Contents

Chapter 1 — Intuitive MCQ Bank (Introduction)

56 multiple-choice questions, intuitive/conceptual level. One correct option each. Cover the answer (blockquote) while testing yourself. Bilingual explanations (English + বাংলা).

Exam style: "Which statement best describes…", exactly one correct option; use full technical terms, no abbreviations.

বাংলা ব্যাখ্যা: এই ব্যাংকে অধ্যায় ১-এর ৫০+ ধারণামূলক MCQ আছে প্রতিটি প্রশ্নের নিচে উত্তর ও সংক্ষিপ্ত ব্যাখ্যা (ইংরেজি + বাংলা) দেওয়া

Topic: The AI Revolution (Section 1.1)

Q1. Which statement best describes what the course means by "the AI revolution" triggered by ChatGPT?

- A. Artificial intelligence became usable by anyone with a browser, removing the need for a developer to write application programming interface calls.
- B. A new neural network architecture was invented that no prior system had used.
- C. Governments first began to regulate artificial intelligence systems.
- D. Models started training themselves without any human-written data.

> **Answer: A.** The revolution was about *access*: before ChatGPT, using artificial intelligence required developer glue code and an application programming interface call; ChatGPT let anyone simply ask a question. B is tempting but wrong — the Transformer architecture already existed since 2017. **বাংলা:** বিপ্লবটা ছিল সহজলভ্যতার — আগে ডেভেলপার লাগত, এখন শুধু ব্রাউজার আর একটা প্রশ্নই যথেষ্ট; নতুন স্থাপত্য নয়

Q2. A friend says, "ChatGPT must be conscious because it remembers everything we discussed earlier in the chat." Using the chapter, what is the best refutation?

- A. The model is conscious, but only within a single session.
- B. The model itself is stateless; the chat application re-sends the full conversation history as context every turn, so the memory is an application-layer effect, not evidence of consciousness.
- C. The model has a permanent internal database that it queries each turn.
- D. The model rebuilds its weights after every message to store the conversation.

> **Answer: B.** The model is stateless — apparent memory comes from the application re-sending the prefix each turn and is bounded by the context length. D is wrong because weights are fixed at inference; nothing is retrained per message. **বাংলা:** মডেল নিজে কিছু মনে রাখে না; অ্যাপ প্রতিবার পুরনো কথোপকথন আবার পাঠায় বলে মনে রাখার ভ্রম — এটা চেতনার প্রমাণ নয়

Q3. In the term "Chat Generative Pretrained Transformer", which component names the neural network *architecture*?

- A. Chat
- B. Generative
- C. Transformer
- D. Pretrained

> **Answer: C.** Transformer is the attention-based architecture ("Attention is All You Need", 2017). Chat is the interface, Generative means it produces text, and Pretrained refers to the training stage — none of those name the architecture. **বাংলা:** Transformer-ই স্থাপত্য; Chat হলো ইন্টারফেস, Generative মানে টেক্সট তৈরি, Pretrained মানে প্রশিক্ষণের ধাপ

Q4. Why does the chapter call the chat assistant's memory an "illusion"?

- A. Because the model deletes its memory at random intervals.

- B. Because the memory is stored on the user's device, not the server.
- C. Because users imagine conversations that never happened.
- D. Because the model is stateless and the application achieves apparent memory only by re-sending the conversation history with every request.

Answer: D. The model holds no state between calls; the running context (all prior turns) is resent each turn, producing the *illusion* of memory. A is wrong — nothing is randomly deleted; the limit is context length. **বাংলা:** মডেল stateless, তাই প্রতিবার পুরো ইতিহাস আবার পাঠানো হয় — এটাই মনে রাখার ভ্রম তৈরি করে

Q5. Which of the following is the *most accurate* distinction the chapter draws between "ChatGPT" and "GPT-4"?

- A. ChatGPT is the application (a chat front-end), while GPT-4 is the underlying model.
- B. They are two names for the exact same thing.
- C. ChatGPT is the architecture and GPT-4 is the product.
- D. ChatGPT is the model and GPT-4 is the company.

Answer: A. The chapter explicitly warns against confusing the *application* (ChatGPT) with the *model* (GPT-3.5 / GPT-4). C reverses the roles and also misuses "architecture", which is the Transformer. **বাংলা:** ChatGPT হলো অ্যাপ্লিকেশন, GPT-4 হলো ভেতরের মডেল — এ দুটো গুলিয়ে ফেলা সাধারণ ভুল

Q6. A success factor for ChatGPT listed in the chapter is "mature cloud infrastructure." Why would this matter for adoption?

- A. It allowed the model to train itself without data.
- B. It meant the service could scale to serve millions of users quickly without each user owning powerful hardware.
- C. It removed the need for any neural network at all.
- D. It guaranteed the model would never produce false statements.

Answer: B. Reaching 100 million users in about two months required infrastructure that could serve that demand; mature cloud platforms made that feasible. D is wrong — hallucination remains an inherent failure mode regardless of infrastructure. **বাংলা:** পরিণত ক্লাউড অবকাঠামো ছিল বলেই দ্রুত কোটি ব্যবহারকারীকে সেবা দেওয়া গেছে; এটা ভুল উত্তর ঠেকায় না

Topic: Next-Word Prediction & the Chain Rule (Section 1.2.1)

Q7. Which statement best describes what a language model fundamentally does?

- A. It stores a dictionary of word definitions and looks them up.
- B. It translates every sentence into a fixed numeric identifier.
- C. It assigns probabilities to sequences of tokens, capturing how likely a word sequence is.
- D. It corrects grammar by applying a fixed rule set.

Answer: C. A language model assigns probabilities to word sequences — that is its defining job. A and D describe rule-based or lookup systems, which is precisely what language modelling replaced. **বাংলা:** ভাষা মডেল শব্দের ধারাকে সম্ভাবনা দেয় — এটাই মূল কাজ; অভিধান-খোঁজা বা নিয়ম প্রয়োগ নয়

Q8. The chain rule of probability rewrites the probability of a whole sentence as a product of conditional probabilities. Why does the chapter stress that the chain rule is "exact"?

- A. Because it always gives a probability of exactly 1.
- B. Because it ignores the order of the words.
- C. Because it only works for sentences of length two.
- D. Because it introduces no approximation — it is just a re-expression of the joint probability using the definition of conditional probability.

Answer: D. The chain rule merely rewrites a joint distribution as a product of conditionals via $P(A, B) = P(A)P(B \mid A)$; no information is lost. Approximation enters *only later* when the history

is truncated (the Markov assumption). **বাংলা:** chain rule শুধু যৌথ সম্ভাবনাকে শর্তাধীন গুণফল আকারে লেখে — কোনো আনুমানিকতা নেই; আনুমানিকতা আসে পরে, ইতিহাস ছাঁটলে

Q9. Given $P(\text{I}) = 0.05$, $P(\text{love} \mid \text{I}) = 0.10$, and $P(\text{AI} \mid \text{I, love}) = 0.20$, what is $P(\text{I love AI})$ under the chain rule?

- A. 1.00×10^{-3}
- B. 0.35
- C. 0.05
- D. 0.20

Answer: A. Multiply the conditional probabilities: $0.05 \times 0.10 \times 0.20 = 1.00 \times 10^{-3}$. B is the *sum* of the three numbers — a classic trap, since the chain rule multiplies, it does not add. **বাংলা:** chain rule-এ গুণ করতে হয়: $0.05 \times 0.10 \times 0.20 = 1.00 \times 10^{-3}$; যোগ করলে ভুল

Q10. Why do practical language-model systems work with log probabilities (sums of logarithms) instead of multiplying raw probabilities?

- A. Logarithms make the probabilities larger, which is easier to read.
- B. Multiplying many small probabilities underflows floating-point arithmetic, so summing logarithms is numerically safer.
- C. Logarithms convert probabilities into exact integer counts.
- D. The chain rule is only valid in logarithmic form.

Answer: B. Sentence probabilities shrink very fast as a product of many factors below one, risking floating-point underflow; summing logarithms avoids this. D is false — the chain rule is stated in probability form and is exact there. **বাংলা:** অনেক ছোট সম্ভাবনা গুণ করলে floating-point underflow হয়; তাই logarithm যোগ করা নিরাপদ

Q11. A more negative base-2 log probability for a sentence means which of the following?

- A. The sentence is more probable under the model.
- B. The sentence has more tokens.
- C. The sentence is less probable under the model.
- D. The sentence contains no unseen bigrams.

Answer: C. Since probabilities below one have negative logarithms, a *more negative* log probability corresponds to a *smaller* probability — a less likely sentence. B is unrelated; log probability reflects likelihood, not length directly. **বাংলা:** log probability যত বেশি ঋণাত্মক, সম্ভাবনা তত কম — বাক্যটি তত কম স্বাভাবিক

Topic: The Bigram Model & Maximum Likelihood Estimation (Section 1.2.2)

Q12. Which statement best describes the first-order Markov (bigram) assumption?

- A. Every word is statistically independent of every other word.
- B. The probability of the next word depends on the entire preceding history.
- C. The probability of a word depends on the word that comes after it.
- D. The probability of the next word depends only on the immediately preceding word.

Answer: D. A bigram model truncates the chain-rule history to just the previous token: $P(w_t \mid w_1, \ldots, w_{t-1}) \approx P(w_t \mid w_{t-1})$. A is the unigram/independence assumption; B is the exact chain rule with no approximation. **বাংলা:** bigram শুধু আগের একটি শব্দ দেখে; A হলো unigram ধারণা, B হলো পূর্ণ chain rule

Q13. In the maximum likelihood estimate $\hat{P}(w_t \mid w_{t-1}) = C(w_{t-1}, w_t)/C(w_{t-1})$, what is the correct denominator?

- A. The count of the previous word w_{t-1} in the corpus.
- B. The size of the vocabulary V.
- C. The total number of tokens in the corpus.
- D. The count of the bigram (w_{t-1}, w_t).

Answer: A. The denominator is the count of the *previous word*. Dividing by corpus length (C) or vocabulary size (B) is the single most common exam trap; D would make every estimate equal to one. **বাংলা:** ভাজক হলো আগের শব্দের গণনা $C(w_{t-1})$ — মোট শব্দসংখ্যা বা vocabulary দিয়ে ভাগ করা সবচেয়ে কমন ভুল

Q14. For the corpus "the cat sat on the mat the cat slept on the mat", the word "the" occurs 4 times and the bigram "the cat" occurs 2 times. What is $\hat{P}(\text{cat} \mid \text{the})$?

- A. 0.25
- B. 0.50
- C. 2.00
- D. 0.17

Answer: B. $\hat{P}(\text{cat} \mid \text{the}) = C(\text{the cat})/C(\text{the}) = 2/4 = 0.50$. A comes from wrongly dividing the bigram count by the corpus length region; D ($\approx 2/12$) is also a wrong-denominator trap. **বাংলা:** $2/4 = 0.50$; ভুল ভাজক ব্যবহার করলে 0.25 বা 0.17 আসে

Q15. In the toy corpus, after the word "cat" the model sees "sat" once and "slept" once. What does $\hat{P}(\text{sat} \mid \text{cat}) = \hat{P}(\text{slept} \mid \text{cat}) = 0.50$ tell us?

- A. The model strongly prefers "sat" over "slept".
- B. The corpus is grammatically incorrect.
- C. The count-based model considers "sat" and "slept" equally likely after "cat" and cannot prefer one over the other from these data.
- D. The bigram "cat sat" never occurred.

Answer: C. Both continuations occur once after "cat", so the maximum likelihood estimates are equal; the model has no basis to prefer either. This is why "the cat sat on the mat" and "the cat slept on the mat" get identical probability. **বাংলা:** "cat"-এর পর দুটোই একবার এসেছে, তাই মডেল কোনোটাকে এগিয়ে রাখতে পারে না — সমান সম্ভাবনা

Q16. Why does the bigram-model probability of "the mat sat on the cat" collapse to 0.00 even though the sentence is grammatical?

- A. Because "the mat sat on the cat" is too long for a bigram model.
- B. Because the model only accepts sentences it has seen verbatim.
- C. Because "cat" cannot appear at the end of a sentence.
- D. Because the bigram "(mat, sat)" never occurs in the corpus, so $\hat{P}(\text{sat} \mid \text{mat}) = 0$, and one zero factor forces the whole product to zero.

Answer: D. A single unseen bigram has probability zero, and since the sentence probability is a *product* of bigram probabilities, one zero factor zeroes the entire sentence — the zero-count problem. B is too strong; the model scores any sentence whose bigrams were all seen. **বাংলা:** "(mat, sat)" জোড়া কর্পাসে নেই, তাই তার সম্ভাবনা শূন্য; গুণফলে একটি শূন্যই পুরো বাক্যকে শূন্য করে দেয়

Q17. A student computes $\hat{P}(\text{cat} \mid \text{the})$ by dividing the bigram count of "the cat" by 12 (the corpus length) and gets about 0.17. What is the conceptual error?

- A. They divided by the corpus length instead of by $C(\text{the})$, the count of the previous word.
- B. They used the wrong logarithm base.
- C. They forgot to add Laplace smoothing.
- D. They counted the bigram wrongly.

Answer: A. The maximum likelihood denominator must be the count of the conditioning word $C(\text{the}) = 4$, not the corpus length 12. The bigram count (2) is correct, so D is wrong; smoothing (C) is a different, optional step. **বাংলা:** ভুলটা ভাজকে — মোট ১২ দিয়ে ভাগ নয়, আগের শব্দ "the"-এর গণনা ৪ দিয়ে ভাগ করতে হবে

Q18. Why must every row of a *complete* bigram transition table sum to 1.00 (a useful exam sanity check)?

- A. Because the vocabulary always has exactly that many words.
- B. Because, given a previous word, the probabilities of all possible next words form a probability distribution that must sum to one.

- C. Because the corpus length is always one.
- D. Because the columns, not the rows, must sum to one.

Answer: B. Each row is a conditional distribution $\hat{P}(\cdot \mid w_{t-1})$ over all possible next words, which must sum to 1.00. D is wrong — it is the rows (fixed previous word), not columns, that normalize. **বাংলা:** প্রতিটি সারি হলো আগের শব্দ স্থির রেখে পরের শব্দের বণ্টন, তাই যোগফল ১; কলাম নয়, সারি স্বাভাবিক হয়

Q19. The chapter notes that the "mat" row of the toy transition table sums to 0.50, not 1.00. What is the reason?

- A. A counting error in the corpus.
- B. The vocabulary size was computed incorrectly.
- C. "mat" is the final token of the corpus and has no successor, so under the convention of dividing by the unigram count its outgoing probabilities do not sum to one.
- D. "mat" is not a real word.

Answer: C. The last corpus token has no following token, so one of its two occurrences contributes no successor; dividing by the unigram count $C(\text{mat}) = 2$ leaves the outgoing mass at 0.50. This is a corpus-boundary subtlety, not an error. **বাংলা:** "mat" কর্পাসের শেষ টোকেন, তার কোনো পরের শব্দ নেই; তাই unigram গণনা দিয়ে ভাগ করলে যোগফল ০.৫০ হয় — ভুল নয়, সীমানার প্রভাব

Q20. Computing $P(\text{the cat sat on the mat})$ uses the unigram start term $P(\text{the}) = 4/12 \approx 0.33$ multiplied by the chained bigrams, giving $1/24 \approx 0.04$. Why is the *unigram* probability used for the first word?

- A. Because the first word is always the most frequent word.
- B. Because unigram probabilities are always larger than bigram probabilities.
- C. Because the chain rule forbids bigrams at position one.
- D. Because the first word has no preceding word, so there is no bigram to condition on; its prior is the unigram probability.

Answer: D. The first token has no predecessor, so it cannot be conditioned on a previous word; its standalone (unigram) probability serves as the start term. A and B are simply false generalizations. **বাংলা:** প্রথম শব্দের আগে কোনো শব্দ নেই, তাই তার জন্য bigram হয় না — তার unigram সম্ভাবনাই শুরুর term

---

Topic: Why n-gram Models Break Down (Section 1.2.3)

Q21. Which statement best describes the central scaling problem of n-gram models as n grows?

- A. The number of possible n-grams grows as V^n, so no realistic corpus can cover a meaningful fraction of them, causing pervasive zero counts.
- B. The model becomes too fast to be useful.
- C. The Markov assumption becomes exact.
- D. The vocabulary size V shrinks toward zero.

Answer: A. With V^n possible contexts (e.g. $50{,}000^5 \approx 3.13 \times 10^{23}$), most grammatical n-grams are never observed, so data sparsity dominates. C is the opposite of the truth — larger n still only approximates the full history. **বাংলা:** সম্ভাব্য n-gram-এর সংখ্যা V^n হারে বিস্ফোরিত হয়, কোনো কর্পাসেই যথেষ্ট উদাহরণ নেই — তাই শূন্য গণনা সর্বত্র

Q22. Why can a 5-gram model never reliably connect "The girl who grew up in Braunschweig speaks fluent ____" to "German"?

- A. Because "German" is not in the vocabulary.
- B. Because the relevant clue ("Braunschweig") lies more than four tokens back, beyond the fixed window the 5-gram can see.
- C. Because 5-gram models cannot handle proper nouns.
- D. Because the sentence is grammatically incorrect.

Answer: B. A 5-gram conditions only on the previous four tokens; the disambiguating clue is further back, so the Markov assumption is hard-limited and the long-range dependency is lost. A and C are

not the issue — it is the limited context window. **বাংলা:** 5-gram শুধু আগের চারটি শব্দ দেখে; "Braunschweig" তার চেয়ে দূরে, তাই দূরসম্পর্কের সূত্রটা ধরা পড়ে না

Q23. To a count-based n-gram model, "the cat sat" and "the dog sat" are unrelated events. What capability does this reveal is *missing*?

- A. The ability to count bigrams.
- B. The ability to store probabilities.
- C. The ability to generalize across semantically similar words, because each word is just a discrete symbol with no shared representation.
- D. The ability to tokenize text.

Answer: C. Count-based models treat each word as an isolated symbol, so statistical strength is not shared between "cat" and "dog"; this lack of generalization is one of the three failures motivating neural language models. **বাংলা:** count-মডেলে প্রতিটি শব্দ আলাদা প্রতীক, তাই "cat" আর "dog"-এর মিল কাজে লাগে না — generalization-এর অভাব

Q24. What does the chapter present as the solution to the three failures of n-gram models (sparsity, limited context, no generalization)?

- A. Using an even larger value of n.
- B. Deleting rare words from the vocabulary.
- C. Switching from bigrams back to unigrams.
- D. A neural language model — a parametric function f_θ that outputs a distribution over the vocabulary — leading ultimately to Large Language Models.

Answer: D. The chapter motivates neural language models as the fix: continuous word representations generalize, attention extends context, and a fixed parameter budget avoids the V^n explosion. A makes sparsity worse, not better. **বাংলা:** সমাধান হলো নিউরাল ভাষা মডেল — যা generalize করে, বড় context নেয়, আর V^n বিস্ফোরণ এড়ায়; n বাড়ানো নয়

Q25. Why does increasing n in an n-gram model *not* solve the long-range dependency problem in a sustainable way?

- A. Because each increase in n multiplies the number of possible contexts by V, so sparsity worsens faster than coverage improves.
- B. Because larger n makes the model ignore recent words.
- C. Because larger n changes the vocabulary.
- D. Because the chain rule breaks for $n > 2$.

Answer: A. Extending the window helps capture slightly longer context, but the count of possible contexts scales as V^n, so data become exponentially sparser — an unwinnable trade. D is false; the chain rule holds for any order. **বাংলা:** n বাড়ালে context সামান্য বড় হয়, কিন্তু সম্ভাব্য context-এর সংখ্যা V গুণে বাড়ে — sparsity দ্রুত বাড়ে, তাই টেকসই সমাধান নয়

Topic: Language Models in Speech Recognition (Section 1.2.4)

Q26. In the speech-recognition decision rule $\hat{W} = \arg\max_W P(X \mid W)\, P(W)$, which statement best describes the role of $P(W)$?

- A. It converts the raw audio into acoustic feature vectors.
- B. It assigns a prior probability to each candidate word sequence, favoring linguistically plausible hypotheses.
- C. It models how well the audio matches a candidate word sequence.
- D. It normalizes the result by dividing by the evidence $P(X)$.

Answer: B. $P(W)$ is the language model — the prior over word sequences, scoring linguistic plausibility independently of the audio. C describes the acoustic model $P(X \mid W)$; D is unnecessary because $P(X)$ is constant across hypotheses. **বাংলা:** $P(W)$ হলো ভাষা মডেল — কোন বাক্য মানুষ বলে তার পূর্ব-সম্ভাবনা; $P(X \mid W)$ হলো acoustic model

Q27. Why may the evidence term $P(X)$ be dropped from the arg max in the noisy-channel decoding rule?

- A. Because $P(X)$ is always equal to one.
- B. Because $P(X)$ is the language model and is therefore optional.
- C. Because $P(X)$ is the same for every candidate hypothesis W (the audio is fixed), so it does not affect which hypothesis attains the maximum.
- D. Because $P(X)$ is too small to compute.

Answer: C. The audio X is fixed, so $P(X)$ is a constant common to all hypotheses; dividing every score by the same constant cannot change the arg max. A is false — $P(X)$ is not generally one. **বাংলা:** অডিও স্থির, তাই সব hypothesis-এর জন্য $P(X)$ একই ধ্রুবক; একই সংখ্যা দিয়ে ভাগ করলে সর্বোচ্চ যেটা সেটাই থাকে

Q28. A decoder compares W_1="recognize speech" (acoustic 0.30, language 1.00×10^{-3}) and W_2="wreck a nice beach" (acoustic 0.36, language 2.00×10^{-5}). Which is chosen and why?

- A. W_2, because it has the higher acoustic likelihood.
- B. W_2, because longer sentences are always preferred.
- C. W_1, because it has the higher acoustic likelihood.
- D. W_1, because the product $P(X \mid W)\,P(W)$ is larger (3.00×10^{-4} vs 7.20×10^{-6}); the language model vetoes the implausible W_2.

Answer: D. The decision uses the *product*: $0.30 \times 10^{-3} = 3.00 \times 10^{-4}$ beats $0.36 \times 2.00 \times 10^{-5} = 7.20 \times 10^{-6}$. Although W_2 fits the audio better, the language model rates it far less probable. C wrongly claims W_1 has higher acoustics. **বাংলা:** সিদ্ধান্ত হয় গুণফলে; W_2 কানে বেশি মিললেও ভাষা মডেল তাকে অগ্রাহ্য করে, তাই W_1 জেতে

Q29. This is exactly the value a language model adds in speech recognition. Which sentence captures that value?

- A. It can override an acoustically better-fitting hypothesis when that hypothesis is a linguistically improbable utterance.
- B. It speeds up the conversion of audio into features.
- C. It removes the need for any acoustic model.
- D. It guarantees the audio is recorded without noise.

Answer: A. The language model contributes linguistic prior knowledge, letting the system reject homophone-like but implausible sequences even when they match the sound slightly better. C is wrong — both models are needed and combined. **বাংলা:** ভাষা মডেলের মূল অবদান — শব্দে মিললেও অস্বাভাবিক বাক্যকে বাতিল করা; acoustic model বাদ দেওয়া নয়

Q30. What would happen to a speech recognizer if the language model were removed (i.e. $P(W)$ made uniform across all hypotheses)?

- A. Nothing would change.
- B. The rule would degenerate to $\arg\max_W P(X \mid W)$ — pure acoustic matching — so homophone confusions like "wreck a nice beach" beating "recognize speech" would increase.
- C. The recognizer would become perfectly accurate.
- D. The acoustic model would also stop working.

Answer: B. A uniform $P(W)$ drops out of the arg max, leaving only acoustic matching; the linguistic prior that suppresses implausible homophones is gone, so error rate rises. C is the opposite of the truth. **বাংলা:** $P(W)$ সমান হলে সিদ্ধান্ত শুধু acoustic মিলে পরিণত হয়; homophone-জনিত ভুল বেড়ে যায়, নির্ভুলতা কমে

Q31. In the noisy-channel model, which pairing of term to meaning is correct?

- A. $P(X \mid W)$ is the language model; $P(W)$ is the acoustic model.
- B. $P(X \mid W)$ is the evidence; $P(W)$ is the posterior.
- C. $P(X \mid W)$ is the acoustic model; $P(W)$ is the language model.
- D. $P(X \mid W)$ is the posterior; $P(W)$ is the evidence.

Answer: C. $P(X \mid W)$ scores how well the audio matches a word sequence (acoustic model); $P(W)$ is the prior over word sequences (language model). A swaps the two — a common confusion. **বাংলা:**

$P(X \mid W)$ = acoustic model (শব্দের মিল), $P(W)$ = language model (বাক্যের পূর্ব-সম্ভাবনা); A উল্টে দিয়েছে

Topic: Perplexity (Section 1.2.5)

Q32. Which statement best describes the intuitive meaning of perplexity?

- A. The number of parameters in the model.
- B. The probability that the model is correct.
- C. The total number of tokens in the test set.
- D. The average branching factor — roughly, among how many equally likely words the model is hesitating at each step.

Answer: D. Perplexity is the average branching factor: lower means the model is more confident and better. A and C confuse it with model or data size; it is a per-step uncertainty measure. **বাংলা:** perplexity হলো গড় branching factor — প্রতি ধাপে মডেল গড়ে কতগুলো শব্দের মধ্যে দ্বিধায়; মান যত কম, মডেল তত ভালো

Q33. A model has lower perplexity on a test set than a competing model. What does this indicate?

- A. The lower-perplexity model is better, because it is on average less uncertain about the next word.
- B. The lower-perplexity model is worse.
- C. The two models are identical.
- D. The lower-perplexity model has more parameters.

Answer: A. Lower perplexity = lower average branching factor = a better language model. B inverts the convention; D is unrelated, since perplexity measures uncertainty, not size. **বাংলা:** কম perplexity মানে ভালো মডেল — গড়ে কম অনিশ্চয়তা; প্যারামিটারের সংখ্যার সাথে সম্পর্ক নেই

Q34. The toy bigram model scores its own training sentence with perplexity about 1.70. Why is this so low?

- A. Because the test sentence had thousands of tokens.
- B. Because the model has effectively memorized its tiny corpus, so it is barely uncertain on a sentence it has already seen.
- C. Because perplexity is always near one for bigram models.
- D. Because the vocabulary size is 1.

Answer: B. On a memorized training sentence the model assigns very high probability and is almost not hesitating, giving very low perplexity. On unseen text with a zero-probability bigram, perplexity would instead be infinite. **বাংলা:** ছোট কর্পাসটা মডেল মুখস্থ করে ফেলেছে, তাই চেনা বাক্যে দ্বিধা প্রায় নেই — perplexity খুব কম (১.৭০)

Q35. A uniform random language model over a vocabulary of size V has what perplexity, and what does this tell us about the metric?

- A. Perplexity 1, meaning it is a perfect model.
- B. Perplexity 0, meaning it never makes errors.
- C. Perplexity exactly V, meaning it is maximally uncertain — choosing among all V words equally; perplexity therefore upper-bounds at the vocabulary size for such a model.
- D. Perplexity infinite, regardless of V.

Answer: C. A uniform model is equally undecided among all V words, so its branching factor — and hence perplexity — is exactly V. This anchors the scale: good models score far below V. **বাংলা:** uniform মডেল Vটি শব্দের মধ্যে সমানভাবে দ্বিধায়, তাই perplexity ঠিক V; ভালো মডেল এর অনেক নিচে থাকে

Q36. Why would the toy bigram model's perplexity be *infinite* on a test sentence containing an unseen word pair?

- A. Because the sentence is too short.
- B. Because perplexity ignores unseen pairs.
- C. Because the vocabulary is too large.
- D. Because the unseen bigram has probability zero, making the sentence probability zero, and perplexity is a (negative) power of that probability — undefined/infinite for zero.

Answer: D. An unseen bigram gives zero sentence probability; since perplexity is $P^{-1/N}$, a zero probability sends perplexity to infinity. This is the perplexity-side face of the zero-count problem. **বাংলা:** অদেখা জোড়ার সম্ভাবনা শূন্য, ফলে বাক্যের সম্ভাবনা শূন্য; $P^{-1/N}$ তখন অসীম — এটাই zero-count সমস্যার আরেক রূপ

Topic: Large Language Models (Section 1.3)

Q37. Which statement best describes how a Large Language Model relates to the bigram model of Section 1.2?

- A. It uses the same next-word prediction objective, but the conditional probability is computed by a deep neural network instead of count tables.
- B. It abandons next-word prediction entirely for a new objective.
- C. It only works on speech, not text.
- D. It replaces probabilities with hand-written grammar rules.

Answer: A. A Large Language Model keeps the chain-rule next-word objective; the difference is that a neural network with parameters θ implements the conditional instead of count tables. B is wrong — the objective is the same, just scaled. **বাংলা:** Large Language Model সেই "পরের শব্দ" লক্ষ্যই রাখে; পার্থক্য — গণনার টেবিলের বদলে গভীর নিউরাল নেটওয়ার্ক

Q38. Why does a neural language model handle the words "cat" and "dog" more gracefully than a count-based n-gram model?

- A. It deletes one of the two words to avoid confusion.
- B. It represents words as continuous vectors, so similar words share statistical strength rather than being treated as unrelated symbols.
- C. It memorizes every sentence containing either word.
- D. It refuses to process rare words.

Answer: B. Continuous word representations let semantically similar words share evidence, directly fixing the n-gram failure of no generalization. C describes count-based memorization, which is the problem, not the fix. **বাংলা:** নিউরাল মডেলে শব্দ ভেক্টর হিসেবে থাকে, তাই কাছাকাছি শব্দরা একে অপরের তথ্য ভাগ করে — n-gram-এর generalization-অভাব দূর হয়

Q39. The chapter says a Large Language Model's memory is "fixed at the parameter count instead of exploding as V^n." Why is this an advantage?

- A. Because it makes the vocabulary smaller.
- B. Because parameters are free to store.
- C. Because the storage requirement no longer grows exponentially with context length; a fixed parameter budget can encode broad statistics without a V^n table.
- D. Because it removes the need for any training data.

Answer: C. Instead of an exponentially growing count table, the neural model uses a fixed set of parameters, sidestepping the n-gram explosion. D is false — these models are trained on trillions of tokens. **বাংলা:** V^n টেবিলের বদলে স্থির সংখ্যক প্যারামিটার ব্যবহার হয়, তাই context বাড়লেও মেমরি বিস্ফোরিত হয় না

Q40. Which statement best describes "autoregressive generation" as used by a Large Language Model?

- A. Generating all tokens of the output simultaneously in one step.
- B. Selecting the output from a fixed list of stored sentences.
- C. Translating the prompt into a single numeric label.
- D. Producing output one token at a time, each conditioned on all previously generated tokens, then appending and repeating.

Answer: D. Autoregressive generation emits one token at a time, feeding each new token back as context — the GPT-style decoding loop. A describes a non-autoregressive scheme, which is not what GPT-style models do. **বাংলা:** autoregressive generation মানে এক টোকেন করে তৈরি, প্রতিটা আগেরগুলোর ওপর শর্ত রেখে, তারপর যোগ করে পুনরাবৃত্তি

Q41. The chapter warns that Large Language Models do not "look up" facts. What does this imply about a failure mode called hallucination?

- A. Because the model produces fluent text from statistics stored in its weights rather than querying a database, it can generate fluent but false statements.
- B. Hallucination only happens when the internet connection fails.
- C. Hallucination means the model refuses to answer.
- D. Hallucination is impossible for neural models.

Answer: A. Facts live as statistics in the weights, not as retrievable database entries; this is why a model can produce fluent yet incorrect output (hallucination), a gap Retrieval-Augmented Generation later addresses. **বাংলা:** তথ্য ওজনের ভেতরে পরিসংখ্যান হিসেবে থাকে, ডেটাবেস নয়; তাই সাবলীল অথচ ভুল উত্তর — hallucination — সম্ভব

Q42. Why does the chapter say capabilities "emerge with scale" (e.g. a 70-billion-parameter model can write a coherent poem while a tiny model cannot)?

- A. Because larger models are explicitly trained on poems and smaller ones are not.
- B. Because greater scale (more data, parameters, compute) lets the model internalize enough structure that complex abilities appear without any task-specific training.
- C. Because small models are deliberately disabled.
- D. Because poems require a separate poem module.

Answer: B. The same next-word objective at large scale yields emergent abilities; no poem-specific training is involved. A and D contradict the chapter's point that competence transfers from general pretraining. **বাংলা:** বড় পরিসরে (ডেটা, প্যারামিটার, কম্পিউট) একই লক্ষ্য থেকেই জটিল দক্ষতা আবির্ভূত হয়, আলাদা কবিতা-প্রশিক্ষণ ছাড়াই

Topic: Foundation Models & the Paradigm Shift (Section 1.4)

Q43. Which statement best describes a Foundation Model?

- A. Any neural network with more than one billion parameters.
- B. A rule-based expert system forming the base layer of a software stack.
- C. A model pretrained on broad data with self-supervision that can be adapted to a wide range of downstream tasks.
- D. A language model fine-tuned exclusively for conversational chat.

Answer: C. A Foundation Model is defined by *broad pretraining data*, *self-supervision*, and *adaptability to many downstream tasks*. A is wrong because parameter count alone does not define the paradigm; D describes one adapted product, not the foundation. **বাংলা:** Foundation Model-এর সংজ্ঞা — বিস্তৃত ডেটায় self-supervised pretraining এবং বহু কাজে অভিযোজনযোগ্যতা; শুধু প্যারামিটারসংখ্যা নয়

Q44. Which statement best describes the relationship between Large Language Models and Foundation Models?

- A. They are exact synonyms.
- B. Foundation models are a special case of Large Language Models.
- C. They are unrelated concepts.
- D. Large Language Models are one type of foundation model — the text modality — while foundation models also span images, audio, and video.

Answer: D. A Large Language Model is the text-modality instance of the broader foundation-model category, which also includes image, audio, and video models. A (synonym) and B (reversed subset) are both common errors. **বাংলা:** Large Language Model হলো foundation model-এর টেক্সট সংস্করণ; foundation model আরও বড় ছাতা — ছবি, অডিও, ভিডিওও এর অন্তর্গত

Q45. Which statement best describes the "pretrain → adapt" paradigm shift?

- A. Pretrain one expensive model on broad unlabelled data, then adapt it cheaply to many tasks via prompting, retrieval, tool use, or fine-tuning.

- B. Collect labelled data for each task and train a fresh task-specific model every time.
- C. Hand-code rules for each new task as it arises.
- D. Train a separate model for every user.

Answer: A. The paradigm is one costly broad pretraining run followed by many cheap adaptations. B describes the *classical* per-task pipeline the paradigm replaced; C is the older symbolic approach. **বাংলা:** নতুন দৃষ্টান্ত — একবার ব্যয়বহুল pretraining, তারপর prompt/retrieval/tool/fine-tuning দিয়ে বহু কাজে সস্তায় অভিযোজন

Q46. Why does the chapter say "almost nobody trains foundation models — almost everybody engineers products on top of them"?

- A. Because training is illegal.
- B. Because pretraining costs tens of millions of dollars in compute, so the economically sensible activity for most is building on pre-trained models — which is AI Engineering.
- C. Because foundation models cannot be improved.
- D. Because there is no demand for new models.

Answer: B. The enormous pretraining cost concentrates model creation among a few, leaving most practitioners to engineer products on existing foundation models — the very definition of AI Engineering. A and D are false. **বাংলা:** pretraining-এর খরচ কোটি ডলার, তাই বেশিরভাগের কাজ তৈরি মডেলের ওপর প্রোডাক্ট বানানো — এটাই AI Engineering

Q47. A startup needs both a legal-document summarizer and a support chatbot. What is the foundation-model approach, and a key trade-off?

- A. Train two separate task-specific models from scratch; trade-off: faster time to market.
- B. Hand-code rules for both tasks; trade-off: requires no data at all.
- C. Use one pretrained foundation model adapted twice (e.g. by prompting or fine-tuning); trade-off: dependence on the model provider and less control over failure modes.
- D. Use the chatbot model for legal summaries without any adaptation; trade-off: none.

Answer: C. One pretrained foundation model can serve both tasks via cheap adaptation, sharing pretraining cost and general language competence; the trade-off is reliance on the provider and reduced control over failures and inference cost. A reverses the cost story. **বাংলা:** একটি pretrained foundation model দুবার অভিযোজন করলেই দুই কাজ চলে; trade-off হলো provider-নির্ভরতা ও failure-নিয়ন্ত্রণ কম

Q48. Both a diffusion image model and a decoder-only Transformer count as foundation models despite completely different architectures. What do they share?

- A. The same neural network architecture.
- B. The same training corpus.
- C. The same number of parameters.
- D. The same paradigm: broad-data self-supervised pretraining followed by cheap adaptation to many downstream tasks.

Answer: D. "Foundation model" describes the *training-and-use workflow*, not a specific architecture; both follow pretrain-then-adapt. A directly contradicts the premise that their architectures differ. **বাংলা:** "Foundation model" স্থাপত্য নয়, প্রশিক্ষণ-ও-ব্যবহারের ধরন বোঝায় — দুটোই pretrain-তারপর-adapt; স্থাপত্য আলাদা

Q49. The chapter shows one fake model function serving as both a translator and a sentiment classifier, selected purely by the system prompt. What capability does this illustrate?

- A. That task selection can happen through the prompt alone — adaptation without retraining — the defining capability of foundation models.
- B. That the model is retrained between the two calls.
- C. That prompting changes the model's parameters.
- D. That the two tasks require two different models.

Answer: A. The same function (model) performs different tasks chosen entirely by the prompt, demonstrating adaptation without retraining. B and C wrongly assume the weights change; the

point is that they do not. **বাংলা:** একই মডেল শুধু prompt বদলে দুই কাজ করে — retraining ছাড়াই অভিযোজন, যা foundation model-এর মূল ক্ষমতা

Topic: Modalities, Tools & Local Inference (Section 1.5)

Q50. Which statement best describes the difference between LangChain and a Large Language Model?

- A. LangChain is the model and the Large Language Model is the framework.
- B. LangChain is a software framework that orchestrates calls to models, prompts, retrieval, and tools; it contains no intelligence itself.
- C. They are the same thing.
- D. LangChain is a quantization method for shrinking models.

Answer: B. LangChain glues together models, prompts, retrieval, and tools behind a Python interface but holds no intelligence — the model does the reasoning. A reverses the roles; D confuses it with quantization. **বাংলা:** LangChain একটি orchestration framework — মডেল, prompt, retrieval, tool জোড়ে; এর নিজস্ব বুদ্ধি নেই, কাজটা মডেল করে

Q51. Which statement best describes Ollama as presented in the chapter?

- A. A hosted application programming interface that sends data to a remote server.
- B. A diffusion model for generating images.
- C. A runtime that lets you run open-weight Large Language Models locally on your own machine.
- D. A dataset of web text.

Answer: C. Ollama is a runtime for local execution of open-weight models (e.g. LLaMA-3 8B). A describes the opposite — a hosted service; the whole point of Ollama is keeping inference local and private. **বাংলা:** Ollama হলো লোকাল রানটাইম — নিজের মেশিনে open-weight মডেল চালায়; এটা remote hosted সেবা নয়

Q52. Why is quantization (e.g. 4-bit or 8-bit integer weights) useful for local inference?

- A. It increases the model's accuracy beyond the full-precision version.
- B. It removes the need for any graphics-processing unit.
- C. It converts the model into a diffusion model.
- D. It reduces the memory the model needs so it can fit in consumer hardware, at some cost to capability.

Answer: D. Quantization shrinks weight precision so the model fits in limited consumer memory; the chapter notes a quantized model is weaker than a frontier model — a capability cost. A wrongly claims accuracy improves. **বাংলা:** quantization ওজনের নির্ভুলতা কমিয়ে মেমরি বাঁচায়, তাই consumer hardware-এ চলে — তবে সক্ষমতা কিছুটা কমে

Q53. What is the central trade-off the chapter draws between local inference and hosted application programming interfaces?

- A. Local inference improves privacy but reduces capability and convenience, whereas hosted interfaces offer higher capability but send data off the premises and charge per token.
- B. Local inference is always free; hosted interfaces always cost more and are slower.
- C. Hosted interfaces keep data fully on the user's device.
- D. There is no meaningful difference between the two.

Answer: A. The trade-off is privacy/control versus capability/convenience: local keeps data private but is weaker, hosted is stronger but data leaves the premises with per-token cost. B is wrong because local still costs memory, compute, and electricity. **বাংলা:** trade-off হলো গোপনীয়তা বনাম সক্ষমতা — লোকালে গোপনীয়তা বেশি কিন্তু দুর্বল; hosted-এ শক্তিশালী কিন্তু ডেটা বাইরে যায় ও per-token খরচ

Q54. Which pair correctly names two foundation models of *different* modalities, as mentioned in the chapter?

- A. GPT-4 (text) and GPT-3.5 (text)
- B. Stable Diffusion (image) and Whisper (audio)
- C. LangChain (framework) and Ollama (runtime)

- D. Common Crawl (data) and Sora (video)

  Answer: B. Stable Diffusion is an image foundation model and Whisper an audio one — two different modalities. A lists two text models; C lists tooling, not models; D pairs a dataset with a model. **বাংলা:** Stable Diffusion (ছবি) আর Whisper (অডিও) — দুটি আলাদা modality; বাকিগুলো একই ধরন বা টুল/ডেটা

Topic: Course Map, Glossary & Engineering Perspective (Sections 1.0, 1.6)

Q55. Which statement best describes the engineering perspective that defines "AI Engineering" in this course?

- A. AI Engineering is mainly about training new foundation models from scratch.
- B. AI Engineering is the study of acoustic models only.
- C. AI Engineering is about building reliable products on top of pre-trained foundation models, rather than training them.
- D. AI Engineering means hand-coding rules for each task.

  Answer: C. The course frames AI Engineering as building reliable products on existing foundation models, since pretraining is prohibitively expensive for most. A states the opposite of the chapter's thesis. **বাংলা:** AI Engineering মানে নতুন মডেল ট্রেইন করা নয়, বরং তৈরি foundation model-এর ওপর নির্ভরযোগ্য প্রোডাক্ট বানানো

Q56. The exam cover sheet instructs: "Use the technical terms from the lecture. Do not use abbreviations." Which answer best reflects this rule?

- A. Write "LLM" because it is shorter and clearer.
- B. Avoid technical terms entirely and use everyday words.
- C. Use abbreviations only for foundation models.
- D. Write "Large Language Model" in full rather than "LLM", and similarly spell out other technical terms.

  Answer: D. The instruction requires the full lecture technical terms with no abbreviations — e.g. "Large Language Model", not "LLM". A and C violate the no-abbreviation rule; B drops the required terminology. **বাংলা:** নিয়ম হলো পূর্ণ পারিভাষিক শব্দ লেখা, সংক্ষিপ্ত রূপ নয় — "LLM" নয়, "Large Language Model"

Answer Key

Q#	Ans	Q#	Ans	Q#	Ans	Q#	Ans
Q1	A	Q15	C	Q29	A	Q43	C
Q2	B	Q16	D	Q30	B	Q44	D
Q3	C	Q17	A	Q31	C	Q45	A
Q4	D	Q18	B	Q32	D	Q46	B
Q5	A	Q19	C	Q33	A	Q47	C
Q6	B	Q20	D	Q34	B	Q48	D
Q7	C	Q21	A	Q35	C	Q49	A
Q8	D	Q22	B	Q36	D	Q50	B
Q9	A	Q23	C	Q37	A	Q51	C
Q10	B	Q24	D	Q38	B	Q52	D
Q11	C	Q25	A	Q39	C	Q53	A
Q12	D	Q26	B	Q40	D	Q54	B
Q13	A	Q27	C	Q41	A	Q55	C
Q14	B	Q28	D	Q42	B	Q56	D

Distribution count: A = 14, B = 14, C = 14, D = 14 (total = 56).