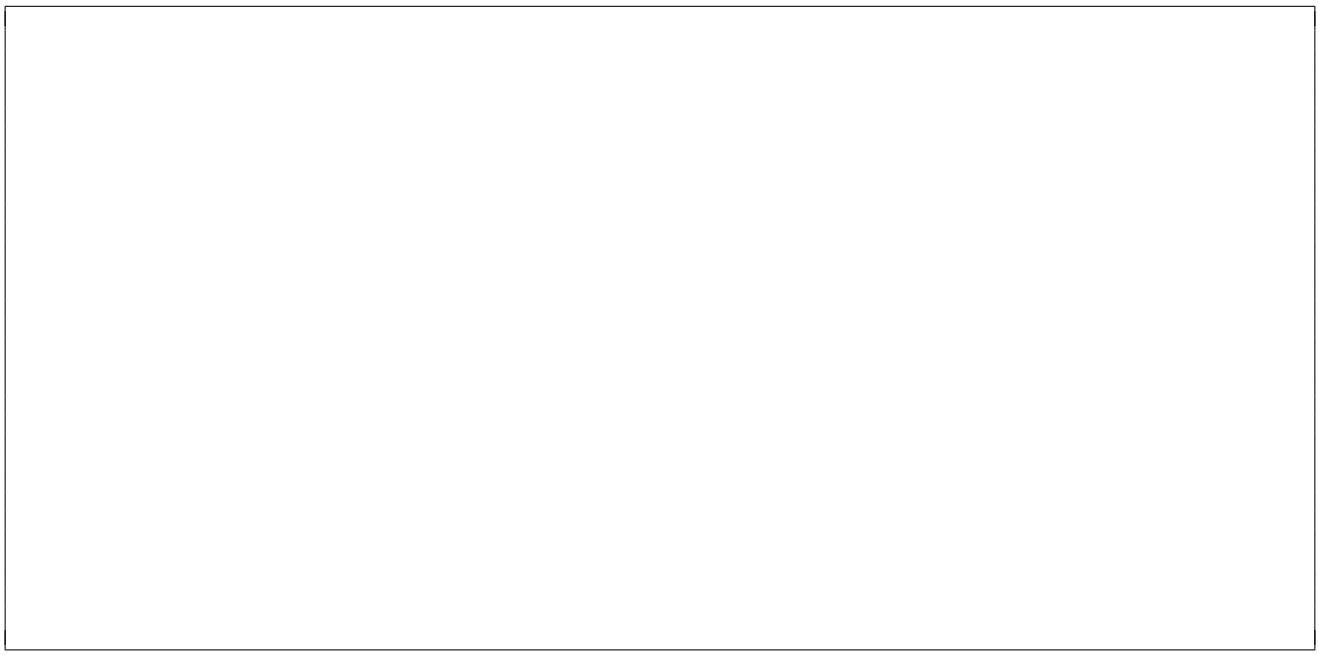
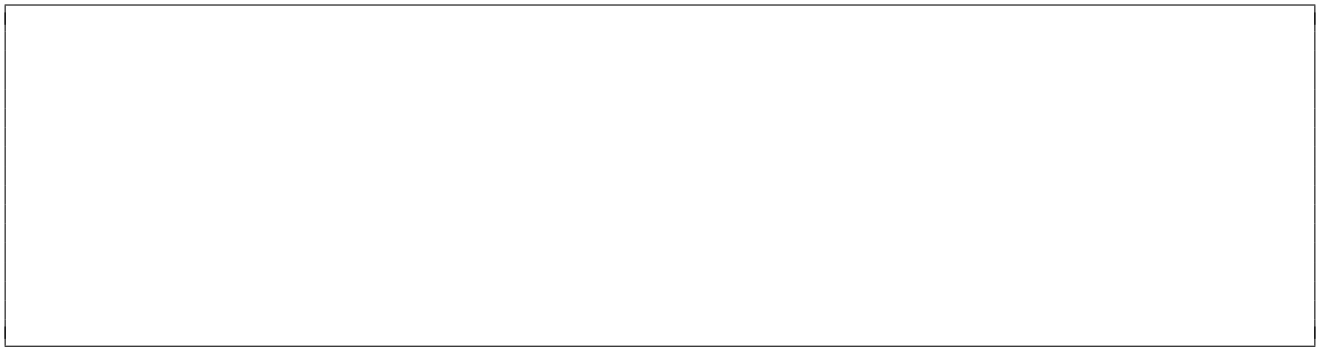
1. (4 Punkte) Ein Bigramm-Modell wird mit den vier Namen `anna`, `nina`, `ina`, `ana` trainiert. Der Punkt markiert Anfang und Ende, aus `anna` wird `.anna.`.
   (a) Zähle alle Bigramme.
   (b) Berechne die Wahrscheinlichkeiten für das Zeichen nach `n` und nach `a` (ohne +1).
   (c) Mit welcher Wahrscheinlichkeit erzeugt das Modell den Namen `nina`? Mit welcher den Namen `nana`, der nicht in der Liste steht?
   (d) Warum würde das Modell den Namen `nia` nie erzeugen? Wozu addiert man im Notebook zu jeder Zählung 1?



2. (3 Punkte) Wir nehmen das Bigramm-Modell aus der vorigen Aufgabe (ohne +1).
   (a) Berechne für den Namen `nina` für jedes der fünf Zeichen nach dem Anfangspunkt den Fehler $-\ln(p)$ und den Mittelwert. Das ist die Cross-Entropy des Modells für diesen Namen.
   (b) Wie groß wäre der mittlere Fehler eines Modells, das jedes der fünf Zeichen `.`, `a`, `i`, `n` und `o` mit gleicher Wahrscheinlichkeit rät?



3. (3 Punkte) Ein Sprachmodell liefert für drei mögliche nächste Zeichen die Logits $(2, 1, 0)$. Vor der Softmax-Funktion werden die Logits durch die Temperatur T geteilt.
   (a) Berechne die Wahrscheinlichkeiten für $T = 0.5$, $T = 1$ und $T = 2$.
   (b) Was bewirkt eine kleine, was eine große Temperatur beim Erzeugen von Text?

4. (4 Punkte) Die Zeichen `.`, `a`, `i`, `n` haben die Nummern 0 bis 3 und die Embeddings `.`: $(0, 0)$, `a`: $(1, 2)$, `i`: $(0.5, -1)$, `n`: $(-1, 0.5)$.
   (a) Schreibe die Embeddings als Matrix E mit 4 Zeilen und berechne den One-Hot-Vektor von `n` mal E.
   (b) Ein Modell bekommt als Kontext die letzten zwei Zeichen `an`. Welcher Vektor geht in die versteckte Schicht? Ein Neuron dort hat die Gewichte $(1, 0, 2, -2)$ und den Bias 0.5. Berechne seine Summe z.
   (c) Das MLP im Notebook hat 30 Zeichen, Embeddings der Länge 10, 3 Zeichen Kontext, 200 versteckte Neuronen und 30 Ausgaben. Berechne die Zahl der Parameter. Wie viele sind es bei 5 Zeichen Kontext?

5. (4 Punkte) Die aktuelle Position hat die Query $q = (1, 0)$. Die drei vorherigen Positionen haben die Keys $k_1 = (2, 0)$, $k_2 = (0, 1)$, $k_3 = (1, 1)$ und die Values $v_1 = (1, 0)$, $v_2 = (0, 1)$, $v_3 = (1, 1)$.
   (a) Berechne die Scores $q \cdot k_i$ (Skalarprodukt).
   (b) Berechne mit der Softmax-Funktion die Gewichte (ohne Teilen durch eine Wurzel).
   (c) Berechne das Ergebnis, die gewichtete Summe der Values.
   (d) Welche Position bekommt die meiste Aufmerksamkeit und warum?

6. (3 Punkte) Im Notebook wurden verschiedene Modelle mit 4378 Trainingsnamen trainiert und mit 487 Testnamen geprüft:

Modell	Fehler Training	Fehler Test
Bigramm	2.47	2.52
MLP, 3 Zeichen Kontext	2.02	2.39
MLP, 5 Zeichen Kontext	1.75	2.61
Attention, 8 Zeichen Kontext	2.06	2.33

   (a) Warum ist das MLP mit 5 Zeichen auf den Testnamen schlechter als mit 3, obwohl es mehr sieht?
   (b) Warum misst man den Fehler auf Testnamen, die beim Training nicht vorkamen?
   (c) Warum kann das Attention-Modell 8 Zeichen nutzen, ohne so stark zu überfitten?

7. (3 Punkte) Führe drei Schritte von Byte Pair Encoding für die Wörter `lena`, `elena`, `helena`, `leon`, `leni` durch: Am Anfang ist jedes Zeichen ein Token. In jedem Schritt wird das Paar nebeneinander stehender Tokens, das am häufigsten vorkommt, zu einem neuen Token verschmolzen. Gib nach dem dritten Schritt die Tokens jedes Wortes an.

8. (4 Punkte) (a) Was lernt ein großes Sprachmodell beim Vortraining? Was kommt bei der Feinabstimmung dazu?
   (b) Was sind Tokens und warum benutzt man keine ganzen Wörter?
   (c) Was ist eine Halluzination und warum entsteht sie?
   (d) Ein Chatbot erzeugt eine Antwort. Beschreibe, was dabei Schritt für Schritt passiert.

9. (4 Punkte) Erweitere das Bigramm-Modell zu einem **Trigramm-Modell**, das die letzten zwei Zeichen kennt: Zähle, wie oft auf zwei Zeichen ein drittes folgt (am Anfang steht `..`, am Ende `.`), mit +1 für jede Kombination. Berechne den Fehler auf den Testnamen (gleiche Aufteilung wie im Notebook) und vergleiche mit dem Bigramm-Modell.