνω σε question answering dataset.

Αποτελέσματα (1)



Όπως ακριβώς και στο paper, αποδυκνείεται μια **log linear** σχέση μεταξύ του **loss** και του αριθμού των **trainable tokens**.

Bonus Task: 4-gram expirement

Bonus Task: 4-gram expirement

- Στο σημείο αυτό θα πειραματιστούμε **αυξάνοντας το πλήθος του vocabulary ακόμη περισσότερο.**

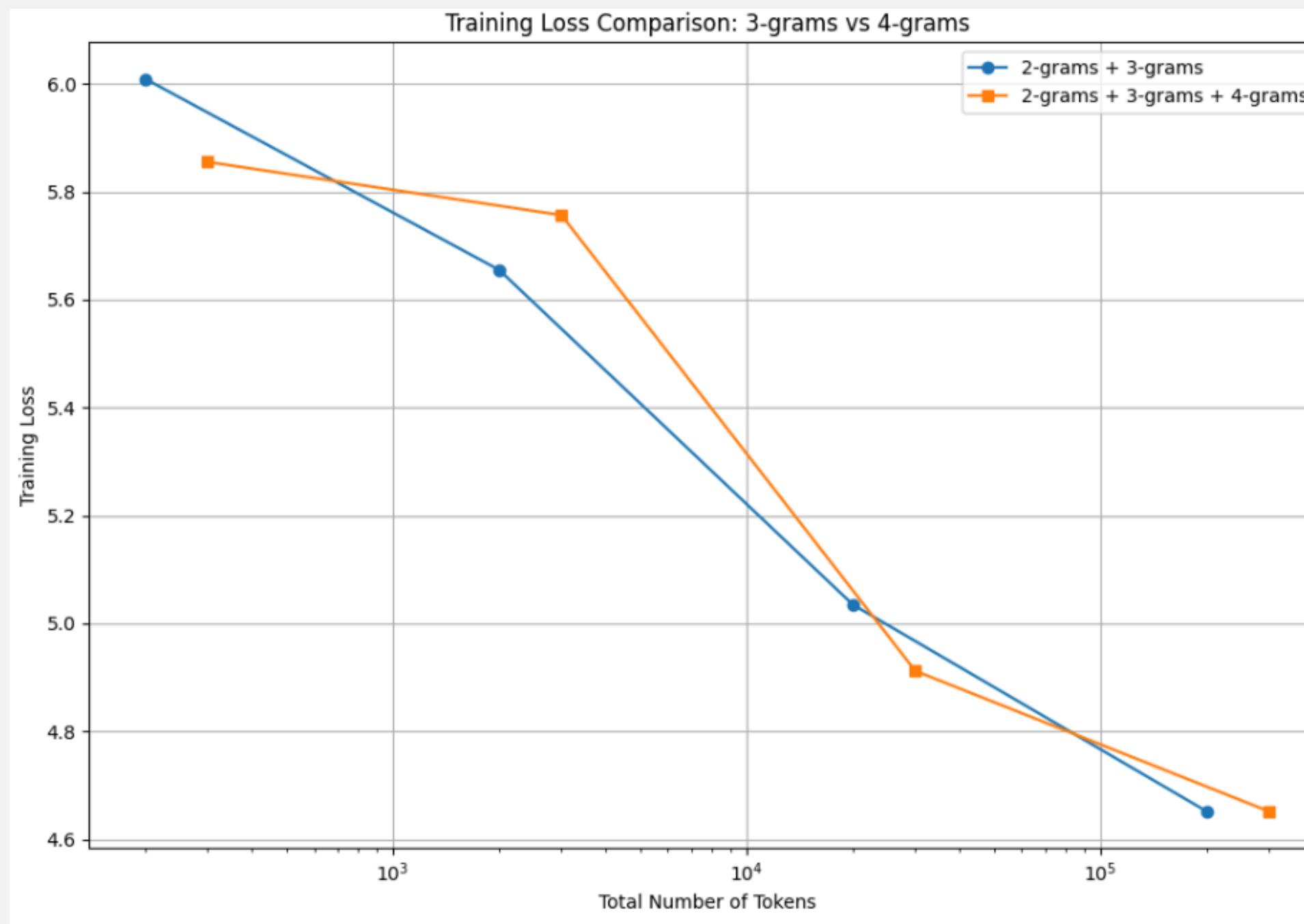
Bonus Task: 4-gram expirement

- Στο σημείο αυτό θα πειραματιστούμε **αυξάνοντας το πλήθος του vocabulary ακόμη περισσότερο.**
- Συγκεκριμένα θα προσθέσουμε στα embeddings και **4-grams και 5-grams** διατηρώντας συγχρόνως και τα 2-grams & 3-grams που ήδη έχουμε.

Bonus Task: 4-gram expirement

- Στο σημείο αυτό θα πειραματιστούμε **αυξάνοντας το πλήθος του vocabulary ακόμη περισσότερο.**
- Συγκεκριμένα θα προσθέσουμε στα embeddings και **4-grams και 5-grams** διατηρώντας συγχρόνως και τα 2-grams & 3-grams που ήδη έχουμε.
- Σκοπός μας να μετρήσουμε πως μεταβλήθηκε το **training loss** με την προσθήκη επιπλέον embeddings καθώς και **επίδοση του συστήματος σε downstream tasks** όπως τα παραπανω.

Bonus Task: 4-gram expirement

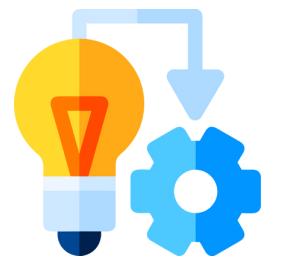


- Δεν παρατηρούμε μείωση του loss συγκριτικά με τα 3-gram μοντέλα και ούτε περεταίρω βελτίωση στο accuracy κατά το testing.
- Τα ίδια παρατηρούμε και για 5-gram μοντέλα και άρα η **υπολογιστική πολυπλοκότητα που επιβαρύνει την εκπαίδευση δεν επιφέρει ανάλογο κέρδος.**

Overview



Paper Analysis



Our Experiments



Project review

Αποτίμηση και μελλοντικές επεκτάσεις

Αποτίμηση και μελλοντικές επεκτάσεις

- Το over encoding βελτιώνει σταθερά την επίδοση αυξάνοντας διαρκώς το generation accuracy όσο το input vocabulary μεγαλώνει

Αποτίμηση και μελλοντικές επεκτάσεις

- Το over encoding βελτιώνει σταθερά την επίδοση αυξάνοντας διαρκώς το generation accuracy όσο το input vocabulary μεγαλώνει
- Μείωση του training loss και άρα δυνατότητα ταχύτερης εκπαίδευσης μέσω early stopping.

Αποτίμηση και μελλοντικές επεκτάσεις

- Το over encoding βελτιώνει σταθερά την επίδοση αυξάνοντας διαρκώς το generation accuracy όσο το input vocabulary μεγαλώνει
- Μείωση του training loss και άρα δυνατότητα ταχύτερης εκπαίδευσης μέσω early stopping.

Μελλοντικές επεκτάσεις

Αποτίμηση και μελλοντικές επεκτάσεις

- Το over encoding βελτιώνει σταθερά την επίδοση αυξάνοντας διαρκώς το generation accuracy όσο το input vocabulary μεγαλώνει
- Μείωση του training loss και άρα δυνατότητα ταχύτερης εκπαίδευσης μέσω early stopping.

Μελλοντικές επεκτάσεις

- Δυναμικό λεξιλόγιο που εξελίσσεται αυτόματα κατά την εκπαίδευση, με στόχο προσαρμογή στα δεδομένα αγνοώντας n-gram τα οποία δεν “βοήθησαν” το μοντέλο.

Αποτίμηση και μελλοντικές επεκτάσεις

- Το over encoding βελτιώνει σταθερά την επίδοση αυξάνοντας διαρκώς το generation accuracy όσο το input vocabulary μεγαλώνει
- Μείωση του training loss και άρα δυνατότητα ταχύτερης εκπαίδευσης μέσω early stopping.

Μελλοντικές επεκτάσεις

- Δυναμικό λεξιλόγιο που εξελίσσεται αυτόματα κατά την εκπαίδευση, με στόχο προσαρμογή στα δεδομένα αγνοώντας n-gram τα οποία δεν “βοήθησαν” το μοντέλο.
- Πειραματισμός με διάφορους συνδιασμούς n-gram, και όχι απαραίτητα hierarchical n-gram modelling, για διαπίστωση του πιο αποδοτικού συνδιασμού.

Retrospective



Retrospective Δυσκολίες και μελλοντικές βελτιώσεις



Retrospective Δυσκολίες και μελλοντικές βελτιώσεις

- Δυσκολία φόρτωσης cfg dataset με 1 εκατομμύριο samples. → **On the fly data loading.**



Retrospective Δυσκολίες και μελλοντικές βελτιώσεις

- Δυσκολία φόρτωσης cfg dataset με 1 εκατομμύριο samples. → **On the fly data loading.**
- Αργή εκπαίδευση μοντέλων. → **Row-wise sharding** για την κατανομή του embedding πίνακα στις GPUs, μειώνοντας το communication cost.

Retrospective Δυσκολίες και μελλοντικές βελτιώσεις

- Δυσκολία φόρτωσης cfg dataset με 1 εκατομμύριο samples. → **On the fly data loading.**
- Αργή εκπαίδευση μοντέλων. → **Row-wise sharding** για την κατανομή του embedding πίνακα στις GPUs, μειώνοντας το communication cost.
- Αδυναμία GPT-2 σε question answering tasks → **Fine tuning** σε κάποιο αντίστοιχο dataset αντί για Wikitext.

Retrospective Δυσκολίες και μελλοντικές βελτιώσεις

- Δυσκολία φόρτωσης cfg dataset με 1 εκατομμύριο samples. → **On the fly data loading.**
- Αργή εκπαίδευση μοντέλων. → **Row-wise sharding** για την κατανομή του embedding πίνακα στις GPUs, μειώνοντας το communication cost.
- Αδυναμία GPT-2 σε question answering tasks → **Fine tuning** σε κάποιο αντίστοιχο dataset αντί για Wikitext.
- Πειραματισμός και με over decoding τεχνικές πέραν του over encoding.

Thank you !

Κατσαϊδώνης Νικόλαος 03121868

Τζαμουράνης Γεώργιος 03121141

Κατσιαδράμης Κυριάκος 03121132

Φωτάκης Ανδρέας 03121100