Induction-Deduction Prompting: Enhancing Hidden-Information Reasoning in Medical LLM QA

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in medical question answering. However, limited attention has been given to the underlying reasoning patterns in LLM generated chain of thoughts. We analyse common medical reasoning scenarios using a Bayesian Network, revealing the prevalence of hidden information, especially in the MedQA dataset. We introduce two simple prompts, induction (inferring hidden information) and deduction (evaluating options based on observed and inferred information). Used together they outperform conventional prompting techniques as well as Med-Palm 2, which relies on complex, expert-crafted prompting and expensive fine-tuning.

1 Introduction

004

007

012

017

019

027

033

The advancement of Large Language Models (LLMs) has spurred increased exploration of their capabilities in medical domain question answering. Leading LLMs, such as GPT-4 and Med-Palm 2 (Nori et al., 2023a; Singhal et al., 2023), have showcased proficiency with high scores in US Medical Licensing Examination (USMLE) style questions on the MedQA dataset (Jin et al., 2021).

In this paper, we seek to further improve medical QA performance by analysing the reasoning patterns in medical QA and developing appropriate prompting strategies based on the analysis. Consider the Bayesian network for a medical case shown in Figure 1. A common type of question would describe the patient history, some symptoms, the results for some tests and ask for an appropriate treatment. In such cases, one way to answer the question would be to infer the disease, which is hidden, before deciding on the treatment using knowledge of both the disease as well as the other information that is provided.

Error analysis on MedMCQA, a commonly used benchmark, reveals that models often mispredict



Figure 1: Bayes network of a medical case showing the hidden variable *disease*.

041

042

043

044

045

047

051

053

054

058

059

060

061

062

063

064

065

066

067

questions related to the right course of action (treatment) in a given medical situation (Pal et al., 2022). Questions of this nature, which may describe a patient's condition and then inquire about appropriate treatment or necessary examinations, commonly follow a pattern where underlying hidden information is necessary to link the provided information to the answer. We define these as hidden-information questions. The hidden-information questions constitute a significant portion of USMLE-style questions, such as those in MedQA (Jin et al., 2021); our estimate shows that over 60% of questions in USMLE-style MedQA have hidden-information.

We call the process of prompting the LLM to infer the underlying hidden information *induction prompting*. A more careful examination of the inference procedure on a Bayes network would also suggest examining the conditional probability of each treatment given the disease and other information and selecting the most likely treatment. This suggests that to select the best option among multiple options, we should prompt the LLM to evaluate the likelihood of each option given the observed and inferred information before deciding the best option. We name this technique as *deduction prompting*, because assuming each option is true and evaluating its likelihood resembles human deduction reasoning. By combining both types

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

of prompting techniques, we obtain what we call *induction-deduction prompting*.

When hidden information is not important, we found that doing induction prompting only is slightly more effective. We develop a prompting method for classifying whether hidden information is required. Combining the classifier with induction and induction-deduction prompting, we obtain substantial improvement over the underlying LLM (GPT-4) that is used (Nori et al., 2023a).

2 Related Works

069

074

077

082

087

094

095

100

101

102

103

106

107

108

109

110

111

112

113

114

115

116

We focus on MedQA due to its close simulation of a real-world medical QA setting - it is openQA and requires the challenging hidden-information reasoning. GPT-4 with simple prompt (Nori et al., 2023a) achieved approximately 80% accuracy on MedQA without fine-tuning.

Med-PaLM 2(Singhal et al., 2023), state-of-theart until Nov 2023, attains an 86.5% accuracy through expensive, task-specific fine-tuning of the general PaLM model (Chowdhery et al., 2023). Additionally, the results on the medical benchmarks for Med-PaLM 2 relied on expert-crafted exemplars and an elaborated two-stage prompt scheme of 44 calls for answering each question.

Our simple prompting strategies with the publicly available general LLM GPT-4 outperform Med-Palm 2.

Medprompt (Nori et al., 2023b), independent contemporaneous work released on Arxiv in Nov 2023, obtains the new state-of-the-art performance. Although not fine-tuned, its preprocessing involves generating a chain-of-thought for all training questions for retrieval when answering test questions. We will briefly compare our prompting strategies with Medprompt in Section 4.6.

Our work is also closely related to multi-hop question answering, where multiple steps of reasoning are required to answer a question (Yang et al., 2018). Our induction prompting can enhance multi-reasoning QA by prompting LLMs to infer the underlying identity of the bridge entity (Trivedi et al., 2022), which is crucial in connected reasoning.

3 Induction-Deduction Prompting

Complex questions in MedQA contain underlying hidden information. For instance, in questions asking for treatment or additional tests given a patient's symptoms, the disease is hidden, as shown in Figure 1. If the hidden variable *disease* is marginalized out, the amount of dependencies among the variables increases substantially, as shown in Figure 2.



Figure 2: Bayesian network for medical QA where the hidden variable *disease* is marginalized out.

Deductive reasoning in the resulting Bayesian Network (BN) poses two significant challenges. Firstly, the large number of dependencies complicates the evaluation of joint likelihoods. Secondly, specifying local conditional probabilities between distantly associated nodes becomes challenging for LLMs. For instance, LLMs are mainly trained on text related to disease-symptom or diseasetreatment topics that are more common, leading to a scarcity of information on direct symptomtreatment relationships.

To address these challenges and enhance deductive reasoning in questions with hidden information, we first infer the hidden information through inductive reasoning and subsequently evaluate the likelihood of all options using deductive reasoning.

3.1 Prompting Strategies

Classifying Hidden-information Questions We first prompt the LLM to automatically classify hidden-information questions. As shown in Table 1, the LLM is prompted to infer underlying information. If the inferred entity exactly matches the question's target, it is classified as a fully-observed question; otherwise, it is considered hidden-information, as the additional entity inferred can act as a hidden node uncovered in the Bayesian network to facilitate reasoning. The classification prompt with examples is shown in the Appendix A.

Deduction and Induction Prompts Deductive prompting instructs LLMs to consider each option in multiple-choice questions, assess the joint likelihoods of observing the given information, and select the most probable answer. For inductive

Hidden-information Classification Prompt
Instructions:
Classify the following medical questions. Based on the given information,
infer any additional information such as underlying conditions or the identity
of unknown entities. If the inferred entity is in the question stem or options,
output type 1; otherwise, output type 2.
For each question, provide:
Inferred Entity:
Entity is an option: Y or N, explain
Type: 1 or 2
Fully-observed Example Question
Question: A 5-year-old girl is brought to the clinic by her mother for exces-
sive hair growth. [Details of the case] What is the most likely diagnosis?
(A) "Congenital adrenal hyperplasia"
(B) "Granulosa cell tumor"
(C) "Idiopathic precocious puberty"
(D) "McCune-Albright syndrome"
(E) "Sertoli-Leydig tumor"
Inferred Entity: Granulosa cell tumor
Entity is an option: Y (The inferred entity, "Granulosa cell tumor," is listed
as option B.)
Type: 1
Hidden-information Example Question
Question: A 68-year-old man presents to the emergency department with
leg pain. [Details of the case] Which of the following is the best next step
in management?
(A) "CT angiogram"
(B) "Graded exercise and aspirin"
(C) "Heparin drip"
(D) "Surgical thrombectomy"
(E) "Tissue plasminogen activator"
Inferred Entity: Acute limb ischemia
Entity is an option: N (The term "acute limb ischemia" is not mentioned
in the question stem or options, but the symptoms described are indicative
of this condition.)
Туре: 2

Table 1: Classification Prompt for Hidden-Information and Fully-Observed Questions with Examples

prompting, the LLM is prompted to infer any underlying information and select the option based 155 on both observed and inferred information, without being asked to assess each option separately. Deduction and induction prompts with their examples are shown in the Appendix A.

> Induction-deduction Combining strengths of both, induction-deduction prompting first infers underlying information and finally evaluates all options given the observed and inferred information to choose the most likely answer. A prompt with an example is shown in Table 2.

4 **Experiments**

4.1 Dataset

154

156

157

159

160

161

162

163

164

165

166

167

169

170

171

172

173

174

175

176

177

MedQA comprises English multiple-choice questions. The dataset includes a standard version with 5 options per question and a simplified 4-option version with the same 1273 questions, where one incorrect option is removed.

4.2 Models

We utilized GPT-4-Turbo (GPT-4-1106-preview) as the base model for our study.

As baseline we use few-shot chain-of-thought prompting. Few-shot prompting (Brown et al.,

Induction-Deduction Prompt				
Address the medical multiple-choice questions methodically. Infer any				
additional information and evaluate all options based on both the observed				
and inferred information. Choose the most probable answer.				
Question: A 48-year-old woman comes to the emergency department				
because of a photosensitive blistering rash on her hands, forearms, and face				
for 3 weeks. [details of the case] Laboratory studies show a normal serum				
ferritin concentration. Which of the following is the most appropriate next				
step in management to induce remission in this patient?				
(A) "Pursue liver transplantation"				
(B) "Begin oral thalidomide therapy"				
(C) "Begin phlebotomy therapy"				
(D) "Begin oral hydroxychloroquine therapy"				
(E) "Begin subcutaneous deferoxamine therapy"				
Entity inferred: Porphyria Cutanea Tarda (PCT)				
Explanation: The patient's presentation suggests PCT, the most common				
porphyria type.				
Evaluate all options:				
(A) Liver transplantation is not first-line for PCT.				
(B) Thalidomide is not a standard PCT treatment.				
(C) Phlebotomy reduces iron overload in PCT.				
(D) Hydroxychloroquine is used after phlebotomy.				
(E) Deferoxamine is not first-line for PCT.				
Answer: (C)				

Table 2: Induction-Deduction Prompt with Example and CoT Explanation

2020) involves prompting LLMs by adding example inputs and outputs before the actual question is asked. Chain-of-thought (CoT), introduced by Wei et al. (2022), extends this by augmenting each few-shot example in a prompt with a step-by-step explanation towards the final answer. For baseline, we used the same few-shot CoT prompts crafted by human experts as employed by Singhal et al. (2022).

For our Induction and Deduction prompts, we manually selected fixed examples and autogenerated the explanations and answers with GPT-4-Turbo.

In addition, we also apply Self-consistency (SC), introduced by Wang et al. (2022). SC enhances LLM reasoning performance by sampling multiple explanations and answers from the model. The final answer is determined by the majority (or plurality) vote. We applied SC with 3 samplings for all base prompts evaluated.

4.3 Results

	One-Hop	Multi-Hop
5-Option	63.0	37.0
4-Option	66.7	33.3

Table 3: Percentage of one- and multi-hop questions in MedQA standard (5-options) and simplified (4-options) versions.

Our classification prompt accurately classified hidden-information questions in MedQA, which make up over 60% of all questions. The accuracy is verified through manual inspection of random samples. This indicates that specific prompting

199

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

204

205

206

208

209

210

211

212

214

215

216

217

218

219

220

221

229

strategies targeting hidden-information reasoning will substantially impact the overall accuracy.

	No Hidden	With Hidden	Total
5-shot CoT	90.6	81.4	84.5
Deduction	93.4	81.9	85.7
Induction	92.0	83.0	86.0
Induction-Deduction	92.2	84.5	87.1
GPT-4 0-shot	-	-	78.9
GPT-4 5-shot	-	-	81.4
Med-Palm 2 ER	-	_	86.5

4.4 Induction-Deduction

Table 4: GPT-4-Turbo performance on MedQA 4-option questions with 5-shot Chain-of-Thought, Deduction, Induction, Induction-deduction prompting, all with Self Consistency with 3 samples. Baseline results are reported by Nori et al. (2023b) and Singhal et al. (2023), without differentiation of hidden-information questions.

	No Hidden	With Hidden	Total
5-shot CoT	87.5	78.8	82.0
Deduction	88.5	79.3	82.7
Induction	87.9	80.7	83.4
Induction-Deduction	89.4	82.8	85.2
GPT-4 0-shot	-	-	74.7
GPT-4 5-shot	-	-	78.6

Table 5: GPT-4-Turbo performance on MedQA 5-option questions with 5-shot Chain-of-Thought, Deduction, Induction, Induction-deduction prompting, all with Self Consistency with 3 samples. Baseline results are reported by Nori et al. (2023b), without differentiation of hidden-information questions.

Induction-deduction prompting improves the baseline GPT-4-Turbo with 5-shot CoT crafted by experts by around 3% in both standard and simplified versions. In the simplified version, it outperforms the previous state-of-the-art Med-Palm 2, which used expensive medical fine-tuning and expert-crafted prompts with tens of samples for self-consistency. Med-Palm 2 did not report performance on the standard version.

4.5 Ablation Studies

Induction and deduction prompting both improve upon the baseline, although not as significantly as the induction-deduction prompting. Deduction prompting performs better in fully observed scenarios, while induction is more effective in hiddeninformation questions. This observation aligns with the nature of hidden-information questions, where underlying entities can be uncovered through induction, making it a more effective strategy. Conversely, in fully observed questions, a thorough evaluation of all options tends to lead to the correct answer. These results further affirm the effectiveness of our classification prompting strategy.

4.6 Comparison with Medprompt

In independent contemporaneous work released on Arxiv in Nov 2023, Medprompt (Nori et al., 2023b) obtains a state-of-the-art accuracy score of 90.2 on the MedQA 4-option questions. In Medprompt, the chain of thought reasoning examples for few-shot learning were generated by GPT-4 instead of medical specialists. With a 5-shot selfgenerated CoT prompt, Medprompt achieves an 86.9% accuracy on the MedQA 4-option test set which is substantially higher than that achieved using few-shot examples created by medical specialists¹. This performance is only marginally worse than our induction-deduction prompt and surpasses the performance of induction or deduction alone.

231

232

233

234

235

236

237

238

239

240

241

242

243

245

246

247

248

249

250

251

252

253

254

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

The reasons for the good performance of GPT-4's self-generated examples was not discussed in detail in (Nori et al., 2023b). Interestingly, examining the examples reveals that GPT-4's autonomously generated CoT consistently employs both induction and deduction strategies, involving the inference of diagnoses and the evaluation of all options to identify the best answer. Our work can hence also be viewed as providing an explanation for the strong performance of GPT-4's autonomously generated CoT examples.

5 Conclusion

Our analysis of USMLE-style MedQA dataset revealed the prevalence of complex questions requiring hidden-information reasoning. We devised a simple prompt for Language Models (LLMs) to identify and classify whether hiddeninformation is required in a question. To address these complex hidden-information questions, we introduced a new type of prompt for LLMs that follows an inductive-deductive reasoning pattern. The induction-deduction prompt on GPT-4-Turbo demonstrated substantial improvement over the basic prompt on MedQA performance. The performance surpassed the previous state-of-the-art Med-Palm 2, which relies on more complex fine-tuning. Extension of the work to handle questions that requires more than a single inductive reasoning step would be interesting.

¹Not only did GPT-4 outperform medical specialists, we have been scooped by GPT-4 in coming up with a new technique as well! Not really, ... the work is independent and contemporaneous. Furthermore, we explain the underlying mechanism for our example prompts whereas GPT-4 just generates them.

274 Limitations

In this work, we focused on hidden-information 275 questions, a challenging question type prevalent 276 in the USMLE-style medical QA dataset MedQA. 277 However, we acknowledge that different question 278 types within medical datasets may benefit from distinct prompting strategies. As demonstrated in the results in Section 4.3, induction and deduction 281 prompts show varying improvements in hiddeninformation and fully-observed questions. This underscores the importance of automated question 284 classification, performed here with GPT-4-Turbo, and the automatic selection of optimal prompts. A similar approach is concurrently explored in Medprompt (Nori et al., 2023b) and previously in Auto-CoT (Zhang et al., 2022), where they retrieve 289 the most similar Chain-of-Thought examples for prompting. Future work may involve classifying additional question types and developing tailored prompts accordingly.

> In future work, we aim to assess the effectiveness of induction-deduction prompting across diverse datasets within and beyond medical domain, potentially uncovering its utility in multi-hop QA datasets.

References

294

296

299

302

307

310

311

312

313

314

316

317

318

319

322

323

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al.

2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*. 324

325

326

327

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

354

355

356

357

358

359

360

361

362

363

364

365

366

- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health*, *Inference, and Learning*, pages 248–260. PMLR.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

A Appendix

Hidden-information Classification Prompt Instructions: Classify medical questions and provide additional information. For each question, output: Inferred Entity, Entity is an option ('Y' or 'N'), Type (One or Two) **Baseline CoT Prompt** Question: A 5-year-old girl with excessive hair growth, Tanner stage 4 development, ovarian mass, elevated estrogen. Most likely diagnosis? (A) "Congenital adrenal hyperplasia" (B) "Granulosa cell tumor" - Y (Inferred entity listed as option B) (C) "Idiopathic precocious puberty" (D) "McCune-Albright syndrome' (E) "Sertoli-Leydig tumor" Inferred Entity: Granulosa cell tumor scalene" Entity is an option: Y (B) "Latissimus dorsi"(C) "Pectoralis minor" Type: One (D) "Quadratus lumborum" Question: A 30-year-old man with a history of bilateral cryptorchidism treated with orchidopexy. Increased risk for which of the following? (A) "Teratocarcinoma" from the outer surfaces of the 3rd to 5th ribs. (B) "Sertoli cell tumor" Answer: (C) (C) "Leydig cell tumor' (D) "Yolk sac tumor" (E) "Testicular lymphoma" Inferred Entity: Testicular cancer Entity is an option: N Type: Two Question: A 16-year-old boy with behavior changes, expelled from school, torsion" history of excellent grades. Change related to his father moving out. Defense (B) "Left-on-right sacral torsion" mechanism describing the change in behavior? (C) "Right unilateral sacral flexion" (A) "Suppression" (D) "Right-on-right sacral torsion" (B) "Acting out" - Y (Question is asking to identify the defense mechanism) (C) "Projection" (D) "Passive aggression" (E) "Regression" options have a deep sulcus on the right. Inferred Entity: Defense mechanism explaining the change in behavior Answer: (D) Entity is an option: Y Type: One Question: A 68-year-old man with sudden leg pain, cold and pale left leg, diminished sensation, muscle strength 1/5, history of diabetes, hypertension, (A) "Allergic rhinitis" obesity, atrial fibrillation. Best next step in management? (B) "Epstein-Barr virus" (A) "CT angiogram" (C) "Mycoplasma pneumonia" (B) "Graded exercise and aspirin"(C) "Heparin drip" (D) "Rhinovirus" (D) "Surgical thrombectomy" (E) "Tissue plasminogen activator" likely cause is Rhinovirus. Inferred Entity: Acute limb ischemia Answer: (D) Entity is an option: N Type: Two Ouestion: A 38-year-old woman undergoes hemithyroidectomy for papillary thyroid carcinoma. Structure adjacent to superior thyroid artery dam-"Dopamine" aged. Likely symptom? (B) "Glutamate" (C) "Norepinephrine" (A) "Voice pitch limitation" (B) "Ineffective cough" (D) "Serotonin" (C) "Weakness of shoulder shrug" (D) "Difficulty swallowing' (E) "Shortness of breath" Serotonin can help increase positive emotions. Inferred Entity: External branch of the superior laryngeal nerve Answer: (D) Entity is an option: N Type: Two Table 6: Classifier prompt with examples

Address the medical multiple-choice questions methodically. Infer any additional information and evaluate all options based on both the observed and inferred information. Choose the most probable answer.

Ouestion: A 22-year-old male marathon runner presents to the office with the complaint of right-sided rib pain when he runs long distances. [details of the case] Which of the following muscles or muscle groups will be most useful in correcting this dysfunction utilizing a direct method? (A) "Anterior

Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. Among the options, only pectoralis minor muscle originates

Question: A 36-year-old male presents to the office with a 3-week history of low back pain. He denies any recent trauma but says that he climbs in and out of his truck numerous times a day for his job. Examination of the patient in the prone position reveals a deep sacral sulcus on the left, a posterior inferior lateral angle on the right, and a lumbosacral junction that springs freely on compression. The most likely diagnosis is (A) "Left-on-left sacral

Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. The deep sulcus on the left, a posterior ILA on the right, with a negative spring test suggests a right-on-right sacral torsion. All other

Question: A 44-year-old man comes to the office because of a 3-day history of sore throat, nonproductive cough, runny nose, and frontal headache. Which of the following is the most likely cause of this patient's symptoms?

Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. The symptoms, especially the headache, suggest that the most

Question: A previously healthy 32-year-old woman comes to the physician 8 months after her husband was killed in a car crash. Pharmacotherapy should be targeted to which of the following neurotransmitters? (A)

Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. The patient feels sad and among the options, only Dopamine and

Question: A 42-year-old man comes to the office for preoperative evaluation prior to undergoing adrenalectomy scheduled in 2 weeks. Initial preoperative preparation should include treatment with which of the following? (A) "Labetalol"

(B) "A loading dose of potassium chloride"
(C) "Nifedipine"

(D) "Phenoxybenzamine"

Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. The symptoms and the adrenal mass suggested pheochromocytoma, and the blood pressure indicates hypertension. Phenoxybenzamine is used to treat hypertension caused by pheochromocytoma. Answer: (D)

Table 7: CoT prompt with examples and explanations.

Deduction Prompt

Instructions: Evaluate all options and output a single answer in the format: Answer: ([ABCDE])

Ouestion: A 2-month-old at well-child exam, born at 39 weeks gestation. exclusively breastfed, normal vitals, weight 5,200 g, length 57.5 cm. Likely developmental milestone met?

(A) "Absence of asymmetric tonic neck reflex" - Possible, as this reflex usually disappears by 2-3 months.

(B) "Monosyllabic babble" - Unlikely, as it typically begins around 6 months.

(C) "Reaches for objects" - Unlikely, as infants start around 3-4 months.
(D) "Smiles in response to face" - Very likely, as social smiles appear

between 6 weeks and 2 months. (E) "Stares at own hand" - Possible, as infants show interest in hands by 2-3 months.

Answer: (D)

Question: A 51-year-old woman with right flank pain, bloody urine, lower extremity swelling, weight gain, chronic hepatitis B, periorbital edema, distended abdomen, 2+ lower extremity edema. CT scan shows liver nodules, ascites, right renal vein filling defect. Urinalysis: 4+ protein, positive glucose, fatty casts. Most likely underlying cause of renal vein findings? (A) "Acquired factor VIII deficiency" - Unlikely, as it presents with bleeding.

(B) "Loss of antithrombin III" - Likely, as symptoms suggest nephrotic syndrome with loss of antithrombin III in urine.

(C) "Impaired estrogen degradation" - Not typically associated with renal findings.

(D) "Antiphospholipid antibodies" - No specific indication for antiphospholipid syndrome

(E) "Paraneoplastic erythropoietin production" - Not directly related to renal vein thrombosis.

Answer: (B)

Ouestion: A 56-year-old woman with mammography showing dense, spiculated mass, clustered microcalcifications. Former flight attendant, strict Mediterranean diet, mother of two, breastfed each infant for 18 months, breast augmentation, tubal ligation, no family history of cancer. Most significant risk factor for breast cancer?

(A) "Sedentarism" - Unlikely, as patient compensates with Mediterranean diet

(B) "Breastfeeding" - Protective against breast cancer.

(C) "Mediterranean diet" - Associated with lower breast cancer risk.

(D) "Breast implants" - Controversial, not definitive evidence of increased breast cancer risk.

(E) "Occupation" - Flight attendant associated with increased breast cancer risk

Answer: (E)

Question: Investigator studying biomolecular mechanisms in human cells. Radioactive isotope unable to cross organelles introduced, isolated from cell components. Which reaction is most likely present?

(A) "Glucose-6-phosphate to glucose" - Cytoplasmic reaction, not organelle.
(B) "Fatty acyl-CoA to acetyl-CoA" - Mitochondrial reaction, organelle.

(C) "Carbamoyl phosphate to citrulline" - Urea cycle, starts in mitochondria. (D) "Glucose-6-phosphate to 6-phosphogluconolactone" - Pentose phosphate pathway in cytoplasm.

(E) "Isocitrate to α-ketoglutarate" - Citric acid cycle in mitochondria. Answer: (D)

Question: A 38-year-old woman undergoes hemithyroidectomy for papillary thyroid carcinoma. During surgery, structure adjacent to superior thyroid artery damaged. Likely symptom?

(A) "Voice pitch limitation" - Possible if external branch of superior laryngeal nerve damaged.

(B) "Ineffective cough" - Not typically associated with structures near superior thyroid artery.

(C) "Weakness of shoulder shrug" - Associated with accessory nerve damage, not near superior thyroid artery.

(D) "Difficulty swallowing" - Possible if recurrent laryngeal nerve damaged, (E) "Shortness of breath" - Possible with bilateral recurrent laryngeal nerve

damage, less likely with superior thyroid artery.

Answer: (A)

Table 8: Deduction prompt with example and CoT explanation.

Induction Prompt

Instructions: Choose the most probable answer based on observed and inferred information. Output a single option as the final answer in the format: Answer: ([ABCDE])

Question: A 68-year-old man with leg pain, sudden onset, cold and pale left leg, diminished sensation, muscle strength 1/5, past medical history of diabetes, hypertension, obesity, atrial fibrillation. Best next step in management

(A) "CT angiogram" - Unlikely, as the patient's presentation is suggestive of acute limb ischemia.

(B) "Graded exercise and aspirin" - Not appropriate for acute limb ischemia. (C) "Heparin drip" - Correct, as immediate anticoagulation is indicated for suspected acute limb ischemia.

(D) "Surgical thrombectomy" - May be needed but not initial management. (E) "Tissue plasminogen activator" - Not the first-line treatment for acute limb ischemia.

Answer: (C)

Question: A 48-year-old woman with photosensitive blistering rash, dark ruine, history of Coats disease, currently on hormonal replacement therapy. Family history of similar skin lesions. Most appropriate next step in management?

 (A) "Pursue liver transplantation" - Not indicated for the current condition.
(B) "Begin oral thalidomide therapy" - Not the first-line treatment for the suspected condition.

(C) "Begin phlebotomy therapy" - Correct, as the patient's presentation suggests Porphyria Cutanea Tarda (PCT), and reducing iron overload is essential.

(D) "Begin oral hydroxychloroquine therapy" - Not the preferred treatment for PCT

(E) "Begin subcutaneous deferoxamine therapy" - Not the preferred treatment for PCT.

Answer: (C)

Question: A 31-year-old woman at 26 weeks gestation with abnormal glucose tolerance test (156 mg/dL 1 hour after 50 g oral glucose challenge). Most appropriate next step in management?

(A) "Repeat the 50 g oral glucose challenge" - Not the standard approach after an abnormal result.

(B) "Administer an oral, 3-hour 100 g glucose dose" - Correct, as it is the standard diagnostic test for gestational diabetes mellitus (GDM).

(C) "Advise the patient to follow an American Diabetic Association diet plan" - Not sufficient for the diagnosis of GDM.

(D) "Begin insulin treatment" - Premature without confirming the diagnosis. (E) "Order a fetal ultrasound examination" - Not indicated for the current situation.

Answer: (B)

Question: Investigator studying neuronal regeneration. Aniline stain visualizes only soma and dendrites, not axon. Cellular element responsible for this staining pattern?

(A) "Microtubule" - Not responsible for Nissl staining.

(B) "Nucleus" - Not responsible for the described staining pattern.

(C) "Lysosome" - Not responsible for Nissl staining.

(D) "Golgi apparatus" - Not responsible for Nissl staining.

(E) "Rough endoplasmic reticulum" - Correct, as Nissl staining targets the rough endoplasmic reticulum present in soma and dendrites but not the axon.

Answer: (E)

A 38-year-old woman undergoes hemithyroidectomy for papillary thyroid carcinoma. Structure adjacent to superior thyroid artery damaged. Likely symptom?

(A) "Voice pitch limitation" - Correct, as the structure damaged is the (B) "Ineffective cough" - Not associated with the damaged structure.

(C) "Weakness of shoulder shrug" - Not associated with the damaged structure

(D) "Difficulty swallowing" - More associated with recurrent laryngeal nerve damage.

(E) "Shortness of breath" - Less likely with superior thyroid artery damage. Answer: (A)

Table 9: Induction prompt with example and CoT explanation.

Induction-Deduction Prompt

Instructions: Address the medical multiple-choice questions methodically. Based on the given information, infer any additional information such as underlying conditions or the identity of unknown entities.

Evaluate all options based on both the observed and inferred information. Choose the most probable answer.

You must output a single option as the final answer in the format: Entity inferred:

Explanation: Evaluate all options: Answer: ([ABCDE])

Question: A 68-year-old man presents to the emergency department with leg pain. He states that the pain started suddenly while he was walking outside. The patient has a past medical history of diabetes, hypertension, obesity, and atrial fibrillation. Which of the following is the best next step in management?

(A) "CT angiogram"

(B) "Graded exercise and aspirin"

(C) "Heparin drip"

(D) "Surgical thrombectomy"(E) "Tissue plasminogen activator"

Entity inferred: Acute limb ischemia

Explanation: The patient's presentation with sudden onset of leg pain, a cold and pale extremity, diminished sensation, and markedly reduced muscle strength in the affected leg is highly suggestive of acute limb ischemia. This is a medical emergency that requires immediate intervention to restore blood flow and prevent tissue death. The patient's history of atrial fibrillation is a risk factor for the formation of blood clots, which can embolize and cause arterial occlusion.

Evaluate all options:

(A) "CT angiogram" - This diagnostic test can help visualize the blood vessels and identify the location and extent of the occlusion. However, given the acute presentation and the need for immediate treatment. this may not be the best initial step.

(B) "Graded exercise and aspirin" - This is more appropriate for chronic management of peripheral artery disease, not for acute limb ischemia.

(C) "Heparin drip" - Anticoagulation with heparin is an important part of the management of acute limb ischemia to prevent further clotting, but it may not be sufficient on its own to restore perfusion in the acute setting. (D) "Surgical thrombectomy" - This is a procedure to remove the clot surgically and is a potential treatment for acute limb ischemia, especially if the limb is immediately threatened.

(E) "Tissue plasminogen activator" - This is a thrombolytic agent used to dissolve clots. It may be considered in certain cases of acute limb ischemia, but the decision to use thrombolytics would typically be made after imaging studies have confirmed the diagnosis and location of the occlusion. Answer: (C)

Question: A 48-year-old woman comes to the emergency department because of a photosensitive blistering rash on her hands, forearms, and face for 3 weeks. The lesions are not ichy. She has also noticed that her urine has been dark brown in color recently. Twenty years ago, she was successfully treated for Coats disease of the retina via retinal sclerotherapy. She is currently on hormonal replacement therapy for perimenopausal symptoms. Her aunt and sister have a history of a similar skin lesions. Examination shows multiple fluid-filled blisters and oozing erosions on the forearms, dorsal side of both hands, and forehead. There is hyperpigmented scarring and patches of bald skin along the sides of the blisters. Laboratory studies show a normal serum ferritin concentration. Which of the following is the most appropriate next step in management?

(A) "Pursue liver transplantation"

(B) "Begin oral thalidomide therapy"

(C) "Begin phlebotomy therapy"(D) "Begin oral hydroxychloroquine therapy"

(E) "Begin subcutaneous deferoxamine therapy"

Entity inferred: Porphyria Cutanea Tarda (PCT)

Explanation: The patient's presentation suggests Porphyria Cutanea Tarda (PCT). The most appropriate next step in management is phlebotomy therapy to reduce iron overload, which can exacerbate the condition.

Evaluate all options: (A) Liver transplantation is not the first-line treatment for PCT and is reserved for end-stage liver disease. (B) Thalidomide is not a standard treatment for PCT. (C) Phlebotomy therapy is a treatment for PCT. (D) Hydroxychloroquine is used in the management of PCT, typically considered after phlebotomy. (E) Deferoxamine is not the first-line treatment for inducing remission in PCT. Answer: (C)

Inducton-Deduction Prompt Continued Question:A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Which of the following is the best treatment for this patient?

(A) "Ampicillin"

- (B) "Ceftriaxone'
- (C) "Ciprofloxacin"
- (D) "Doxycycline" (E) "Nitrofurantoin"

Entity inferred: Urinary tract infection (UTI) in pregnancy

Explanation: The patient's symptoms suggest a lower urinary tract infection (UTI). The best treatment is Nitrofurantoin, considered safe during the second trimester.

Evaluate all options: (A) Ampicillin is not the first-line treatment for UTIs in pregnant women. (B) Ceftriaxone is typically reserved for more complicated infections or inpatient treatment. (C) Ciprofloxacin is generally avoided in pregnancy. (D) Doxycycline is contraindicated in pregnancy. (E) Nitrofurantoin is commonly used to treat uncomplicated UTIs in pregnancy. Answer: (E)

Question: An investigator is studying biomolecular mechanisms in human cells. A radioactive isotope that is unable to cross into organelles is intro-duced into a sample of cells. Which of the following reactions is most likely to be present in this cell component?

(A) "Glucose-6-phosphate to glucose"

(B) "Fatty acyl-CoA to acetyl-CoA"

(C) "Carbamoyl phosphate to citrulline"

(D) "Glucose-6-phosphate to 6-phosphogluconolactone"

(E) "Isocitrate to α -ketoglutarate"

Entity inferred: Cytosol

Explanation: The radioactive isotope introduced into the cells cannot cross into organelles, indicating it is confined to the cytosol. The most likely reaction is the first step of the pentose phosphate pathway, occurring in the cytosol.

Evaluate all options: (A) Gluconeogenesis occurs in the cytosol. (B) Beta-oxidation occurs in the mitochondria. (C) Urea cycle involves both mitochondria and cytosol. (D) Pentose phosphate pathway starts in the cytosol. (E) Citric acid cycle occurs in the mitochondria. Answer: (D)

Question: A 38-year-old woman undergoes hemithyroidectomy for treatment of localized, well-differentiated papillary thyroid carcinoma. The lesion is removed with clear margins. However, during the surgery, a structure lying directly adjacent to the superior thyroid artery at the upper pole of the thyroid lobe is damaged. This patient is most likely to experience which of the following symptoms?

(A) "Voice pitch limitation"(B) "Ineffective cough"(C) "Weakness of shoulder shrug"

(D) "Difficulty swallowing"

(E) "Shortness of breath"

Entity inferred: External branch of the superior laryngeal nerve

Explanation: Damage to the external branch of the superior laryngeal nerve can lead to voice pitch limitation due to the inability to tense the vocal cords. Evaluate all options: (A) Consistent with damage to the external branch. (B) More likely related to recurrent laryngeal nerve damage. (C) Related to accessory nerve, not in close proximity. (D) Could be related to recurrent laryngeal nerve. (E) Typically not associated with damage to the external branch.

Answer: (A)

Table 10: Induction-deduction prompts with examples and CoT explanations.