

Appendix A. Instruction prompt for QA generation from text fragments

You are an expert medical curriculum developer. Given a medical or medically-related text passage, you must generate question-answer pairs. The questions MUST be

- self-contained: answerable by the information in the passage but understandable on their own without needing direct reference to the passage (ie. "what is this passage about?" is NOT a valid question).
- grounded: directly related to the information in the passage
- research-able: have enough information to be able to retrieve the relevant passages or information in the future to help answer the question.

Generate a set of question-answer types including yes/no answer, short phrases, and longer sentences.

Think step-by-step and reason through the content of the passage to hypothesize potential question-answer pairs and then respond. Think of at least 8 question-answer pairs but more, if possible.

Respond in this format:

```
<think> You reasoning here... </think>
<qa-pair> [question here...] | [answer here...] </qa-pair>
<qa-pair> [question here...] | [answer here...] </qa-pair>
```

Here is an example:

PASSAGE:

The goal of transvesical prostatectomy is to remove the hypertrophic transitional zone of the prostate to reduce urinary obstruction. It is not a surgical procedure for malignant lesions. Appropriate history (including functional

status inquiry) and physical examination are done prior to any investigation. Physical exam always includes digital rectal exam. Enlarged lateral lobes can be easily felt and the rectal mucosa should move freely over the prostate. Any nodularity or fixation of the rectal mucosa should alert you to the possibility of prostatic cancer and core needle biopsy should be done before proceeding with the admission.

RESPONSE:

<think>

Okay, so I need to generate question-answer pairs based on the given medical passage. The questions should be answerable from the passage and understandable on their own. Let me read through the passage carefully to understand the key points.

The passage is about transvesical prostatectomy. It mentions that the goal is to remove the hypertrophic transitional zone of the prostate to reduce urinary obstruction. It's not for malignant lesions. Before any investigation, they do a history and physical exam, including a digital rectal exam. Enlarged lateral lobes can be felt, and the rectal mucosa should move freely. If there's nodularity or fixation, it might indicate cancer, so a biopsy is needed before admission.

Alright, so the main topics here are the purpose of the procedure, what it's not used for, the pre-procedure assessments, the physical exam findings, and when to suspect cancer.

I should think of questions that cover these areas. Let's start with the purpose. A question like, "What is the primary goal of transvesical prostatectomy?" The answer would be about removing the hypertrophic zone to relieve obstruction.

843 Next, since it's not for malignant
844 lesions, a question could be, "Is
845 transvesical prostatectomy used to
846 treat malignant lesions of the
847 prostate?" The answer would be no.
848 Then, the pre-procedure steps. Maybe,
849 "What assessments are typically
850 performed before transvesical
851 prostatectomy?" The answer would
852 include history, physical exam, and
853 digital rectal exam.

854 For the physical exam findings, a question
855 could be, "What is a normal finding
856 during the digital rectal exam in a
857 patient undergoing transvesical
858 prostatectomy?" The answer would
859 mention enlarged lateral lobes and
860 freely moving rectal mucosa.

861 When to suspect cancer: "What signs during
862 a digital rectal exam may indicate the
863 need for a core needle biopsy?" The
864 answer would be nodularity or fixation
865 of the rectal mucosa.

866 I can also think about the implications of
867 the findings. For example, "What
868 should be done if nodularity or
869 fixation is detected during the
870 digital rectal exam?" The answer would
871 be to perform a biopsy before
872 proceeding.

873 Another question could be about the
874 importance of the digital rectal exam
875 in this context. "Why is a digital
876 rectal exam important before
877 transvesical prostatectomy?" The
878 answer would be to assess for signs of
879 cancer.

880 I should make sure each question is clear
881 and can be answered without directly
882 referencing the passage, just the
883 information in it. Let me see if I can
884 come up with more. Maybe about the
885 specific zones of the prostate
886 involved. "Which part of the prostate
887 is targeted in transvesical
888 prostatectomy?" The answer is the
889 transitional zone.

890 Also, the procedure's focus: "What is the
891 focus of transvesical prostatectomy in
892 treating urinary obstruction?" The
893 answer is removing the hypertrophic
894 transitional zone.

895 I think that's a good set. I have eight
896 pairs, but maybe I can find a couple
897 more. Perhaps about the consequences
898 of not doing the biopsy. "What is the
899 consequence of not performing a core
900 needle biopsy when suspicious findings
901 are present?" The answer would be
902 proceeding without knowing if cancer
903 is present, which could be risky.

904 Wait, the passage doesn't explicitly state
905 the consequences, so maybe that's
906 beyond the given info. I should stick
907 to what's in the passage.

908 Another angle: the patient's history.
909 "What is included in the patient's
910 history before transvesical
911 prostatectomy?" The answer would be
912 functional status inquiry.

913 I think that's a solid list. I'll make
914 sure each question is clear and each
915 answer is accurate based on the
916 passage.

917 </think>

918 <qa-pair> What is the primary goal of
919 transvesical prostatectomy? | The
920 primary goal is to remove the
921 hypertrophic transitional zone of the
922 prostate to reduce urinary
923 obstruction. </qa-pair>

924 <qa-pair> Is transvesical prostatectomy
925 used to treat malignant lesions of the
926 prostate? | No, it is not a surgical
927 procedure for malignant lesions.
928 </qa-pair>

929 <qa-pair> What assessments are typically
930 performed before transvesical
931 prostatectomy? | Appropriate history,
932 including functional status inquiry,
933 and physical examination are done
934 prior to any investigation. </qa-pair>

935 <qa-pair> What is a normal finding during
936 the digital rectal exam in a patient
937 undergoing transvesical prostatectomy?
938 | Enlarged lateral lobes can be easily
939 felt, and the rectal mucosa should
940 move freely over the prostate.
941 </qa-pair>

942 <qa-pair> What signs during a digital
943 rectal exam may indicate the need for

944	a core needle biopsy? Nodularity or	correctness. Your evaluation must	992
945	fixation of the rectal mucosa may	result in a binary decision: "Correct"	993
946	indicate the possibility of prostatic	or "Incorrect".	994
947	cancer. </qa-pair>		995
948	<qa-pair> What should be done if	### EVALUATION CRITERIA ###	996
949	nodularity or fixation is detected	You must adhere to the following strict	997
950	during the digital rectal exam? A	criteria:	998
951	core needle biopsy should be performed		999
952	before proceeding with the admission.	1. **Factual Consistency:** The "Generated	1000
953	</qa-pair>	Answer" must be factually consistent	1001
954	<qa-pair> Why is a digital rectal exam	with the "Ground Truth Answer". It	1002
955	important before transvesical	must not contain any information that	1003
956	prostatectomy? It is important to	contradicts the ground truth.	1004
957	assess for signs of prostatic cancer,	2. **Completeness:** The "Generated	1005
958	such as nodularity or fixation of the	Answer" must address all parts of the	1006
959	rectal mucosa. </qa-pair>	"Query". It is considered "Incorrect"	1007
960	<qa-pair> Which part of the prostate is	if it omits critical information that	1008
961	targeted in transvesical	is present in the "Ground Truth	1009
962	prostatectomy? The transitional zone	Answer" and is necessary for a full	1010
963	of the prostate is targeted. </qa-pair>	response.	1011
964	<qa-pair> What is the focus of	3. **Relevance:** The "Generated Answer"	1012
965	transvesical prostatectomy in treating	must directly answer the user's	1013
966	urinary obstruction? The focus is on	"Query". An answer that is factually	1014
967	removing the hypertrophic transitional	correct but irrelevant to the question	1015
968	zone of the prostate. </qa-pair>	is "Incorrect".	1016
969	<qa-pair> What is included in the		1017
970	patient's history before transvesical	**Important Note:** Differences in	1018
971	prostatectomy? The patient's history	phrasing, verbosity, or style between	1019
972	includes a functional status inquiry.	the "Generated Answer" and the "Ground	1020
973	</qa-pair>	Truth Answer" are acceptable as long	1021
974		as the core semantic meaning is the	1022
975		same and the criteria above are met.	1023
976	Here is the passage for which you need to		
977	generate question-answer pairs		
978		B.1. Instruction prompt prefix for	1024
979	PASSAGE:	LLM-as-judge	1025
980	{text_chunk}		
981		### INSTRUCTIONS ###	1026
982	RESPONSE:	Follow these steps to generate your	1027
		evaluation:	1028
			1029
983	Appendix B. System prompt for	1. **Analyze the Query:** Carefully read	1030
984	LLM-as-judge	the user's "Query" to understand the	1031
		user's full intent and all a-priori	1032
985	### ROLE AND OBJECTIVE ###	information.	1033
986	You are an impartial and meticulous AI	2. **Study the Ground Truth:** Read the	1034
987	evaluator. Your objective is to	"Ground Truth Answer" to understand	1035
988	determine if a "Generated Answer"	the complete and factually correct	1036
989	correctly and completely answers a	response. This is your gold standard.	1037
990	given "Query", using the "Ground Truth	3. **Compare:** Compare the "Generated	1038
991	Answer" as the definitive source of	Answer" against the "Ground Truth	1039

1040	Answer" based on the **Factual	1091
1041	Consistency** and **Completeness**	1092
1042	criteria.	
1043	4. **Final Judgment:** Based on your	1093
1044	comparison, make a final binary	1094
1045	judgment.	1095
1046	5. **Output Format:** Provide your	1096
1047	response in the following format	1097
1048	<think> Your reasoning here... </think>	1098
1049	<answer> correct OR incorrect </answer>	1099
1050		1100
1051	Here are some examples	1101
1052		1102
1053	**Example 1**	1103
1054		1104
1055	QUERY	1105
1056	What is the escape velocity from the	1106
1057	surface of Earth?	1107
1058		1108
1059	GROUND TRUTH ANSWER	1109
1060	The escape velocity from Earth's surface	1110
1061	is approximately 11.2 kilometers per	1111
1062	second (km/s), which is about 40,320	1112
1063	kilometers per hour (km/h) or 25,000	1113
1064	miles per hour (mph).	1114
1065		1115
1066	GENERATED ANSWER	1116
1067	To escape Earth's gravity from the	1117
1068	surface, an object needs to reach a	1118
1069	speed of about 11.2 km/s.	1119
1070		1120
1071	YOUR RESPONSE	1121
1072	<think> Step 1: The query asks for Earth's	1122
1073	escape velocity. Step 2: The ground	1123
1074	truth states it is 11.2 km/s. Step 3:	1124
1075	The generated answer also states it is	1125
1076	11.2 km/s. Step 4: The generated	1126
1077	answer is factually consistent and	1127
1078	directly answers the query. While less	1128
1079	detailed than the ground truth, it	1129
1080	does not omit the critical piece of	1130
1081	information and contains no	1131
1082	contradictions. Therefore, it is	1132
1083	correct. </think>	1133
1084	<answer> correct </answer>	1134
1085		1135
1086	**Example 2**	1136
1087		1137
1088	QUERY	1138
1089	Who was the second person to walk on the	1139
1090	moon?	1140
		1141

GROUND TRUTH ANSWER

Neil Armstrong was the first; Buzz Aldrin was second.

GENERATED ANSWER

The second man on the moon was Michael Collins.

YOUR RESPONSE

<think> Step 1: The query asks for the second person on the Moon. Step 2: The ground truth identifies this person as Buzz Aldrin. Step 3: The generated answer incorrectly identifies the person as Michael Collins. Step 4: This is a direct factual contradiction with the ground truth. Therefore, the answer is incorrect. </think>
<answer> incorrect </answer>

****Example 3****

QUERY

What are Newton's first two laws of motion?

GROUND TRUTH ANSWER

Newton's first law states that an object will not change its motion unless a force acts on it. The second law states that the force on an object is equal to its mass times its acceleration.

GENERATED ANSWER

Newton's first law is the law of inertia, stating an object in motion stays in motion.

YOUR RESPONSE

<think> Step 1: The query asks for Newton's first AND second laws. Step 2: The ground truth provides both laws. Step 3: The generated answer only provides the first law. Step 4: The generated answer is incomplete as it omits a critical part of the information required by the query and present in the ground truth. Therefore, it is incorrect. </think>
<answer> incorrect </answer>

1142		Create new sub-questions in the following	1186
1143	Here are the "Query", the "Ground Truth	format but do NOT answer the question.	1187
1144	Answer", and the "Generated Answer".	Respond in the following format:	1188
1145			1189
1146	QUERY	<think> Your reasoning here... </think>	1190
1147	{query}	<sub-question> The first sub-question...	1191
1148		</sub-question>	1192
1149	GROUND TRUTH ANSWER	<sub-question> The second sub-question...	1193
1150	{ground_truth_answer}	</sub-question>	1194
1151		...	1195
1152	GENERATED ANSWER	<sub-question> The last sub-question...	1196
1153	{generated_answer}	</sub-question>	1197
1154			
1155	YOUR RESPONSE		
1156	Appendix C. Prompts Used in the	C.2. Answer Generation from Context	1198
1157	Surgical Information	Prompt	1199
1158	Assistant		
1159	This appendix contains all the prompts used in	This prompt is used to generate answers from re-	1200
1160	the Surgical Information Assistant codebase. These	trieved context for each sub-question. This results in	1201
1161	prompts are used to guide the language models in	multiple LLM calls, but can be run asynchronously:	1202
1162	performing various tasks such as question decomposi-		1203
1163	tion, information retrieval, answer synthesis, search-	Based on the given question and context,	1204
1164	ing a fallback data-store, best-effort answer genera-	generate an answer.	1205
1165	tion, and follow-up question generation.	Question: {question}	1206
		Context: {context}	1207
			1208
1166	C.1. Question Decomposition Prompt	Think step-by-step and make sure to reason	1209
1167	This prompt is used to break down complex surgical	through how to generate an answer.	1210
1168	questions into simpler sub-questions:	ONLY rely on the given context to	1211
		generate the answer.	1212
			1213
1169	You are an expert at breaking complex	Include snippets of the context that	1214
1170	surgical questions into simpler ones.	support your answer. Do NOT use any	1215
1171	Break the following question into	information outside of the given	1216
1172	smaller sub-questions:	context to generate the answer.	1217
1173			1218
1174	Question: {question}	Respond in the following format:	1219
1175			1220
1176	Each sub-question should be independent	<think> Your reasoning here... </think>	1221
1177	and answerable on it's own without	<answer> The generated answer... </answer>	1222
1178	needing reference to other	<snippet> First relevant snippet from the	1223
1179	sub-questions. Think of at least 3	context... </snippet>	1224
1180	sub-questions but no more than 7.	<snippet> Second relevant snippet from the	1225
1181		context... </snippet>	1226
1182	Think step-by-step and make sure to reason	...	1227
1183	through how break the question into	<snippet> The last relevant snippet from	1228
1184	sub-questions.	the context </snippet>	1229
1185			

1230	C.3. Prompt for answer generation without	what additional questions should be	1277
1231	context	asked.	1278
1232	This prompt is used to generate answers for sub-		1279
1233	questions when no context is available (for ablation	Original Question:	1280
1234	experiments).	{original_question}	1281
			1282
1235	You are a medical expert specializing in	Knowledge:	1283
1236	surgery. Answer the following question	{answers}	1284
1237	using your knowledge of surgical		1285
1238	procedures, anatomy, and medical	Think step-by-step to reason through you	1286
1239	practices.	answer and consider the relevant	1287
1240		information. Respond in the following	1288
1241	Question: {question}	format:	1289
1242		<think> Your reasoning here... </think>	1290
1243	Think step-by-step and provide a	<can_answer> yes OR no </can_answer>	1291
1244	comprehensive answer based on your	<answer> The answer to the original	1292
1245	medical knowledge. If you're uncertain	question... </answer>	1293
1246	about any aspect, please indicate that	<new_questions> The first new	1294
1247	in your response.	sub-question... </new_questions>	1295
1248		<new_questions> The second new	1296
1249	Respond in the following format:	sub-question... </new_questions>	1297
1250		...	1298
1251	<think> Your reasoning here... </think>	<new questions> The last new sub-question	1299
1252	<answer> The generated answer based on	</new_questions>	1300
1253	your medical knowledge... </answer>		
1254	<confidence> High/Medium/Low - your		
1255	confidence level in this answer		
1256	</confidence>		
		C.5. Wikipedia Search Prompt	1301
		This prompt is used to generate search queries for	1302
		Wikipedia:	1303
1257	C.4. Answer Synthesis Prompt	Given the following question, generate 3	1304
1258	This prompt is used to synthesize answers from mul-	search queries that would help find	1305
1259	multiple sub-questions or route for further iteration in	relevant information on Wikipedia. The	1306
1260	the DeRetSyn system.	queries should be specific and focused	1307
		on the key concepts in the question.	1308
			1309
1261	You are a reasoning engine. Given the	Question: {question}	1310
1262	following sub-question answers,		1311
1263	determine whether they are enough to	Respond in the following format:	1312
1264	fully answer the original question.	<query>first search query</query>	1313
1265	ONLY rely on the knowledge to	<query>second search query</query>	1314
1266	determine whether the question can be	<query>third search query</query>	1315
1267	answered.		
1268			
1269	If yes, then provide the answer. Make your	C.6. Wikipedia Context Synthesis Prompt	1316
1270	answer detailed and structured with	This prompt is used to synthesize information from	1317
1271	sections, as appropriate. Include as	Wikipedia contexts:	1318
1272	much relevant information as possible		
1273	from the knowledge provided.	You are a reasoning engine. Given the	1319
1274		following question and Wikipedia	1320
1275	If you determine that you cannot answer	contexts, synthesize the information	1321
1276	the original question, then suggest	to provide a comprehensive answer.	1322

1323			1369
1324	Question: {question}	Final Answer:	1370
1325		{final_answer}	1371
1326	Wikipedia Contexts:		1372
1327	{contexts}	Think step-by-step to reason through your	1373
1328		answer and consider the relevant	1374
1329	Think step-by-step to reason through your	information. Respond in the following	1375
1330	answer and consider the relevant	format:	1376
1331	information from the contexts. Respond	<think> Your reasoning here... </think>	1377
1332	in the following format:	<follow_up_questions> follow-up question	1378
1333	<think> Your reasoning here... </think>	here... </follow_up_questions>	1379
1334	<answer> The synthesized answer...	<follow_up_questions> follow-up question	1380
1335	</answer>	here... </follow_up_questions>	1381
		<follow_up_questions> follow-up question	1382
		here... </follow_up_questions>	1383
1336	C.7. Best Effort Answer Generation Prompt		
1337	This prompt is used when the system needs to gen-	C.9. Chain-of-Thought Generation Prompt	1384
1338	erate a best-effort answer using Wikipedia:	This prompt is used to generate detailed reasoning	1385
1339	You are a reasoning engine. Given the	for answers:	1386
1340	following original question and		
1341	sub-question answers, formulate an	You are a reasoning engine. Based on the	1387
1342	answer to the best of your ability.	following question and knowledge,	1388
1343		provide a detailed, step-by-step	1389
1344	Original Question:	reasoning to arrive at an answer.	1390
1345	{original_question}	Include at least 3 steps, but more as	1391
1346		needed.	1392
1347	Knowledge:		1393
1348	{state["answers"]}	Question:	1394
1349	{state["wikipedia_results"]}	{state["original_question"]}	1395
1350			1396
1351	Think step-by-step to reason through you	Knowledge:	1397
1352	answer and consider the relevant	{state["answers"]}	1398
1353	information. Respond in the following	{state["wikipedia_results"] if	1399
1354	format:	"wikipedia_results" in state else ""}	1400
1355	<think> Your reasoning here... </think>		1401
1356	<answer> The answer to the original	Provide your response in this format:	1402
1357	question... </answer>		1403
		<think> Your reasoning here... </think>	1404
		<answer> The final answer here... </answer>	1405
1358	C.8. Follow-up Question Generation Prompt		
1359	This prompt is used to generate follow-up questions:	Appendix D. Chat interaction	1406
1360	You are a reasoning engine. Given the	management prompts	1407
1361	following original question and final		
1362	answer, generate 3 follow-up questions	The following prompts are used to manage the chat-	1408
1363	that help expand on the original	interface with the Surgical Information Assistant.	1409
1364	question and the answer in a step-wise		
1365	manner.	D.1. Surgery Topic Classification Prompt	1410
1366		This prompt is used to determine if a question is	1411
1367	Original Question:	about surgery. If so, the DeRetSyn system is trig-	1412
1368	{original_question}		

1413	gered. If not, the LLM answers with general knowl-	rather requesting to explain a	1463
1414	edge acknowledging it may hallucinate and directs	previous response. </thinking>	1464
1415	the user to ask a more relevant question:	<answer> no </answer>	1465
			1466
1416	Determine if the following question	Question:	1467
1417	requires access to specific medical	What are some core differences between	1468
1418	documents related to detailed surgical	robotic and laparoscopic inguinal	1469
1419	information to be answered accurately.	hernia?	1470
1420	Think step-by-step and reason through your	Response:	1471
1421	answer. Respond in the following	<thinking> The question is asking about	1472
1422	format:	the differences between robotic and	1473
1423		laparoscopic inguinal hernia. Both	1474
1424	<thinking> Your reasoning here...	robotic and laparoscopic surgery	1475
1425	</thinking>	involve surgery on the inguinal	1476
1426	<answer> yes OR no </answer>	hernia. This is clearly a question	1477
1427		about surgery and requires	1478
1428	Here are some examples:	investigation into documentation to	1479
1429		answer. </thinking>	1480
1430	Question:	<answer> yes </answer>	1481
1431	What is the primary purpose of the		1482
1432	coronary artery bypass graft?	Question:	1483
1433	Response:	{question}	1484
1434	<thinking> The question is asking about a	Response:	1485
1435	coronary bypass graft which is related		
1436	to surgery. So yes, this question is		
1437	about surgery. </thinking>		
1438	<answer> yes </answer>		
1439			
1440	Question:		
1441	Is machine learning useful for solving		
1442	complex medical problems?		
1443	Response:		
1444	<thinking> The question is asking about		
1445	machine learning as a tool for solving		
1446	complex medical problems. While the		
1447	question does specify medical		
1448	problems, it does not refer to surgery		
1449	or topics related to surgery.		
1450	</thinking>		
1451	<answer> no </answer>		
1452			
1453	Question:		
1454	Can you elaborate on your previous		
1455	response about suturing or rephrase it?		
1456	Response:		
1457	<thinking> The question is asking for an		
1458	explanation of suturing that was given		
1459	earlier in the conversation. While the		
1460	question does mention suturing which		
1461	is related to surgery, it is not		
1462	asking about suturing specifically but		

D.2. Context-Aware Input Update Prompt

This prompt is used to update user input with context from chat history:

Given the following chat history and a new user input, update the user input to include any relevant context from the chat history. Only add necessary context, and keep the updated input concise.

Think step-by-step and reason through your answer. Respond in the following format:

<thinking> Your reasoning here...
</thinking>
<updated_input> Updated user input here...
</updated_input>

Here are some examples:

CHAT HISTORY:

User: What are the main types of sutures used in surgery?

Assistant: The main types of sutures used in surgery are absorbable and

1511	non-absorbable sutures. Absorbable	User: Are there any alternatives to this	1561
1512	sutures dissolve over time, while	procedure?	1562
1513	non-absorbable sutures need to be	Assistant: Yes, there are alternatives to	1563
1514	removed. Common absorbable sutures	laparotomy, particularly minimally	1564
1515	include catgut and synthetic materials	invasive techniques. The main	1565
1516	like polyglycolic acid. Non-absorbable	alternative is laparoscopy, also known	1566
1517	sutures are often made of materials	as keyhole surgery. In laparoscopy,	1567
1518	like nylon or silk.	several small incisions are made	1568
1519	User: How long do absorbable sutures	instead of one large incision. A	1569
1520	typically last?	camera and specialized instruments are	1570
1521	Assistant: The duration of absorbable	inserted through these small incisions	1571
1522	sutures varies depending on the	to perform the surgery. This technique	1572
1523	material used. Generally, they can	often results in less pain, faster	1573
1524	last anywhere from a few days to	recovery, and smaller scars compared	1574
1525	several months. For example,	to traditional laparotomy.	1575
1526	fast-absorbing sutures may dissolve		1576
1527	within 7-10 days, while	NEW USER INPUT:	1577
1528	slower-absorbing sutures can take	What are the risks?	1578
1529	60-90 days or even longer.		1579
1530	NEW USER INPUT:	RESPONSE:	1580
1531	What about non-absorbable ones?	<thinking> The user is asking about risks,	1581
1532	RESPONSE:	but it's not clear whether they're	1582
1533	<thinking> Considering the previous	referring to laparotomy or	1583
1534	discussion the user is referring to	laparoscopy, both of which were	1584
1535	non-absorbable sutures. Specifically,	discussed in the previous messages.	1585
1536	they are asking about the lifespan or	Since laparotomy was the original	1586
1537	removal time-frame for non-absorbable	topic and laparoscopy was introduced	1587
1538	sutures. I'll restate the question to	as an alternative, it would be helpful	1588
1539	include this context from the chat	to ask about the risks of both	1589
1540	history. </thinking>	procedures for a comprehensive answer.	1590
1541	<updated_input> What is the typical	</thinking>	1591
1542	lifespan or removal timeframe for	<updated_input> What are the risks	1592
1543	non-absorbable sutures used in	associated with both laparotomy and	1593
1544	surgery? </updated_input>	laparoscopy procedures?	1594
1545		</updated_input>	1595
1546			1596
1547			1597
1548	CHAT HISTORY:		1598
1549	User: What is a laparotomy?	CHAT HISTORY:	1599
1550	Assistant: A laparotomy is a surgical	User: What are the potential complications	1600
1551	procedure that involves making a large	of laparoscopic surgery?	1601
1552	incision in the abdomen to gain access	Assistant: Laparoscopic surgery can have	1602
1553	to the abdominal cavity. It allows	several potential complications,	1603
1554	surgeons to examine the abdominal	including:	1604
1555	organs, diagnose problems, and perform	1. Infection: Laparoscopic surgery can	1605
1556	necessary treatments. Laparotomies are	lead to infections, especially in	1606
1557	used for various purposes, including	patients with infections that can be	1607
1558	exploratory surgery, treatment of	spread through the abdominal cavity.	1608
1559	abdominal trauma, and removal of	2. Stool bleeding: Laparoscopic surgery	1609
1560	tumors.	can lead to stool bleeding, especially	1610

1611	if the surgical instrument is inserted	1658
1612	directly into the abdominal cavity.	1659
1613	3. Pain: Laparoscopic surgery can be	1660
1614	painful, especially for patients with	1661
1615	pre-existing conditions or those who	1662
1616	have had previous laparoscopic	1663
1617	surgeries.	1664
1618		1665
1619	NEW USER INPUT:	
1620	Can you think of any more?	
1621		
1622	RESPONSE:	
1623	<thinking> The user is asking if there are	
1624	any more complications, but it's	
1625	important to note that the original	
1626	question was about laparoscopic	
1627	surgery, not about potential	
1628	complications. I'll restate the	
1629	question to clarify that the original	
1630	topic was laparoscopic surgery.	
1631	</thinking>	
1632	<updated_input> Are there any more	
1633	potential complications associated	
1634	with laparoscopic surgery besides	
1635	infection, stool bleeding, and pain?	
1636	</updated_input>	
1637		
1638		
1639		
1640	CHAT HISTORY:	
1641	{formatted_history}	
1642		
1643	New User Input:	
1644	{user_input}	
1645		
1646	RESPONSE:	
1647	Appendix E. Prompts used to test	
1648	alternative RAG	
1649	paradigms	
1650	E.1. Chain-of-thought prompting without	
1651	context	
1652	You are a medical expert. Please answer	
1653	the following question:	
1654	Think step-by-step and provide a detailed	
1655	reasoning process to arrive at your	
1656	answer. Include at least 3 steps in	
1657	your reasoning, but more as needed.	
	Respond in the following format:	
	<think> Your reasoning here... </think>	
	<answer> Your final answer here...	
	</answer>	
	Question: {question}	
	E.2. Chain-of-thought prompting with	
	context	
	You are a medical assistant specializing	
	in surgical information. Use the	
	following context to answer the	
	question.	
	If you cannot find the answer in the	
	context, say "I don't have enough	
	information to answer this question."	
	Context:	
	{context_docs}	
	Question: {question}	
	Think step-by-step to reason through your	
	answer and consider the relevant	
	information from the context. Respond	
	in the following format:	
	<think> Your reasoning here... </think>	
	<answer> The answer to the question...	
	</answer>	
	E.3. ReAct Agent System Prompt	
	This prompt is used for the ReAct agent in the eval-	
	uation scripts:	
	You are a medical assistant specializing	
	in surgical information. Your goal is	
	to provide accurate and helpful	
	information about surgical procedures,	
	techniques, and related medical	
	knowledge.	
	When answering questions:	
	1. Use the DocumentSearch tool to find	
	relevant information in the surgical	
	database	
	2. Analyze the retrieved information	
	carefully	

3. Provide comprehensive and accurate answers based on the retrieved content
 4. If the information is not available in the search results, acknowledge the limitations
 5. Think step-by-step to reason through complex questions
 6. Cite specific parts of the retrieved documents when appropriate
 7. Focus on providing factual medical information rather than opinions
- Remember accuracy is crucial. Provide all reasoning and the final answer.

Appendix F. Prompt Design Principles

The prompts used in the Surgical Information Assistant follow several key design principles:

F.1. Step-by-Step Reasoning

All prompts encourage the language model to think step-by-step and provide detailed reasoning before arriving at an answer. This approach, often referred to as chain-of-thought prompting, has been shown to improve the accuracy and reliability of language model outputs, especially for complex tasks.

F.2. Structured Output Format

The prompts use a consistent structured output format with XML-like tags (e.g., `<think>`, `<answer>`, `<sub-question>`) to clearly separate different components of the response. This structured format makes it easier to parse and extract specific information from the language model's output.

F.3. Few-Shot Examples

Several prompts include few-shot examples to demonstrate the expected reasoning process and output format. These examples help guide the language model to produce responses that follow the desired pattern and meet the quality standards.

F.4. Task-Specific Instructions

Each prompt contains clear, task-specific instructions that define the role of the language model (e.g., "You

are an expert at breaking complex surgical questions into simpler ones") and the specific task it needs to perform. This helps focus the model's attention on the relevant aspects of the task.

F.5. Context Awareness

The prompts are designed to make the language model aware of the context in which it operates, including the original question, retrieved information, and previous interactions. This context awareness is crucial for generating coherent and relevant responses.

Appendix G. Example responses from different experimental designs

G.1. Question 1

Why should soiled instruments be cleaned within twenty minutes of contamination?

Known answer: Soiled instruments should be cleaned within twenty minutes because stainless steel can experience corrosion from biological and chemical liquids.

Llama32 Deretsyn: The soiled instruments should be cleaned within twenty minutes of contamination to prevent the potential risks associated with them, including corrosion, injury from sharp instruments, and the spread of infection due to the presence of biological and chemical liquids. This time frame is significant because it is the time frame within which organic matter such as blood or tissue can be prevented from drying on the surface of the instrument. If not cleaned within this time frame, the consequences can include damage to the instruments, corrosion, and the potential for injury. Furthermore, there are specific guidelines that dictate the cleaning of soiled instruments within twenty minutes of contamination.

Llama32 RAG: Soiled instruments should be cleaned within twenty minutes of contamination to prevent the growth of

microorganisms and the drying of organic matter, which can lead to the formation of biofilms and make the instrument more difficult to clean and disinfect.

Llama32 React: Soiled instruments should be cleaned within twenty minutes of contamination to prevent the growth and spread of microorganisms. This is because microorganisms can multiply rapidly, especially in warm, moist environments, and can cause infections if not removed promptly.

GPT40: Soiled instruments should be cleaned within twenty minutes of contamination to prevent organic material from drying and adhering to the surface, to reduce the risk of biofilm formation that can make sterilization less effective, to minimize the risk of cross-contamination and infection, to maintain the integrity and functionality of the instruments, and to adhere to best practices in infection control.

Llama31: Soiled instruments should be cleaned within twenty minutes of contamination to prevent bacterial growth, biofilm formation, and subsequent infection transmission. This timeframe allows for effective cleaning and disinfection, reducing the risk of hospital-acquired infections and ensuring patient safety.

G.2. Question 2:

What factors can put the liver remnant at risk for further ischemic injury postoperatively after liver resection surgery?

Known answer: Hypoxia, hypotension, septicemia, or drug toxicity can put the liver remnant at risk for further ischemic injury postoperatively.

Llama32 Deretsyn: The factors that can put the liver remnant at risk for further ischemic injury postoperatively include:

1. **Hemodynamic instability**: This can affect blood flow to the liver remnant and increase the risk of ischemic injury.
2. **Hyperdynamic circulation**: This can also affect blood flow to the liver remnant and increase the risk of ischemic injury.
3. **Septicemia**: This is a potential cause of ischemic injury in the liver remnant postoperatively.
4. **Drug toxicity**: This is another potential cause of ischemic injury in the liver remnant postoperatively.
5. **Volume expansion**: This can affect blood flow to the liver remnant and increase the risk of ischemic injury.
6. **Glycemic control**: Poor glycemic control can lead to hyperglycemia, which can contribute to ischemic injury in the liver remnant.
7. **Liver regeneration**: While liver regeneration is a natural process, it can also lead to changes in blood flow to the liver remnant, increasing the risk of ischemic injury.
8. **Ascites**: This can put pressure on the liver remnant and impair venous return to the heart, increasing the risk of ischemic injury.
9. **Bleeding**: This can lead to hemodynamic instability and affect blood flow to the liver remnant, increasing the risk of ischemic injury.
10. **Portal hypertension**: This can affect blood flow to the liver remnant and increase the risk of ischemic injury.
11. **Cardiac function**: Poor cardiac function can lead to hypotension and affect blood flow to the liver remnant, increasing the risk of ischemic injury.
12. **Coagulation status**: Abnormal coagulation status can lead to bleeding and affect blood flow to the liver remnant, increasing the risk of ischemic injury.
13. **Warm ischemia time**: Prolonged warm ischemia time can increase the risk of ischemic injury in the liver remnant.

1890	14. **Liver function** : Poor liver	systemic and local hypoxia,	1941
1891	function can affect blood flow to the	inflammatory and oxidative stress	1942
1892	liver remnant and increase the risk of	related to ischemia-reperfusion	1943
1893	ischemic injury.	injury, pre-existing conditions that	1944
1894	15. **Postoperative factors** : Hypoxia,	impair liver function (e.g.,	1945
1895	hypotension, septicemia, drug	cirrhosis, steatosis, diabetes), poor	1946
1896	toxicity, hyperdynamic circulation,	postoperative hemodynamic management	1947
1897	and hyperglycemia can all contribute	(e.g., hypotension, inadequate	1948
1898	to ischemic injury in the liver	resuscitation), and a small or	1949
1899	remnant postoperatively.	insufficiently functional liver	1950
1900	16. **Surgical procedure** : The surgical	remnant that cannot meet metabolic	1951
1901	procedure itself can increase the risk	demands.	1952
1902	of hypoxia and hypotension, prolong	Llama31: The liver remnant is at risk for	1953
1903	warm ischemia time, and cause	further ischemic injury	1954
1904	hemodynamic instability, all of which	postoperatively due to a combination	1955
1905	can increase the risk of ischemic	of factors, including:	1956
1906	injury in the liver remnant.		1957
1907		- Hemodynamic factors (e.g., changes in	1958
1908	These factors can all contribute to	blood flow, blood pressure, or cardiac	1959
1909	ischemic injury in the liver remnant	output)	1960
1910	postoperatively, either by affecting	- Surgical factors (e.g., extent of liver	1961
1911	blood flow to the liver remnant or by	resection, presence of major blood	1962
1912	increasing the risk of hypoxia,	vessels in the area of resection)	1963
1913	hypotension, septicemia, and drug	- Patient-related factors (e.g., overall	1964
1914	toxicity.	health, age, presence of comorbidities)	1965
1915	Llama32 RAG: Hypoxia, hypotension,	- Specific conditions (e.g., small liver	1966
1916	septicemia or drug toxicity, and	remnant, portal hypertension, biliary	1967
1917	hyperdynamic circulation due to	obstruction, dehydration or	1968
1918	cirrhosis or impaired liver function	hypovolemia)	1969
1919	can put the liver remnant at risk for	- Inadequate intraoperative and	1970
1920	further ischemic injury	postoperative care (e.g., fluid	1971
1921	postoperatively.	management, blood pressure management,	1972
1922	Llama32 React: The following factors can	monitoring and management of	1973
1923	put the liver remnant at risk for	complications)	1974
1924	further ischemic injury		
1925	postoperatively: 1. Warm ischemia		
1926	time, 2. Fatty or cirrhotic livers, 3.		
1927	Large resections, 4. Prolonged		
1928	surgical time, 5. Blood loss and		
1929	cardiovascular instability, 6.		
1930	Non-steroidal anti-inflammatory drugs		
1931	(NSAIDs), 7. Paracetamol		
1932	(acetaminophen) use, 8. Hypoglycemia,		
1933	9. Hypothermia, 10. Poor patient		
1934	selection.		
1935	GPT40: Factors that can put the liver		
1936	remnant at risk for further ischemic		
1937	injury postoperatively include		
1938	vascular and perfusion issues (e.g.,		
1939	portal vein thrombosis, hepatic artery		
1940	injury, small-for-size syndrome),		