LIMITATIONS

While our work enhances SLM performance by using keyword representations of raw data, it only mitigates but does not eliminate privacy concerns. Given that FTC is based on GPT3.5, the medical knowledge it generates may be inaccurate or biased, which can impact SLM performance. Moreover, the inference time for LLMs may be slower than that of SLMs, leading to longer overall inference times compared to models solely reliant on local SLMs. Furthermore, due to the training cut-off time, the medical knowledge in LLM could be outdated, potentially hindering medical decision-making. We aim to integrate LLM with the internet and knowledge graph in future work to generate more reliable medical knowledge for enhancing SLM decision-making. These issues underscore the need for further research on the use of LLMs in privacy-restricted medical scenarios.

A APPENDIX

A.1 DATA AND CODE

Our codes and generated data are public at https://anonymous.4open.science/r/PrivacyBoost-SLM-91F2.

A.2 SLM IMPLEMENTATION AND TRAINING DETAILS

We implement both SFT and FTC based on huggingface transformers Wolf et al. (2020), and train on NVIDIA A40-48GB GPUs. For all datasets, we utilize AdamW (Loshchilov and Hutter, 2019) as optimizer. For MedQA and HEADQA, we set learning rates of 5×10^{-5}, 5×10^{-5}, and 2×10^{-6} for BioLinkBERT-Base, BioLinkBERT-Large, and BioMedLM in both FTC and SFT settings. For MedMCQA, we set learning rates of 2×10^{-5}, 2×10^{-5}, and 2×10^{-6} for BioLinkBERT-Base, BioLinkBERT-Large, and BioMedLM in both FTC and SFT settings. For BioLinkBERT-Base and BioLinkBERT-Large, we limit training to 100 epochs with a 200-step warm-up and apply early stopping after 5 epochs without validation improvement. Batch sizes are 8 for few-shot and full-training scenarios across all datasets. For BioMedLM, we set the training epochs to 10 for all datasets. We run experiments with three random seeds {0, 1, 2} and report mean results and standard deviations.

A.3 ADDITIONAL EXPERIMENTAL RESULTS

Development sets results of MedQA and HeadQA. We report development sets results of MedQA and HeadQA in Table 8 and 9, and Figure 7.

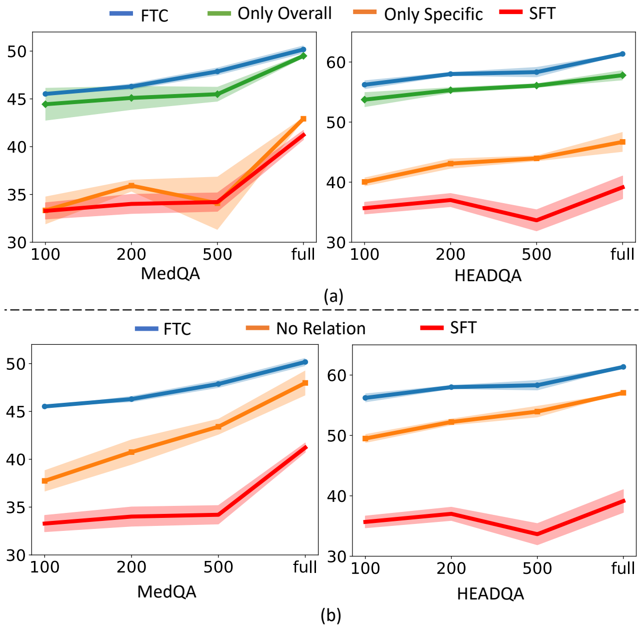
Table 8: Results (%) on development sets of MedQA and HEADQA between LLM, SFT, and FTC under different training sizes.

	MedQA				**HEADQA**			
	100	200	500	full	100	200	500	full
LLM	38.30				47.60			
SFT	$33.28_{0.85}$	$34.01_{1.00}$	$34.20_{0.96}$	$42.21_{0.91}$	$36.54_{0.30}$	$39.00_{0.61}$	$34.88_{2.00}$	$41.48_{0.48}$
FTC	$\mathbf{43.66}_{1.26}$	$\mathbf{45.70}_{0.10}$	$\mathbf{45.62}_{0.70}$	$\mathbf{50.73}_{0.35}$	$\mathbf{55.03}_{1.10}$	$\mathbf{56.18}_{0.76}$	$\mathbf{57.45}_{0.53}$	$\mathbf{60.21}_{1.47}$

Table 9: Results comparison of FTC and FTCR in the full-training setting on development sets of MedQA and HEADQA.

	MedQA	**HEADQA**
FTCR	$48.56_{0.62}$	$58.17_{0.78}$
FTC	$\mathbf{50.73}_{0.35}$	$\mathbf{61.35}_{0.16}$

Figure 7: Accuracy comparison (%) of ablation studies on development sets of MedQA and HEADQA. The upper part of the table examines the effect of different context components on SLM training, while the lower part investigates the impact of relationships within the context.



A.4 Alternative measurement for privacy budget

We further conducted BPC measurement experiments on the training, validation, and test sets of the MedQA dataset, respectively, to demonstrate the effectiveness of using keywords to represent raw medical data while keeping privacy. To evaluate the BPC of raw data and keywords, we separately input keywords with various proportions and raw data into BioMedLM across different sets. Specifically, for each list of keywords, we form a sentence of keywords by concatenating the list of keywords and separating each pair of keywords with an empty space. Subsequently, we calculate the corresponding BPC values to assess the outcomes. The results are shown in Table 10.

Table 10: Comparison of BPC values for raw data and concatenated keywords across different data splits of the MedQA dataset

	Raw Data		Keywords		
	-	25%	50%	75%	100%
Training	12.41	10.37	10.04	9.91	9.87
Validation	12.44	10.31	10.01	9.92	9.89
Test	12.44	10.32	10.06	9.94	9.91

The BPC values of keywords consistently exhibit lower values than those of the raw data.

A higher BPC is consistently observed across various subsets of the MedQA dataset when compared to the approach of inputting keywords into BioMedLM. This comparison implies that, on average, raw data holds a greater level of uncertainty in contrast to the utilization of keywords. This disparity could be attributed to the fact that raw data encompasses a larger amount of medical-unrelated information, which includes privacy-related data, despite its comprehensive information coverage.

A decrease in the proportion of keywords results in an increase in BPC.

Decreasing the number of keywords fed into the BioMedLM leads to an increase in BPC. This indicates that a higher volume of keywords contributes to a more meaningful representation of medical information in raw data, subsequently enhancing the performance of both LLM and FTC.

Keywords representation obtains the lowest BPC compared to random words and random span. We further evaluate the BPC of different raw data representation methods (random span, random words and keywords) while maintaining the privacy budget the same. Specifically, we feed different data representation content into BioMedLM on the training, validation and test sets of the MedQA dataset, and calculate the corresponding BPC values respectively. The results are presented in Table 11.

Table 11: BPC values different raw data representation methods.

Representation Methods	Keywords	Random words	Random span
Training	9.87	10.04	12.25
Validation	9.89	10.05	12.22
Test	9.91	10.04	12.22

Keyword representation consistently obtains the lowest BPC compared to the other two data representation methods, providing a more effective representation of medical knowledge within the same privacy budget. Interestingly, the order of BPC performance aligns with the performance of LLM prompting and SLM fine-tuning in Table 7 of our paper. The method with a lower BPC value achieves better performance in both LLM prompting and SLM fine-tuning. This further highlights the importance of raw data representation in generating high-quality context and effective SLM training.

A.5 General domain experimental setup

We perform experiments on two commonsense datasets to demonstrate the broad applicability of our approach.

Datasets. 1. CommonsenseQA (Talmor et al., 2018) is a multi-choice question-answering dataset featuring 5 options per question, requiring commonsense reasoning. The dataset is split into 9741/1221/1140 instances for training, development, and test sets, respectively. As the test set is not publicly accessible, we follow previous work (Li et al., 2022) and report results on the development set. **2. OpenbookQA** is a 4-option multi-choice question-answering dataset that demands open book facts, broad common knowledge, and multi-hop reasoning (Mihaylov et al., 2018). The dataset is split into 4957/500/500 instances for training, development, and test sets, respectively. We report results on the test set.

Context generation from LLM. We utilize the grounded entities from Zhang et al. (2022) as keywords to query the GPT-3.5 *gpt-3.5-turbo* engine, generating context through a greedy decoding process (by setting the temperature to 0). We employ the same demonstration format as in the medical domain and create a privacy-restricted context in accordance with the in-context learning paradigm. For each dataset, we supply seven-shot hand-crafted examples.

SLM training. We employ T5-base (Raffel et al., 2020) as the SLM backbone for both SFT and FTC. In FTC, we use Fusion-in-Decoder (Izacard and Grave, 2020) to incorporate context information for decision-making. Specifically, each context related to a candidate answer is concatenated with the question and all candidate answers, then processed independently by the encoder. The concatenated representations of all contexts are subsequently fed into the decoder to generate predictions. In both datasets, we utilize AdamW (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-5} for both SFT and FTC. We limit training to 100 epochs with a 200-step warm-up and apply early stopping after 5 epochs without validation improvement. The batch size is set to 8 for both datasets.

A.6 Prompt details

In this section, we present examples of prompts for both medical and general domains. The context of each example consists of three parts: (1) An overall context, which provides high-level information derived from the extracted keywords and candidate answers (red); (2) A specific context, which focuses on the knowledge associated with a candidate answer (blue) and its relation to the overall context (green); and (3) A preliminary decision, which draws a conclusion based on the contexts provided earlier (orange).

Prompts for medical datasets. Three clinicians were involved in the prompt design and writing. We held 5 meetings with clinicians to discuss the prompt design and iterate 4 versions. The final version of context needs around 10 minutes to write per context. Our medical prompts on *MedQA* and *MedMCQA* are based on Singhal et al. (2022), *HEADQA* is based on Wikipedia and written and verified by clinicians. Here we provide the prompts that we used in our experiments.

Table 12: Prompts for MedQA

Question Keywords: male, marathon runner, office, complaint, right-sided rib pain, Physical examination, normal heart, lung findings, exhalation, dysfunction, ribs 4-5, right, muscles, muscle groups, dysfunction, direct method
Candidate Answers: (a) anterior scalene (b) latissimus dorsi (c) pectoralis minor (d) quadratus lumborum

Context: Normal heart and lung findings on a physical exam coupled with evidence of exhalation dysfunction in ribs 4-5 on the right suggest a musculoskeletal cause of exertional chest pain.
(a): The anterior scalene muscle attaches to the first rib. It is not associated with exhalation dysfunction in ribs 4-5.
(b): The latissimus dorsi muscle attaches to ribs 9 and 10. It is not associated with exhalation dysfunction in ribs 4-5.
(c): The pectoralis minor muscle is attached to ribs 3, 4, and 5. Dysfunction in the fourth and fifth ribs can be caused by issues with the pectoralis minor muscle due to its attachment to these ribs. It is associated with exhalation dysfunction in ribs 4-5.
(d): Quadratus lumborum muscle attaches to ribs 11 and 12. It is not associated with exhalation dysfunction in ribs 4-5.
Therefore, the answer is (c).

Question Keywords: male, office, low back pain, denies, any, trauma, says, truck, day, job, Examination, patient, prone position, deep sacral, left, posterior inferior lateral angle, right, lumbosacral junction, springs, freely, compression, diagnosis
Candidate Answers: (a) left-on-left sacral torsion (b) left-on-right sacral torsion (c) right unilateral sacral flexion (d) right-on-right sacral torsion

Context: The physical exam shows the deep sacral sulcus on the left, a posterior inferior lateral angle on the right and normal spring test.
(a): This condition is characterized by the deep sacral sulcus on the right, a posterior inferior lateral angle on the left and normal spring test. It is not consistent with the findings from the physical exam.
(b): The left-on-right sacral torsion would be indicated by a deep sacral sulcus on the right, a posterior inferior lateral angle on the left, and a positive spring test. It is not consistent with the findings from the physical exam.
(c): This condition is characterized by a posterior inferior lateral angle on the right, a deep sacral sulcus on the right, and an absence of normal spring test. It is not consistent with the findings from the physical exam.
(d): This condition is characterized by a deep sacral sulcus on the left, a posterior inferior lateral angle on the right and normal spring test. It is consistent with the findings from the physical exam.
Therefore, the answer is (d).

Question Keywords: man, comes, office, nonproductive cough, runny nose, frontal headache, headache, morning, nd, ibuprofen, relief, not, shortness of breath, Medical history, no medications, ibuprofen, pain, Vital signs, temperature, 37.4, °, 99.4, °, pulse, 88/min, 18/min, blood pressure, 120/84, Examination, nares, erythematous, mucous membranes, Examination, throat, erythema, follicular lymphoid hyperplasia, posterior oropharynx, no, cervical adenopathy, Lungs, clear, auscultation, patient's, symptoms
Candidate Answers: (a) Allergic rhinitis (b) Epstein-Barr virus (c) Mycoplasma pneumonia (d) Rhinovirus

Context: Sore throats are common symptoms in multiple upper respiratory viruses.
(a): A non-productive cough is a common symptom in upper respiratory viruses but is not present in allergic rhinitis. It is not the cause of the symptoms.

Continued on next page

Table 12 – *Continued from previous page*

(b): The absence of shortness of breath indicates mycoplasma is less probable. It is not the cause of the symptoms.
(c): Cervical adenopathy is commonly seen in cases of Epstein Barr virus. The absence of cervical adenopathy indicates Epstein Barr virus is less likely. It is not the cause of the symptoms.
(d): Rhinovirus can cause this patient's symptoms, including sore throat, runny nose and a frontal headache. It is the cause of the symptoms.
Therefore, the answer is (d).

Question Keywords: healthy, woman, comes, physician, 8, months, husband, killed, car crash, decreased, difficulty falling asleep, states, sad, cries, frequently, door lock, five, house, five, pieces, toilet paper, perfectionist, urges, rituals, Pharmacotherapy, neurotransmitters
Candidate Answers: (a) Dopamine (b) Glutamate (c) Norepinephrine (d) Serotonin

Context: The woman is exhibiting symptoms of major depressive episodes, such as difficulty falling asleep, frequent crying, and a persistent feeling of sadness.
(a): Dopamine is a neurotransmitter that increases positive emotions. It is implicated in many disease processes, including Parkinson's and ADHD, and is targeted by antipsychotic medications but not used as a sleep aid. It is not a treatment for the patient's symptoms.
(b): Glutamate is a neurotransmitter that is associated with multiple neurological disorders including epilepsy, stroke, and autism. It is not a treatment for the patient's symptoms.
(c): Norepinephrine is a catecholamine with adrenergic properties. It is not a treatment for the patient's symptoms.
(d): Serotonin is a neurotransmitter which is the target for multiple antidepressants, anxiolytics, and antipsychotics. It could be a treatment to address the patient's symptoms of depression and anxiety.
Therefore, the answer is (d).

Question Keywords: man, comes, office, preoperative, evaluation, adrenalectomy, scheduled, 2, weeks, One, month, care, emergency department, pain, right flank, motor vehicle collision, blood pressure, 160/100, mm Hg and CT scan, abdomen, incidental, left adrenal mass, laboratory studies, complete blood count, serum electrolyte concentrations, liver function tests, reference ranges, patient, healthy, elevated blood pressure, no medications, follow-up visit, office 2, weeks, disclosed, elevated, urinary normetanephrine, metanephrine, plasma, concentrations, patient, surgeon, recommended, adrenalectomy, vital signs, temperature, 36.6, 97.9, pulse, 100/min, 14/min, blood pressure, 170/95, Physical examination, no significant, findings, preoperative, preparation, treatment
Candidate Answers: (a) Labetalol (b) A loading dose of potassium chloride (c) Nifedipine (d) Phenoxybenzamine

Context: The patient is being evaluated for adrenalectomy due to a large left adrenal mass, which is likely causing elevated blood pressure as a symptom of pheochromocytoma. Elevated urinary normetanephrines confirm the diagnosis.
(a): This beta-blocker works by blocking the effects of adrenaline and other stress hormones on the heart and blood vessels. It is not a treatment for pheochromocytoma.
(b): The use of a potassium chloride loading dose is a treatment specifically for hypokalemia, which is a condition where there are abnormally low levels of potassium in the blood. It is not a treatment for pheochromocytoma.
(c): This drug is commonly prescribed to treat high blood pressure and angina. It can also help relieve symptoms of Raynaud's phenomenon. It is not a treatment for pheochromocytoma.
(d): This medication is used as a preoperative preparation treatment to block alpha-adrenergic receptors in the body and it effectively treats hypertension caused by pheochromocytoma. It is a treatment for pheochromocytoma.
Therefore, the answer is (d).

Table 13: Prompts for HEADQA

Question Keywords: autosomal dominant trait

Continued on next page

Table 13 – *Continued from previous page*

Candidate Answers: (a) The trait appears more frequently in males. (b) The unaffected people do not transmit the trait. (c) The trait tends to skip generations. (d) The affected people have both affected parents. (e) The trait tends to appear in the progeny of related parents.

Context: Autosomal dominant inheritance is a mode of genetic transmission in which a trait or condition can be passed down from parent to child. One copy of a mutated gene from one parent can cause the genetic condition. For example, let ’A’ represent the affected allele and ’a’ represent the unaffected allele. An affected person may have the genotype AA or Aa, while an unaffected person has the genotype aa. Consequently, an individual with genotype AA has a 100% chance of passing on the affected allele, and someone with genotype Aa has a 50% chance of doing so.
(a): Autosomal dominant inheritance is not influenced by an individual’s sex, as it is not sex-dependent. The expression of the trait occurs regardless of gender. It is not a characteristic of autosomal dominant inheritance.
(b): Unaffected individuals do not have the mutated gene and therefore cannot transmit the trait. It is a characteristic of autosomal dominant inheritance.
(c): Autosomal dominant traits can be passed down through multiple generations. Since the affected allele is dominant, an individual will express the trait as long as they inherit the affected gene. It is not a characteristic of autosomal dominant inheritance.
(d): Only one affected parent is needed to transmit on the autosomal dominant trait to their child. It is not a characteristic of autosomal dominant inheritance.
(e):A dominant gene can appear in any progeny, regardless of the parent. It is not a characteristic of autosomal dominant inheritance.
Therefore, the answer is (b).

Question Keywords: caring, patient, supraglottic laryngectomy
Candidate Answers: (a) He has lost the ability to speak by extirpation of the true vocal cords. (b) The tracheostomy they have performed will be permanent. (c) You have a risk of bronchoaspiration due to difficulty swallowing. (d) You may have constipation due to cervical dissection. (e) A portion of the larynx has been removed along with a vocal cord.

Context: Supraglottic laryngectomy or horizontal partial laryngectomy is an operation to remove the epiglottis, false vocal cords, and superior half of the thyroid cartilage.
(a): Supraglottic laryngectomy removes the false vocal cords, but the true vocal cords are not affected, and the patient’s ability to speak should not be significantly impacted. It is not typical to lose the ability to speak by extirpation of the true vocal cords.
(b): If a tracheostomy tube is in place after the procedure, it is typically removed within 24-48 hours of surgery. It is not typical to involve A permanent tracheostomy as a part of the supraglottic laryngectomy process.
(c): Supraglottic laryngectomy results in severe disturbance to the swallowing mechanism by removal of protective layers and sensation. There is an increased risk of bronchoaspiration. It is related to care of patients with supraglottic laryngectomy.
(d): Constipation is not a side effect of supraglottic laryngectomy. It is not related to care of patients with supraglottic laryngectomy.
(e): Supraglottic laryngectomy is an operation to remove the epiglottis, false vocal cords, and superior half of the thyroid cartilage. In this procedure, the true vocal cords are not typically affected, preserving the patient’s ability to speak as much as possible. It is not common to remove a portion of the larynx along with a true vocal cord during this procedure.
Therefore, the answer is (c).

Question Keywords: estrogenic treatment, adverse effects, NOT, adverse effect, pharmacological action
Candidate Answers: (a) Edema (b) Breast pain (c) Ovarian cancer (d) Sickness (e) Headaches

Context: Estrogen therapy involves supplementing a patient with estrogen, the primary female sex hormone. Potential side effects include breast tenderness or swelling, edema, nausea, leg cramps, endometrial cancer, and more.

Continued on next page

Table 13 – *Continued from previous page*

(a): Edema is a potential adverse effect of estrogen therapy. Estrogen and aldosterone both originate from cholesterol, and an excessive amount of estrogen in the body can stimulate aldosterone receptors, leading to water retention in nephrons. This water retention can result in edema. It is a non-adverse effect.
(b): Estrogen promotes ductal growth and fat deposition in the breasts. Excessive estrogen levels can lead to mammary duct hyperplasia, which may result in breast pain. It is not a non-adverse effect.
(c): Ovarian cancer is not known to be an adverse effect of estrogenic treatment. It is a possible choice for a non-adverse effect.
(d): Edema is a possible adverse effect of estrogenic treatment, and swelling in body parts may cause the feeling of sickness. It is not a non-adverse effect.
(e): Headache is a possible adverse effect of estrogenic treatment. It is not a non-adverse effect.
Therefore, the answer is (c).

Question Keywords: cardiac valvular prosthesis, biological, mechanical, implanted, patient, aspects, characteristics, patient, prosthesis, INCORRECT, statement
Candidate Answers: (a) Permanent anticoagulation is necessary in mechanical prostheses. (b) In general, biological prostheses are indicated in young patients, with long life expectancy. (c) Biological prostheses would be indicated in cases that present a formal contraindication for anticoagulation. (d) The rate of structural deterioration of a biological prosthesis is inversely proportional to the age of the subject. (e) Biological prostheses do not require permanent anticoagulation.

Context: Cardiac valvular prostheses (biological or mechanical) are artificial cardiac valves implanted into a patient's heart. Mechanical valves may last a lifetime, but they come with an increased risk of blood clots, necessitating the use of blood thinners such as warfarin. In contrast, biological valves, which are made from pig or cow tissue, do not increase the risk of bleeding or clotting but tend to wear out sooner.
(a): Mechanical valves increase the risk of blood clotting. It is not an incorrect statement.
(b): The latest revisions of the ESC/EACTS guidelines suggest that bioprostheses are acceptable in patients aged between 60 and 65 years at the time of surgery. The reoperation rate for structural valve degeneration (SVD) of bioprostheses occurred exclusively among patients younger than 56 years. Young patients are not typically recommended for a biological prosthesis.
(c): Biological prostheses, which are made from pig or cow tissue, do not increase the risk of either bleeding or clotting but will wear out sooner. It is not an incorrect statement.
(d): The disadvantages of biological heart valves are a smaller valve orifice area and the risk of structural valve degeneration, which may necessitate reoperation. Thus, the younger the patient, the higher risk of structural deterioration. It is not an incorrect statement.
(e): Biological prosthesis do not increase the risk of clotting so do not require permanent anticoagulant. It is not an incorrect statement.
Therefore, the answer is (b).

Question Keywords: connection, automatic, emotional responses, control, behaviors, guiding, behavior, manifestation, emotional responses
Candidate Answers: (a) The angular gyrus of the limbic system. (b) The convolution or lobe of the insula. (c) The prefrontal orbitofrontal or ventromedial cortex. (d) The thalamus (e) The cortex of somatosensory association.

Context: The prefrontal orbitofrontal cortex has multiple functions including mediating context specific responding, encoding contingencies in a flexible manner, encoding value, encoding inferred value, inhibiting responses, learning changes in contingency, emotional appraisal, altering behavior through somatic markers, driving social behavior, and representing state spaces. The orbitofrontal cortex thus plays a key role in emotion, by representing the reward value of the goals for action.
(a): The angular gyrus (AG) is a hub of several networks that are involved in various functions, including attention, self-processing, semantic information processing, emotion regulation, and mentalizing. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
(b): The insula is important for gustatory and sensorimotor processing, risk-reward behavior, autonomics, pain pathways, and auditory and vestibular functioning. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.

Continued on next page

Table 13 – *Continued from previous page*

(c): The prefrontal cortex guides behavior by controlling the manifestation of emotional responses through understanding rewards, encoding values, and driving behaviors. It is the potential correct answer.
(d): The thalamus acts as the body's information relay station. All sensory information (except for olfaction) must be processed through the thalamus before being sent to the cerebral cortex for interpretation. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
(e): The somatosensory cortex is responsible for processing all bodily sensations. These sensations originate from receptors located throughout the body that detect temperature, pain, touch, pressure, and proprioception. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
Therefore, the answer is (c).

Table 14: Prompts for MedMCQA

Question Keywords: Maximum, increase, prolactin level
Candidate Answers: (a) Risperidone (b) Clozapine (c) Olanzapine (d) Aripiprazole

Context: The four drugs in answer choices are all atypical antipsychotics, which are used to treat psychotic conditions like schizophrenia through blockage of dopamine and serotonin receptors. These drugs block dopamine D2 receptors and serotonin 5-HT2 receptors. Maximum increase in prolactin, or hyperprolactinemia, is one of the side effects of atypical antipsychotics, because dopamine tends to inhibit prolactin release from the anterior pituitary. (a): Risperidone is a type of atypical antipsychotics that block dopamine D2 receptor and serotonin 5-HT2 receptor. It is generally used to treat schizophrenia or disorders with concomitant psychosis. Hyperprolactinemia is one of the most common side effects of risperidone. It is the drug to increase prolactin levels.
(b): Clozapine is used to treat schizophrenia or disorders with concomitant psychosis. Clozapine is associated with side effects such as agranulocytosis, seizures, and myocarditis, but it does not appear to elevate prolactin levels. It is not the drug to increase prolactin levels.
(c): Olanzapine is used to treat schizophrenia or disorders with concomitant psychosis. The side effect of olanzapine does not include hyperprolactinemia. It is not the drug to increase prolactin levels.
(d): Aripiprazole is generally used to treat schizophrenia or disorders with concomitant psychosis. The side effect of olanzapine does not include hyperprolactinemia. It is not the drug to increase prolactin levels.
Therefore, the answer is (a).

Question Keywords: male, complains, severe back pain, inability, left lower limb, Radiographic studies, compression, nerve elements, intervertebral, foramen, vertebrae L5, S1, structure, space-occupying lesion
Candidate Answers: (a) Anulus fibrosus (b) Nucleus pulposus (c) Posterior longitudinal ligament (d) Anterior longitudinal ligament

Context: The male is complained of a severe back pain and inability to move, and radiographic evidence shows the compression of a nerve component. This may suggest a herniated intervertebral disk through a tear in the surrounding annulus fibrosus. The soft, gelatinous nucleus pulposus is forced out through a weakened part of the disk, compressing nerve components of the spinal cord and resulting in back pain and nerve root irritation. This impingement is resulting in paralysis, and should be considered a medical emergency.
(a): Annulus fibrosus is a tough, circular exterior of the intervertebral disc, made up of fibrous connective tissue. It surrounds the soft inner core, the nucleus pulposus. It is not the component that is forced out by the tear.
(b): Nucleus pulposus is the inner core of the vertebral disc. The tear in the annulus fibrosus causes it to be forced out. It could result in compression of the nerve components of the vertebrae.

Continued on next page

Table 14 – *Continued from previous page*

(c): Posterior longitudinal ligament connects and stabilizes the bones of the spinal column. It runs almost the entire length of the spine, from the 2nd vertebra in the cervical spine (neck) all the way down to the sacrum (end of the spine). This ligament is located adjacent to the spinal cord. It is not easily teared or curved.
(d): Anterior longitudinal ligament is a ligament that runs down the anterior surface of the spine. It traverses all of the vertebral bodies and intervertebral discs on their ventral side. It has a high tensile strength and is resistant to tearing or deformation. It is not easily teared or curved.
Therefore, the answer is (b).

Question Keywords: Neuroendocrine cells, lungs
Candidate Answers: (a) Dendritic cells (b) Type I pneumocytes (c) Type II pneumocytes (d) APUD cells

Context: Neuroendocrine cells are part of the neuroendocrine system. The neuroendocrine cells of the lung make hormones that control the flow of air and blood in the lungs.This may suggest a herniated intervertebral disk through a tear in the surrounding annulus fibrosus. The soft, gelatinous nucleus pulposus is forced out through a weakened part of the disk, compressing nerve components of the spinal cord and resulting in back pain and nerve root irritation. This impingement is resulting in paralysis, and should be considered a medical emergency.
(a): Dendritic cells are a type of antigen-presenting cell in the immune system that act as messengers between the innate and adaptive immune systems. It is not a type of neuroendocrine cell.
(b): Type I pneumocytes are alveolar cells that line the alveolar surface of the lungs and are responsible for gas exchange. It is not a type of neuroendocrine cell.
(c): Type II pneumocytes are alveolar cells that secrete surfactant to reduce alveolar surface tension and prevent alveolar collapse. It is not a type of neuroendocrine cell.
(d): APUD cells are a type of neuroendocrine cell that function through amine precursor uptake and decarboxylation. It is accurate to say that they are a type of neuroendocrine cell.
Therefore, the answer is (d).

Question Keywords: Presence, remote, contamination,water
Candidate Answers: (d) Streptococci (b) Staphalococci (c) Clastridium pertringes (d) Vibrio

Context: Infections that can be spread through water contamination are generally transmitted orally or via fecal matter. (a): Streptococci are spread through direct contact with the nose and throat discharges of an infected individual or with infected skin lesions. Water is not a medium for the spread of streptococci. It is not related water contamination.
(b): Staphylococci is spread by skin contact, like a bite or cut. It is not related water contamination.
(c): Clostridium perfringens are one of the most common causes of food poisoning. They are environmentally stable and specific to contamination by sewage. Their spread is a indicator of water contamination.
(d): Vibrio species are gram-negative bacteria that spread through foodborne infection, but they are highly salt tolerant and unable to survive in fresh water. It is not related water contamination.
Therefore, the answer is (c).

Question Keywords: True, Mooren's ulcer, 2007, 2013
Candidate Answers: (a) Painless condition (b) Affects cornea (c) Sudden loss of vision (d) Bilateral in majority of cases

Context: Mooren's ulcer is characterized by painful peripheral corneal ulceration of unknown etiology. The disease generally begins with intense limbal inflammation and swelling in the episclera and conjunctiva. Patients often experience severe pain, photophobia, and tearing along with a red inflamed eye.
(a): Mooren's ulcer is a painful ulceration of the eye. It is not the truth of Mooren's ulcer.
(b): Mooren's ulcer is characterized by painful peripheral corneal ulceration of unknown etiology. It is the truth of Mooren's ulcer.
(c): The symptoms of Mooren's ulcer do not include sudden loss of vision. It is not the truth of Mooren's ulcer.
(d): About one third of Mooren's ulcer cases present bilaterally. The proportion is less than half. It is not the majority of cases.

Continued on next page

Table 14 – *Continued from previous page*

Therefore, the answer is (b).

Prompts for general domain datasets. Our prompts on *CommonsenseQA* and *OpenbookQA* are based on (Li et al., 2022).

Table 15: Prompts for Commonsense QA

Question Keywords: fountain pen, people, ink, absorb, pen, hand done, extra, use, fountain
Candidate Answers: (a)shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

Context: Fountain pens need to be filled with ink for writing. Extra ink should be absorbed using special tools.
(a): A fountain pen can be conveniently carried in a shirt pocket. It is not associated with the tool to absorb extra ink from fountain pens.
(b): Calligraphers use fountain pens to create stunning handwriting. It is not associated with the tool to absorb extra ink from fountain pens.
(c): An inkwell serves as a container for the ink used in a fountain pen. It is not associated with the tool to absorb extra ink from fountain pens.
(d): A fountain pen can be kept safely in a desk drawer. It is not associated with the tool to absorb extra ink from fountain pens.
(e): Blotters are designed to absorb excess ink from pens. It is the tool for absorbing extra ink.
Therefore, the answer is (e).

Question Keywords: fox, forest, walk, look, city
Candidate Answers: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook (e) dense forest

Context: Foxes are animals that typically live in forests. They walk from the city to the forest to look for their living place.
(a): Pretty flowers are in forests. It is not a reason for a fox walking into the forest.
(b): Foxes sometimes prey on chickens in hen houses. It is not a reason for a fox to walk into the forest.
(c): Forests are the natural habitat of foxes. Foxes walk from city to forest to look for their natural habitat. (d): Forests and foxes are common subjects in storybooks. It is not a reason for fox walking to the forest.
(e): Dense forest is a type or category of forests characterized by having a high density of trees and vegetation. It is a type of forest.
Therefore, the answer is (c) or (e).

Question Keywords: grape, put, check
Candidate Answers: (a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market

Context: Grapes need to be put into a place for checking out.
(a): Grapes can be eaten by mouth. It is not a place to put grapes for checking out.
(b): Grapes can be brought during grocery shopping and people put groceries into grocery carts before checking out. It could be a potential place to put grape.
(c): Super markets sell grapes. It is not a place to put grapes for checking out.
(d): Fruit markets sell grapes. It is not a place to put grapes for checking out.
(e): Fruit baskets are often used as gifts to hold and present a variety of fresh grapes. It is not a place to put grapes for checking out.
Therefore, the answer is (b).

Question Keywords: drawstring bag, head, woman, bag, drawstring, check, baggage
Candidate Answers: (a) garbage can (b) military (c) jewelry store (d) safe (e) airport

Context: A woman can check baggage such as a drawstring bag at the check-in counter.
(a): A garbage can is a container that is specifically designed to hold and contain trash or waste materials. It is not related to the context.

Continued on next page

Table 15 – *Continued from previous page*

(b): Military refers to the armed forces of a country, which is responsible for defending the nation and its interests against external threats. It is not a place where a woman can check bags.
(c): Jewelry stores sell jewelry. It is not a typical place to check baggage.
(d): Check baggage could keep the bag safe. A woman can check her drawstring bag to keep the bag safe.
(e): Airport is a place where the woman can check her drawstring bag as baggage at the check-in-counter. It is common to check baggage in airport.
Therefore, the answer is (e).

Question Keywords: cable, entertainment, home, require, equipment
Candidate Answers: (a) radio shack (b) substation (c) television (d) cabinet (e) desk
Context: A cable transmits electricity or information and data to home entertainment equipment that requires electricity.
(a): Radio Shack is a retailer that sells cable. It is not a home entertainment equipment used cable.
(b): Cables are used to transmit electrical energy between substations and other parts of the electrical power system. It is not a home entertainment equipment used cable.
(c): Television is a type of home electric entertainment equipment that requires cable. It is a home entertainment equipment used cable.
(d): Cabinet is a place to store cable. It is not a home entertainment equipment used cable.
(e): Desk with built-in cable management features can help keep cables tidy. It is not a home entertainment equipment used cable.
Therefore, the answer is (c).

Question Keywords: people, populate, might, may, sammy, go
Candidate Answers: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

Context: People may like to go to places where people populate together. (a): Populated areas are locations where people gather and live in close proximity to each other. It could be a place where people populate together. (b): Deserts are inhospitable environments for people. It is not a place where people populate together. (c): People go to race competitions on the race track. It could be a place where people populate together. (d): Apartments serve as living spaces for people. It is not a place where people populate together. (e): Roadblocks are structures set up to restrict or regulate the movement of people and vehicles. It is not a place where people populate together. Therefore, the answer is (a) or (c).

Question Keywords: highway, maps, replace, street, google, map, highway, gps, service
Candidate Answers: (a) united states (b) mexico (c) countryside (d) atlas (e) oceans

Context: Google Maps and GPS services have replaced traditional physical maps for navigating highways and streets.
(a): People in the United States use Google Maps and GPS services to navigate highways and streets. It is not the tool that GPS replaced with.
(b): People in Mexico use Google Maps and GPS services to navigate highways and streets. It is not the tool that GPS replaced with.
(c): Google Maps and GPS services cover the countryside. It is not the tool that GPS replaced with.
(d): Google Maps and GPS services have replaced traditional physical maps for navigating highways and streets. Atlases are examples of traditional physical maps. It is the tool that GPS replaced with.
(e): Google Maps and GPS services cover the oceans and are commonly used in marine navigation. It is not the tool that GPS replaced with.
Therefore, the answer is (d).

Table 16: Prompts for Openbook QA

Question Keywords: acid, environment, aquatic, rain, effect, acid rain
Candidate Answers: (a) decrease in plant life (b) increase in fish population (c) increase in plant growth (d) cleaner and clearer water

Continued on next page

Table 16 – *Continued from previous page*

Context: The acid rain is a type of rain that has an acidic effect due to the presence of acid in the atmosphere. Acid rain is harmful to the environment, especially aquatic life.The acid in the rain can have a negative effect on the water quality of aquatic environments.
(a): Acid rain can have a negative effect on plant life. The acid in the rain can damage plant cells and cause a decrease in plant growth, leading to a decrease in plant life. It is likely to have a decrease in plant life by acid rain.
(b): Acid rain can have a harmful effect on aquatic life, including fish. The acid in the water can make it difficult for fish to breathe and can harm their reproductive systems. It is not likely to have an increase in fish population by acid rain.
(c): As previously mentioned, the acid in the rain can damage plant cells and cause a decrease in plant growth. It is not possible to have an increase in plant growth.
(d): Acid rain can have a harmful effect on water quality, making it more acidic and harmful to aquatic life. It is not possible to have cleaner and clearer water.
Therefore, the answer is (a).

Question Keywords: moon, surface
Candidate Answers: (a) is smooth on the entire surface (b) contains large cavities cause by explosions (c) contains an internal core of cheese (d) is filled with lakes

Context: The moon is a natural satellite that orbits around the Earth. Its surface is covered with dead volcanoes, impact craters, and lava flows, some visible to the unaided stargazer.
(a): The moon has mountains, craters, and other features caused by impacts from meteoroids and asteroids. It is not entirely smooth on the surface.
(b): Impact craters are formed when an asteroid craters, each of which was formed when an asteroid or comet collided with the Moon's surface. The moon's surface contains large cavities caused by explosions from impacts.
(c): The core is largely composed of iron and some nickel. The inner core is a solid mass about 480 km in diameter. It does not contain an internal core of cheese. (d): The moon has lunar maria composed of basalt formed from surface lava flows that later congealed. It is not filled with lakes.
Therefore, the answer is (b).

Question Keywords: car, approach, night
Candidate Answers: (a) the headlights become more intense (b) the headlights recede into the dark (c) the headlights remain at a constant (d) the headlights turn off

Context: Headlights of a car are a source of light. As a car approaches, the source of light becomes closer, and that source will appear brighter.
(a): As the car becomes closer, the distance to the source of light decreases. The headlights become brighter and more intense. This is a possible phenomenon.
(b): If the source does not change and the headlights are closer, the headlights cannot become dimmer. This is not a commonsense relation.
(c): If the distance to the source of light changes, the brightness of headlights will change. It is not able to remain constant.
(d): Turning off the headlights would cause the driver to be driving in complete darkness, which is dangerous and can lead to accidents. It is not a reasonable condition.
Therefore, the answer is (a).

Question Keywords: change, easter, weather change, weather, christmas
Candidate Answers: (a) the air may chill (b) the ground may freeze (c) the plants may die (d) the ground may warm
Context: In the US, Christmas falls in the winter season, while Easter arrives at the beginning of spring.
(a): The air becomes chill as temperature drops. The temperature commonly increases from winter to spring. It is not a likely scenario.
(b): During winter, the ground usually freezes, whereas in spring, it does not. It is not a probable scenario.
(c): Extreme cold or hot weather can cause plants to die. The beginning of spring provides suitable weather conditions for plants to grow. It is not common to have plants die.

Continued on next page

Table 16 – *Continued from previous page*

(d): As winter transitions into spring, the weather becomes warmer. The temperature of the ground is influenced by the weather.
Therefore, the answer is (d).

Question Keywords: heat, recipe, moisture, good, ocean
Candidate Answers: (a) a violent storm (b) violent sea animals (c) condensation (d) inland storms

Context: The ocean, a vast body of water that covers a large portion of the Earth's surface, serves as a source of heat and moisture.
(a): The heat and moisture present in the ocean can create ideal conditions for a hurricane or typhoon. Hurricane and typhoon are violent storms.
(b): Violent sea animals are not related to heat and moisture in the ocean. It is not a likely choice.
(c): Condensation is the process by which water vapor becomes liquid, which is the reverse of evaporation. This can happen in one of two ways: either the air is cooled to its dew point or it becomes so saturated with water vapor that it cannot hold any more water. It is not likely to occur in hot conditions.
(d): Although heat and moisture can cause inland storms, they are not directly related to the ocean. It is not a likely choice.
Therefore, the answer is (a).

Question Keywords: hummingbird, take
Candidate Answers: (a) bees (b) energy (c) pollen (d) honey

Context: Hummingbirds dip their long bills into flowers to drink nectar to get energy.
(a): Hummingbirds and bees are both attracted to the sweet nectar produced by flowers, but bees extract the nectar from the base of the flowers, while hummingbirds dip their long bills into the flowers to drink the nectar and obtain energy. No relationship can be found.
(b): Hummingbirds obtain energy by getting nectar from flowers through dipping their long bills into the flowers. No relationship can be found.
(c): When hummingbirds drink nectar, they also inadvertently take grains of pollen which stick to their feathers and bills, and get carried to the next flower they visit. No relationship can be found.
(d): Hummingbirds do not produce or consume honey. This fact is unrelated to their method of obtaining energy by drinking nectar from flowers.
Therefore, the answer is None.

Question Keywords: responsible, sun
Candidate Answers: (a) puppies learning new tricks (b) children growing up and getting old (c) flowers wilting in a vase (d) plants sprouting, blooming and wilting

Context: The sun is the source of energy for physical cycles on Earth.
(a): Puppies learning new tricks involves the acquisition and processing of information, which is essential for the puppies to learn and adapt to their environment. It is not directly related to the effect of the sun.
(b): Children grow up and age over time. The sun is not directly responsible for the passage of time itself. It is not directly related to the effect of the sun.
(c): Flowers in a vase become wilting because they are cut from their original source of nutrients and water and are no longer able to receive the essential nourishment they need to stay healthy and vibrant. It is not directly related to the effect of the sun.
(d): Plants need sunlight to photosynthesize and grow, and the sun's heat and light play a crucial role in the process of plant growth and decay. It is the thing that the sun is responsible for.
Therefore, the answer is (d).