Dartmouth GEISEL SCHOOL OF MEDICINE | ACADEMY OF EDUCATORS AND SCHOLARS

Creating High-Quality Exam Questions with Generative AI

Michael W. Lee, PhD, FAAPE
Professor
Geisel School of Medicine
Dartmouth

A 35-year-old man with a history of depression and anxiety presents to his primary care physician complaining of persistent feelings of sadness, loss of interest in activities, and difficulty concentrating. He reports that these symptoms have been present for several months and are affecting his work performance and relationships. The patient has tried various forms of therapy without significant improvement and is now seeking pharmacological treatment. The physician decides to prescribe a medication that increases the synaptic concentration of norepinephrine and serotonin.

Which of the following drugs would most likely achieve this effect by inhibiting the reuptake of norepinephrine into the adrenergic neuron?

A) Phenelzine
B) Reserpine
C) Amphetamine
D) Cocaine
E) Guanethidine

LO: Describe the mechanism by which drugs inhibit the reuptake of norepinephrine into the adrenergic neuron.

Answer Explanations

D) Cocaine is a drug that inhibits the norepinephrine transporter (NET), which is responsible for the reuptake of norepinephrine from the synaptic cleft back into the presynaptic neuron. By blocking NET, cocaine prevents the reuptake of norepinephrine, thereby increasing its concentration in the synaptic cleft and prolonging its action on postsynaptic adrenoceptors. This effect can lead to increased stimulation of both α and β adrenoceptors, which may improve symptoms of depression and anxiety.

A) Phenelzine is a monoamine oxidase inhibitor (MAOI) that prevents the breakdown of norepinephrine within the neuron. While it does increase synaptic levels of norepinephrine, it does not do so by inhibiting reuptake.

B) Reserpine works by inhibiting the vesicular monoamine transporter (VMAT), which is responsible for sequestering neurotransmitters like norepinephrine into synaptic vesicles. This leads to depletion of norepinephrine stores, not inhibition of reuptake.

C) Amphetamine induces the release of norepinephrine from the presynaptic neuron into the synaptic cleft. Although it increases synaptic norepinephrine levels, its primary mechanism is not the inhibition of norepinephrine reuptake.

E) Guanethidine blocks the release of norepinephrine from the presynaptic neuron by interfering with the action potential-triggered release mechanism. It does not inhibit the reuptake of norepinephrine.

Prompt design is an art unto itself..

DARTMOUTH

Perfecting the Prompt

Intro to Prompt Engineering

Simon Stone

Research Software Engineer for HPC & AI

May 27, 2025

Prompt engineering

Anatomy of a prompt

A good prompt contains some or all of the following components:

- ☝ Instruction
  - "Summarize the following article."
- 👉 Context
  - "The summary will appear on social media and should be engaging and energetic."
- 📄 Input Data
  - the full text of the article
- 🖨 Output indicator
  - "Format your response using markdown and emojis."

dartgo.org/geisel-perfect-prompt

Medical Science Educator (2025) 35:611–613
https://doi.org/10.1007/s40670-025-02334-7

INNOVATION

Check for updates

Development of a Universal Prompt as a Scalable Generative AI-Assisted Tool for USMLE Step 1 Style Multiple-Choice Question Refinement in Medical Education

Youngjin Cho[1] · Grace L. Park[2] · Gabi N. Waite[1] · Abhijay Mudigonda[3] · John L. Szarek[1]

Accepted: 14 February 2025 / Published online: 25 February 2025

Abstract
We developed a generative artificial intelligence (genAI)-assisted tool enabling learners to receive feedback on, revise, and clone multiple-choice questions aligned with learning objectives. Initially designed as a custom GPT, we adapted it to a universal prompt for platform-agnostic, equitable access. This innovation exemplifies readily adaptable genAI-enhanced learning driven by pedagogy.

Keywords Multiple-choice questions · Generative AI · Medical students · Prompts for medical education · Assessment writing

Medical Science Educator (2024) 34:1571–1576
https://doi.org/10.1007/s40670-024-02146-1

COMMENTARY

Beginner-Level Tips for Medical Educators: Guidance on Selection, Prompt Engineering, and the Use of Artificial Intelligence Chatbots

Yavuz Selim Kıyak[1]

Accepted: 8 August 2024 / Published online: 17 August 2024

Table 1 Prompt engineering techniques, their explanations, and weak and improved prompt examples in the context of medical education

Technique	Explanation	Weak prompt	Improved prompt
Be clear and direct	Clearly describe what you need. Consider adding contextual details about how the output will be used, who the output is meant for, which part of the workflow the task belongs to, and what successful output looks like	Analyze this new clinical guideline and give me the key points. Here is the guideline: [File]	I am a clinician. Analyze this new clinical guideline. Keep your response concise and list only the necessary information. Include: 1. Main changes or updates 2. Rationale behind the changes Here is the guideline: [File]
Use examples	Examples help clarify expectations and demonstrate the desired output format or style	Here are the learning objectives for my hypertension course: [Objectives] Correct their style	Here are the learning objectives for my hypertension course but I am not sure about the style: [Objectives] Here are the successful examples of learning objectives for Type 2 Diabetes course, generate learning objectives by following the same style: [Examples]
Give the model a role (persona)	Assigning a specific role to the model can enhance its responses by aligning them with a particular expertise or perspective. Useful in role-specific tasks or simulations	Explain the process for cardiac pacemaker insertion	You are a cardiology professor specialized in pacemakers. Explain the process for cardiac pacemaker insertion.
Chain prompts	Break complex tasks into smaller parts when a single prompt is not sufficient. It aids in managing complexity and improves response accuracy	Explain surgical treatment of direct inguinal hernia	First, explain the anatomy of the groin. Next, describe the mechanism behind direct inguinal hernia. Finally, explain its surgical treatment based on the anatomy.
Let the model think	Promote step-by-step reasoning to enhance the depth and analytical quality of responses, ideal for complex problem-solving	How to diagnose appendicitis?	How to diagnose appendicitis? Think step-by-step
Control output format	Specify the output format to maintain consistency and readability, important in tasks requiring structured data like reports or summaries	What are the main groups of antihypertensive drugs and their common side effects?	Provide a table listing the main groups of antihypertensive drugs and their common side effects
Ask the model for rewrites	Requesting rewrites based on specific feedback can refine and improve the model's output, useful in iterative editing or content improvement processes	Rewrite this case	Rewrite this case to make it understandable for first-year medical students.
Long context window (input and output) tips	The amount of content that an LLM can consider at one time is limited. Therefore, put the long document at the top and ask model to quote the relevant parts of the document	What are the practical tips for conducting simulated patient encounters based on this 500-page book?: [File]	Here is a 500-page book: [File] Find recommendations in the book that are relevant to conducting simulated patient encounters and quote them in bullet points. Then, based on these recommendations, provide a list of practical tips for medical educators.

Recent studies suggest LLM-generated medical questions can approximate the quality of faculty-authored questions

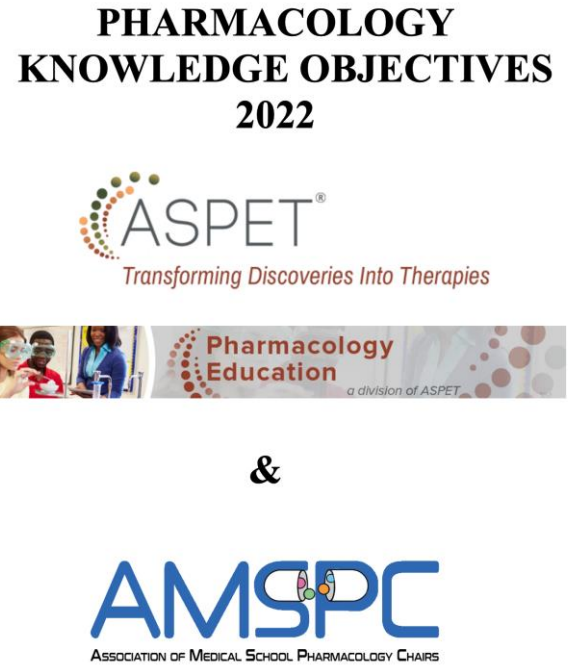
- Cheung et al.[1] compared 50 MCQs generated by ChatGPT (GPT-3.5) against 50 questions crafted by medical faculty
  - Three expert assessors evaluated these questions across five quality domains, finding statistical equivalence in overall quality and four specific domains, with only marginal inferiority in content relevance.
  - LLM **question generation** required approximately **20 minutes** compared to **211 minutes for human experts**—a tenfold efficiency improvement.
- ChatGPT-4-generated clinical vignettes are largely indistinguishable from human-authored questions when evaluated by medical experts[2]
  - Despite these promising results, 36% of ChatGPT-4-generated questions contained flaws identified by physician reviewers that make their deployment reliant on a human-in-the-loop check.

Citations in notes

Back to the Cocaine Question..

- Our item at the beginning was generated using RAG (Retrieval-Augmented Generation) which pulled the content from a sandboxed, curated set of resources:
  - Katzung and Trevor Basic and Clinical Pharmacology 16e
  - Goodman and Gilman's The Pharmacological Basis of Therapeutics 15e
  - NBME Writing Guide

The source of the prompts was:



Study Design

- Compared items generated from RAG and Non-RAG
  - RAG uses a knowledge base (i.e. database) to provide context for generating responses, while Non-RAG models rely solely on the pre-trained knowledge of the AI program.

- Over 156 learning objectives from the ASPET/AMSPC Knowledge Objectives served as prompts to generate 312 matched items (156 RAG and 156 Non-RAG items)

- Of these 100 matched items were randomly selected for each RAG and Non-Rag which were furnished to 3 expert human raters for review

- Review Criteria:
  - (1) content accuracy, (2) clinical relevance (5-point scale), (3) question clarity, (4) educational value, and (5) distractor quality, and (6) overall quality.

So what about the matched Non-RAG item...

A 36-year-old woman presents to her primary care physician with complaints of chronic fatigue, difficulty concentrating, and a persistent feeling of sadness. She mentions that these symptoms have been present for several months and have progressively worsened. She denies any significant changes in her life or recent stressors. Her medical history is unremarkable, and she is not currently taking any medications. After a thorough evaluation, the physician diagnoses her with major depressive disorder and decides to initiate pharmacotherapy aimed at increasing synaptic norepinephrine levels.

Which of the following medications is most likely to inhibit the reuptake of norepinephrine into the adrenergic neuron, thereby treating the patient's depressive symptoms?

A. Sertraline

B. Venlafaxine

C. Bupropion

D. Amitriptyline

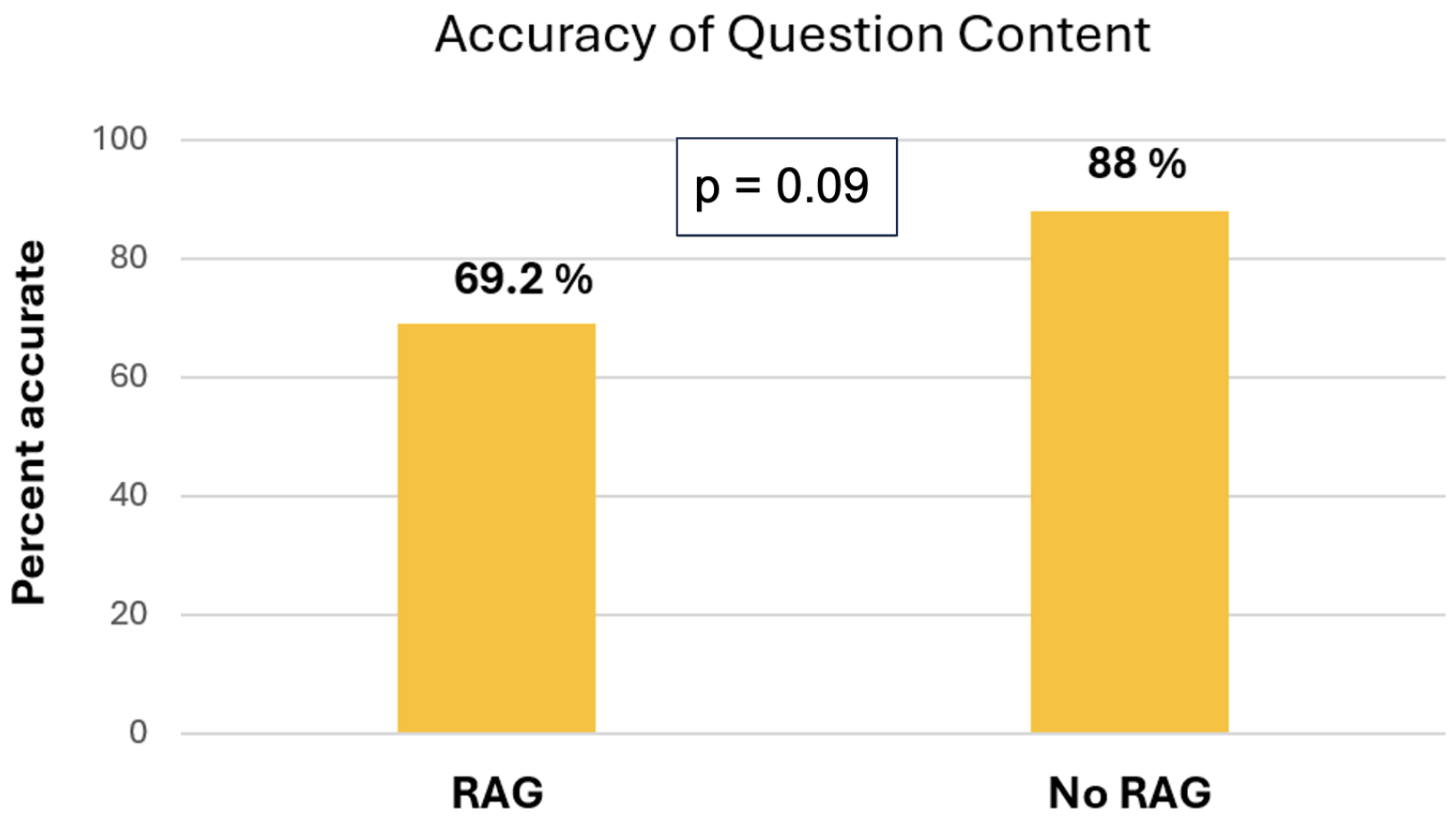
E. Phenelzine

LO: Describe the mechanism by which drugs inhibit the reuptake of norepinephrine into the adrenergic neuron.

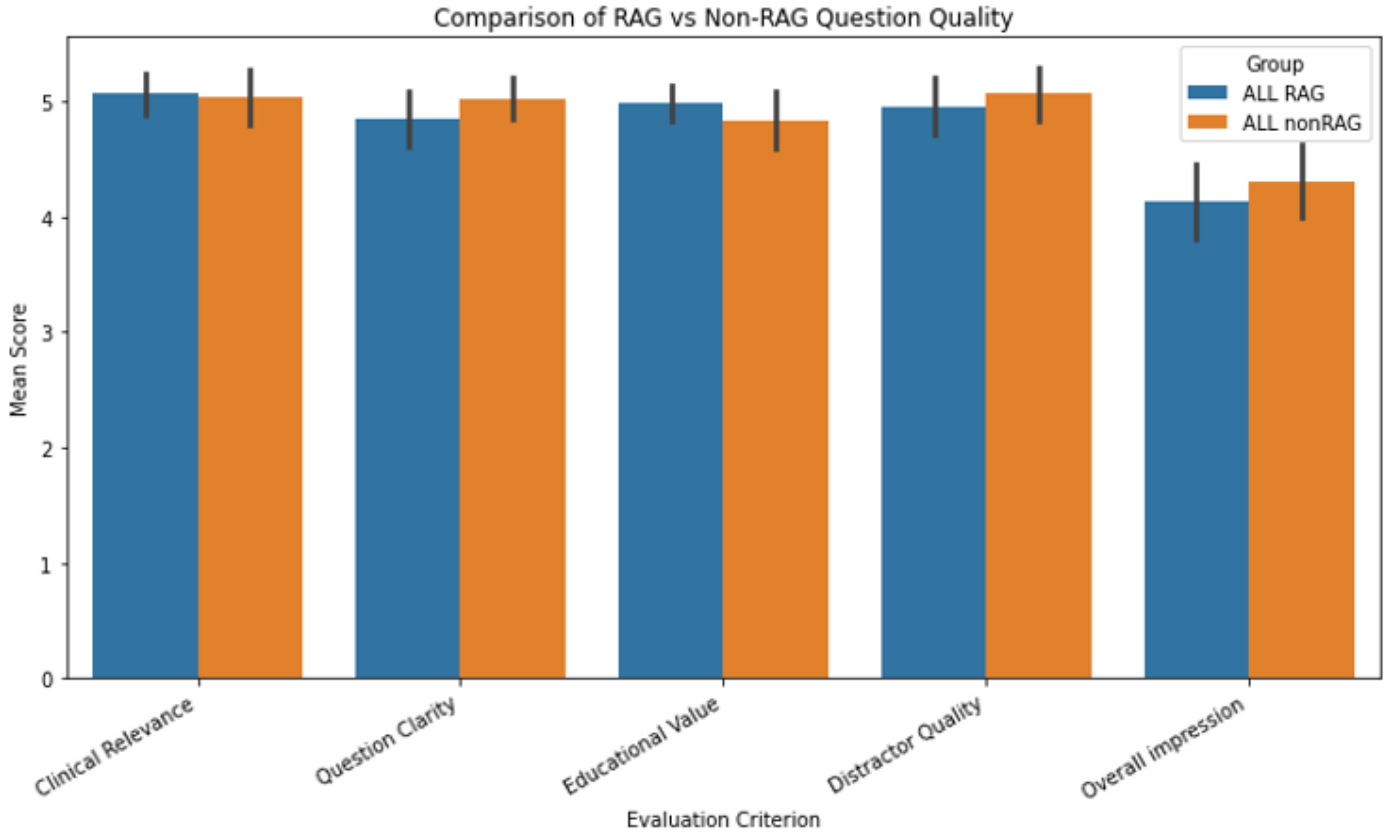
Explanation for the Correct Answer:
B. Venlafaxine is a serotonin-norepinephrine reuptake inhibitor (SNRI) that is commonly used in the treatment of major depressive disorder. It works by blocking the reuptake of both serotonin and norepinephrine into presynaptic neurons, which increases the concentration of these neurotransmitters in the synaptic cleft and enhances neurotransmission. The inhibition of norepinephrine reuptake is particularly relevant to the vignette's focus on increasing synaptic norepinephrine levels.

A. Sertraline is a selective serotonin reuptake inhibitor (SSRI). While it is used to treat depression by increasing synaptic serotonin levels, it does not primarily inhibit the reuptake of norepinephrine.
C. Bupropion is an atypical antidepressant that primarily inhibits the reuptake of dopamine and norepinephrine but has a minimal effect on serotonin reuptake. Although bupropion does inhibit norepinephrine reuptake, venlafaxine is more commonly associated with this mechanism in the context of treating depression.

D. Amitriptyline is a tricyclic antidepressant (TCA) that inhibits the reuptake of both norepinephrine and serotonin. However, it is not as selective as venlafaxine and is associated with a broader range of anticholinergic side effects.
E. Phenelzine is a monoamine oxidase inhibitor (MAOI) that prevents the breakdown of monoamine neurotransmitters (including norepinephrine, serotonin, and dopamine) within the neuron. While this increases the levels of norepinephrine, it does not do so by inhibiting its reuptake.



Comparison of question content accuracy between retrieval-augmented generation (RAG) and non-RAG A.I. frameworks (n= 50 items, n= 3 reviewers). Questions generated without RAG demonstrated higher accuracy (88%) compared to those with RAG-based generation (69.2%), approaching significance (p = 0.0981, Fisher's exact test).



Mean quality scores with standard error of the mean (SEM) for RAG vs. Non-RAG generated questions across five evaluation criteria. No statistically significant differences were observed (n= 3 reviewers).

Who are you and what are you doing here?

Small Group Exercise (30-40min):

1. Join a group (3-4 people if possible).
2. Research if your discipline of clinical specialty society curates content learning objectives or knowledge objectives.
3. Pick 1 learning objective for each unique discipline or clinical specialty and use the LO to generate a USMLE style item in ChatGPT *(or any AI service you prefer or sign up for free access to ChatGPT right now: https://chatgpt.com/)*
4. Share the AI generated item with the group, review for accuracy and clinical appropriateness, edit it if necessary, or repeat the process if it is a lost cause item.
5. Run the edited item back through ChatGPT to check for accuracy.
5. Pick an item to present to the larger group and share your experience.

Activity	Description	Outcome	Time
Part 1: Prompt Identification or Generation	Identify if your discipline/clinical society has curated learning objectives. If not, select a personal learning objective you wrote.	Individuals identify 1-2 prompts for AI item generation.	~10min
Part 2: AI Item Generation	In groups of 3–4, participants compare prompts and generate USMLE-style MCQs using ChatGPT or a similar AI tool for each prompt.	Groups produce several AI-generated draft items and review.	~15min
Part 3: Peer Review	Groups exchange questions and evaluate/refine each other's MCQ for: (1) content accuracy, (2) clinical relevance (5-point scale), (3) question clarity, (4) educational value, and (5) distractor quality, and (6) overall quality. Return revised item to group.	Groups revise the MCQs based on feedback, improving question quality.	~15nin
Part 4: Group Share and Wrap-Up	Groups present refined items, highlight changes, strengths, and areas for improvement. Optional reflection on the use of AI.	Shared learning, deeper understanding of strengths/limits of AI assistance and value of human raters.	~20min

Summary


Step 1: Design of RAG and Non-RAG A.I. Frameworks
Step 2:Generation of matched sets of RAG and Non-RAG neuropharmacology items using ASPET/AMSPC Knowledge Objectives
Step 3: Item Review to remove incomplete or identical items within each framework
Step 4: Recruitment of reviewers and item review
Step 5: Data cleaning and analysis

Summary

- Consult the educational literature on prompt design and educational theory.
- Collaboration with colleagues is the key!
- A team including a clinician and basic scientist can tackle multiple angles of item review.
- Experiment!

Type	Sample explanation
Clinical feature list (FL)	Stroke is the fourth leading cause of death in Western societies. Like other manifestations of arteriosclerosis, risk of stroke increases with age. Brainstem strokes are due to thrombosis of either the basilar artery or one of its branches. It typically evolves to its maximal deficit within a few hours although rarely the onset of the deficit is intermittent or stuttering. While the exact symptoms and signs of a stroke in the brainstem may vary depending on the area of damage, most patients will develop double vision (diplopia) and difficulty swallowing (dysphagia). In addition, patients have difficulty maintaining their posture and if able to walk, are very unsteady on their feet (ataxia). Patients with brainstem strokes are incontinent of urine. Neurological examination shows muscle weakness with increased stiffness or tone of the muscles (spasticity). The weakness is more obvious proximally. There are increased deep tendon reflexes. While the physical signs of weakness may involve one side of the body and the face, it is possible with a brainstem stroke to see involvement of all four limbs including the face. Vertigo is reported by some patients.
Science-based causal explanations (SC)	Most brainstem strokes are due to a blood clot or thrombosis blocking one of the intracranial blood vessels. While this can present as a sudden event, the symptoms may follow a "stuttering" course when the circulation is only partially blocked. Nerves which control the eyeball muscles arise in the brainstem and loss of their function leads to double vision (diplopia). Similarly, the swallowing muscles are compromised (dysphagia) and incoordination of the muscles around the mouth produces slurred and clumsy speech (dysarthria). The brainstem is also the higher control center for the bladder and thus incontinence can develop. Coordination of balance is located here with the result that there is unsteady gait or posture. All the motor and sensory pathways from the cerebral cortex pass down through the brainstem which is part of the central nervous system. Injury here removes central nervous system control of lower peripheral neurons in the spinal cord, resulting in a pattern of upper motor neuron dysfunction characterized by increased muscle tone (spasticity), increased deep tendon reflexes and muscle weakness that is worse in the proximal muscles of the hip and shoulder girdle. Because motor pathways traveling to each side of the body are close together in the brainstem, the overall pattern of weakness can involve either one side of the body including the face, or all four limbs including the face.
Separated—basic science prior to clinical features (BC)	Intracranial blood vessels feed the brainstem which is an important part of the central nervous system (CNS) for coordination of several functions. Nerves which control the eyeball muscles arise in the brainstem. The muscles that control swallowing and the movement of the mouth are also coordinated from the brainstem. Other organs controlled by the brainstem include the bladder. However, the most noticeable function of the brainstem is coordination of balance. Motor and sensory pathways from cerebral cortex to limbs on both sides of the body pass directly through the brainstem. This makes the brainstem crucial for CNS control of lower peripheral neurons of the spinal cord. Brainstem strokes are usually due to clots or thrombosis. Partial blocking leads to a stuttering presentation of symptoms. Symptoms can present on one side of the body and face or all four limbs and the face. These generally include increased muscle tone (spasticity), increased deep tendon reflexes, and muscle weakness that is worse in the proximal muscles of the hip and shoulder girdle. Incontinence is also a possibility. Symptoms in the face are likely to be slurred or clumsy speech (dysarthria), difficulty swallowing (dysphagia), and double vision (diplopia). Difficulty with balance is also very likely.
Separated—clinical features prior to basic science (CB)	Brainstem strokes are usually due to clots or thrombosis. Partial blocking leads to a stuttering presentation of symptoms. Symptoms can present on one side of the body and face or all four limbs and the face. These generally include increased muscle tone (spasticity), increased deep tendon reflexes, and muscle weakness that is worse in the proximal muscles of the hip and shoulder girdle. Incontinence is also a possibility. Symptoms in the face are likely to be slurred or clumsy speech (dysarthria), difficulty swallowing (dysphagia), and double vision (diplopia). Difficulty with balance is also very likely. Intracranial blood vessels feed the brainstem which is an important part of the central nervous system (CNS) for coordination of several functions. Nerves which control the eyeball muscles arise in the brainstem. The muscles that control swallowing and the movement of the mouth are also coordinated from the brainstem. Other organs controlled by the brainstem include the bladder. However, the most noticeable function of the brainstem is coordination of balance. Motor and sensory pathways from cerebral cortex to limbs on both sides of the body pass directly through the brainstem. This makes the brainstem crucial for CNS control of lower peripheral neurons of the spinal cord.

Kulasegaram, Kulamakan; Manzone, Julian C.; Ku, Cheryl; Skye, Aimee; Wadey, Veronica; Woods, Nicole N. Academic Medicine90(11):S63-S69, November 2015. doi: 10.1097/ACM.0000000000000896