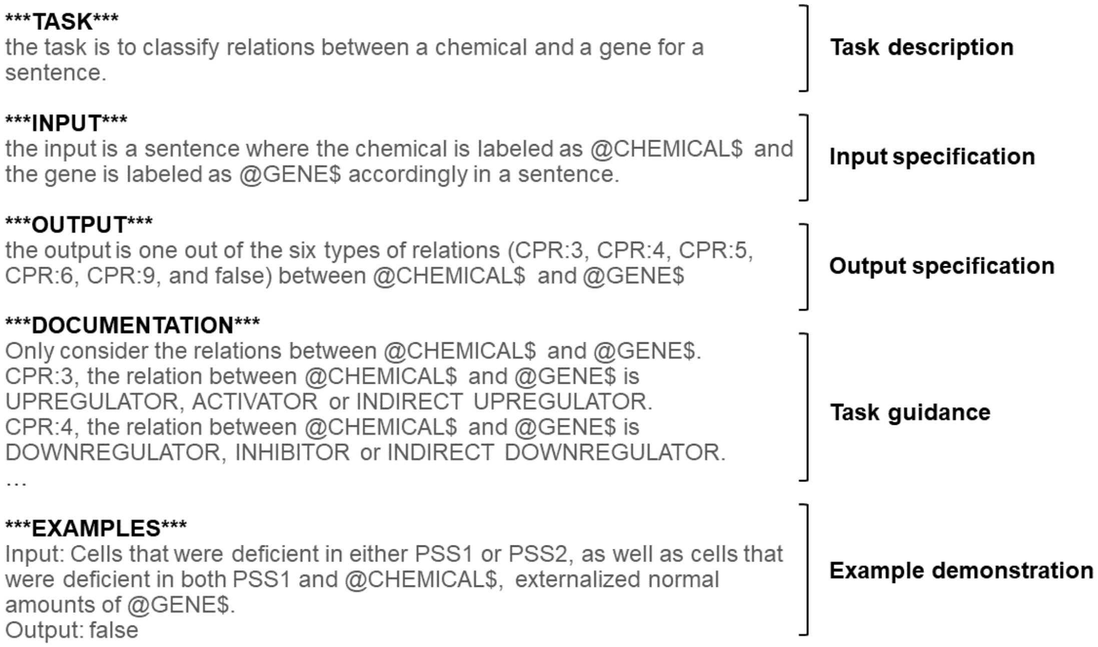
Supplementary Material

S1. Prompt engineering



*****TASK*****
the task is to classify relations between a chemical and a gene for a sentence.

*****INPUT*****
the input is a sentence where the chemical is labeled as @CHEMICAL$ and the gene is labeled as @GENE$ accordingly in a sentence.

*****OUTPUT*****
the output is one out of the six types of relations (CPR:3, CPR:4, CPR:5, CPR:6, CPR:9, and false) between @CHEMICAL$ and @GENE$

*****DOCUMENTATION*****
Only consider the relations between @CHEMICAL$ and @GENE$.
CPR:3, the relation between @CHEMICAL$ and @GENE$ is UPREGULATOR, ACTIVATOR or INDIRECT UPREGULATOR.
CPR:4, the relation between @CHEMICAL$ and @GENE$ is DOWNREGULATOR, INHIBITOR or INDIRECT DOWNREGULATOR.
…

*****EXAMPLES*****
Input: Cells that were deficient in either PSS1 or PSS2, as well as cells that were deficient in both PSS1 and @CHEMICAL$, externalized normal amounts of @GENE$.
Output: false

Figure S1. An annotated one-shot prompt example for biomedical relation extraction.

S1.1 An annotated prompt example

Figure S1 shows an example prompt for relation extraction in ChemProt. The prompt consists of (1) task descriptions (e.g., classifying relations), (2) input specifications (e.g., a sentence with labeled entities), (3) output specifications (e.g., the relation type), (4) task guidance (e.g., detailed descriptions on relation types), and (5) example demonstrations if an example is provided.

All the prompts used for the 12 benchmarks are also publicly available via https://github.com/BIDS-Xu-Lab/Biomedical-NLP-Benchmarks.

S1.2 Parameters

For zero-, one-, and few-shot approaches, we used a temperature of 0 to minimize variance for both GPT and LLaMA-based models. Additionally, for LLaMA models, we kept other parameters unchanged and set the maximum number of generated tokens (the max_new_tokens parameter) per task. The maximum number of generated tokens was 512, 20, 64, 128, 512, and 512 for named entity recognition, relation extraction, multi-label document classification, question answering, text summarization, and text simplification, respectively.

The related codes are also available via the repository.

S2. Quantitative evaluation results

S2.1. Result reporting

For quantitative evaluation, to quantify statistical significance between the LLM performance on the 12 datasets, we performed a two-tailed Wilcoxon rank-sum test with bootstrapping. We used a subsample size of 30 and repeated the process 100 times at a 95% confidence interval.

We employed the two-sided Wilcoxon rank-sum test to conduct comparative analyses between two lists of bootstrapped metrics from different LLMs. Specifically, we utilized the ranksums function from the scipy.stats package with parameters alternative='two-sided' and nan_policy='omit' for statistical analysis.

S2.2 Primary and secondary evaluation metric results

Table S1. Quantitative evaluations of the LLMs on the 12 benchmarks under zero-shot, one-shot, and fine-tuned settings. Both primary metric and secondary metric results are reported. State-of-the-art (SOTA) results, representing the reported best performance of studies using fine-tuned (domain-specific) language models before the LLMs and their backbone models, are also provided. The SOTA results are directly extracted from the studies. [1]the study reported accuracy on MedQA (4-option); we applied the released model for inference on MedQA (5-option).

		SOTA results before LLMs	**Zero-shot**			**One-shot**			**Fine-tuned**	
			GPT-3.5	**GPT-4**	**LLaMA 2 13B**	**GPT-3.5**	**GPT-4**	**LLaMA 2 13B**	**LLaMA 2 13B**	**PMC LLaMA 13B**
Named entity recognition										
BC5CDR-chemical	Entity F1	0.9500	0.6274	**0.7993**	0.3944	0.7133	**0.8327**	0.6276	**0.9149**	0.9063
NCBI Disease	Entity F1	0.9090	0.4060	**0.5827**	0.2211	0.4817	**0.5988**	0.3811	**0.8682**	0.8353
Relation extraction										
ChemProt	Macro F1	0.7344	0.1345	**0.3250**	0.1392	0.1280	**0.3391**	0.0718	**0.4612**	0.3111
	Micro F1		0.2376	**0.3538**	0.1232	0.2011	**0.4109**	0.0174	**0.8006**	0.7659
DDI2013	Macro F1	0.7919	0.2004	**0.2968**	0.1305	0.2126	**0.3312**	0.1779	**0.6218**	0.5700
	Micro F1		**0.3862**	0.3830	0.1650	0.2850	**0.4819**	0.4518	**0.8503**	0.8438
Multi-label document classification										
HoC	Macro F1	0.8882	0.6722	**0.7109**	0.1285	0.6671	**0.7093**	0.3072	**0.6957**	0.4221
	Micro F1		0.6605	**0.7166**	0.1495	0.6711	**0.7205**	0.3813	**0.6787**	0.4536
LitCovid	Macro F1	0.8921	**0.5967**	0.5883	0.3825	**0.6009**	0.5901	0.4808	**0.5725**	0.4273
	Micro F1		0.6707	**0.6809**	0.5400	0.6656	**0.6839**	0.5997	**0.6668**	0.5632
Question answering										
MedQA (5-Option)	Accuracy	0.4195[1]	0.4988	**0.7156**	0.2522	0.5161	**0.7439**	0.2899	**0.4462**	0.3975
	Macro F1		0.4096	**0.5104**	0.2226	0.4241	**0.6171**	0.2750	**0.4394**	0.3932
PubMedQA	Accuracy	0.7340	**0.6560***	0.6280	0.5520	0.4600	**0.7100**	0.2660	**0.8040**	0.7680
	Macro F1		**0.4648**	0.4604	0.2959	0.3761	**0.5692**	0.1916	**0.5628**	0.5358
Text summarization										
PubMed	Rouge-L	0.4316	0.2274	**0.2419**	0.1190	0.2351	**0.2427**	0.0989	**0.1857**	0.1684
	BERT score		0.7195	**0.7242**	0.6546	0.7227	**0.7231**	0.6467	**0.6623**	0.6542
	BART score		-4.8334	-4.8375	**-4.7504**	**-4.8114**	-4.8548	-5.8035	-4.9772	**-4.9212**
MS^2	Rouge-L	0.2080	0.0889	**0.1224**	0.0948	0.1132	**0.1248**	0.0320	**0.0934**	0.0059

	BERT score		0.6652	**0.6904**	0.6431	0.6775	**0.6907**	0.5921	**0.6265**	0.0275
	BART score		**-5.0578**	-5.1767	-5.2804	**-5.1830**	-5.1909	-5.9733	**-5.4586**	-6.7826
Text simplification										
Cochrane	Rouge-L	0.4476	0.2365	**0.2375**	0.2081	**0.2447***	0.2385	0.2207	0.2355	**0.2370***
	FKG		11.8815	11.0179	**12.4385**	**12.3490**	11.6752	12.2585	**11.8200**	11.0660
	DCR		**10.1221**	9.5580	9.8786	**10.1796**	10.0239	9.8024	**9.8072**	9.7508
PLOS	Rouge-L	0.4368	**0.2323***	0.2253	0.2121	**0.2449***	0.2386	0.1836	**0.2583**	0.2577
	FKG		12.5182	11.1895	**13.1722**	13.2913	12.0144	**13.3804**	**13.9517**	13.6370
	DCR		**11.0496**	9.9077	10.1079	**10.8338**	10.1151	9.6846	**10.6670**	10.5341

Table S1 shows the detailed results of LLMs on the 12 benchmarks using primary and secondary evaluation metrics.

S2.3 Performance mean, variance, and confidence intervals

Table S2. Mean, variance, and confidence intervals (shown in brackets) of the results obtained using bootstrapping with a subsample size of 30 and repeated the process 100 times at a 95% confidence interval.

		Zero-shot			**One-shot**			**Fine-tuned**	
		GPT-3.5	**GPT-4**	**LLaMA2 13B**	**GPT-3.5**	**GPT-4**	**LLaMA2 13B**	**LLaMA2 13B**	**PMC LLaMA 13B**
Named entity recognition									
BC5CDR-chemical	Entity F1	0.6111±0.0925 (0.4460, 0.7934)	0.8026±0.0788 (0.6376, 0.9632)	0.3774±0.0943 (0.1947, 0.5460)	0.7037±0.0966 (0.5128, 0.8947)	0.8398±0.0748 (0.6941, 0.9663)	0.6192±0.0912 (0.4499, 0.7947)	0.9145±0.0522 (0.7769, 0.9857)	0.9048±0.0535 (0.7785, 0.9916)
NCBI Disease	Entity F1	0.4046±0.0920 (0.2051, 0.5799)	0.5958±0.0994 (0.3837, 0.7710)	0.2364±0.1020（0.0645，0.4796）	0.4930±0.0976 (0.2727, 0.6729)	0.6172±0.0985 (0.4287, 0.8286)	0.3903±0.0796 (0.2205, 0.5385)	0.8728±0.0595 (0.7302, 0.9693)	0.8442±0.0817 (0.6302, 0.9569)
Relation extraction									
ChemProt	Macro F1	0.2104±0.0945 (0.0650, 0.4399)	0.2941±0.1094 (0.1103, 0.5110)	0.0969±0.0739 (0.0059, 0.2946)	0.1888±0.0789 (0.0594, 0.3702)	0.3047±0.1123 (0.1153, 0.5405)	0.0361±0.0503 (0.0000, 0.1871)	0.4312±0.1460 (0.1762, 0.7412)	0.3534±0.1500 (0.1475, 0.7129)
DDI2013	Macro F1	0.2291±0.0987 (0.0776, 0.4159)	0.2671±0.0941 (0.1258, 0.4400)	0.1001±0.0733 (0.0271, 0.2768)	0.1831±0.0914 (0.0414, 0.3786)	0.3121±0.0997 (0.1463, 0.5121)	0.1879±0.0906 (0.0888, 0.4259)	0.5630±0.1445 (0.2592, 0.8223)	0.5612±0.1718 (0.2296, 0.9080)
Multi-label document classification									
HoC	Macro F1	0.6115±0.0723 (0.4816, 0.7543)	0.6679±0.0694 (0.5504, 0.7922)	0.1243±0.0511 (0.0384, 0.2406)	0.6119±0.0774 (0.4611, 0.7773)	0.6719±0.0680 (0.5484, 0.8028)	0.2851±0.0476 (0.1977, 0.3788)	0.6206±0.0821 (0.4719, 0.7687)	0.3840±0.0713 (0.2524, 0.5283)
LitCovid	Macro F1	0.5626±0.0712 (0.4296, 0.6883)	0.5583±0.0695 (0.4424, 0.7075)	0.3747±0.0308 (0.3046, 0.4356)	0.5788±0.0704 (0.4264, 0.7181)	0.5652±0.0677 (0.4592, 0.7103)	0.4589±0.0566 (0.3493, 0.5591)	0.5265±0.0758 (0.3844, 0.6826)	0.4114±0.0442 (0.3305, 0.5056)
Question answering									
MedQA (5-Option)	Accuracy	0.4960±0.0962 (0.2825, 0.6667)	0.7193±0.0854 (0.5333, 0.8842)	0.1903±0.0659 (0.0667, 0.3175)	0.5163±0.0987 (0.2983, 0.6667)	0.7523±0.0735 (0.6000, 0.9000)	0.2827±0.0853 (0.1158, 0.4333)	0.4480±0.0950 (0.2667, 0.6175)	0.3957±0.0926 (0.2333, 0.5667)
PubMedQA	Accuracy	0.6107±0.0869 (0.4667, 0.7842)	0.6480±0.0955 (0.4667, 0.8000)	0.5303±0.0958 (0.3333, 0.7000)	0.4607±0.0862 (0.3000, 0.6333)	0.7020±0.0918 (0.5333, 0.8667)	0.2707±0.0880 (0.1333, 0.4333)	0.8083±0.0719 (0.7000, 0.9333)	0.7563±0.0761 (0.6000, 0.8842)
Text summarization									
PubMed	Rouge-L	0.2278±0.0135 (0.2045, 0.2535)	0.2407±0.0115 (0.2177, 0.2628)	0.1197±0.0129 (0.0974, 0.1461)	0.2361±0.0122 (0.2165, 0.2618)	0.2408±0.0122 (0.2157, 0.2626)	0.0990±0.0032 (0.0935, 0.1048)	0.1819±0.0212 (0.1477, 0.2274)	0.1697±0.0159 (0.1416, 0.2028)

MS^2	Rouge-L	0.0895±0.0075 (0.0736, 0.1031)	0.1218±0.0058 (0.1121, 0.1344)	0.0948±0.0073 (0.0833, 0.1115)	0.1146±0.0089 (0.0989, 0.1311)	0.1238±0.0060 (0.1118, 0.1350)	0.0320±0.0031 (0.0257, 0.0386)	0.0931±0.0114 (0.0738, 0.1189)	0.0059±0.0053 (0.0000, 0.0175)
Text simplification									
Cochrane	Rouge-L	0.2359±0.0130 (0.2099, 0.2582)	0.2361±0.0110 (0.2162, 0.2575)	0.2085±0.0121 (0.1865, 0.2344)	0.2445±0.0128 (0.2171, 0.2702)	0.2387±0.0106 (0.2198, 0.2596)	0.2190±0.0138 (0.1973, 0.2481)	0.2309±0.0166 (0.2044, 0.2659)	0.2404±0.0182 (0.2108, 0.2836)
PLOS	Rouge-L	0.2312±0.0117 (0.2100, 0.2585)	0.2255±0.0087 (0.2104, 0.2459)	0.2110±0.0089 (0.1970, 0.2281)	0.2462±0.0107 (0.2267, 0.2651)	0.2382±0.0110 (0.2182, 0.2636)	0.1836±0.0148 (0.1553, 0.2123)	0.2569±0.0149 (0.2306, 0.2907)	0.2570±0.0145 (0.2312, 0.2837)

Table S2 further shows the performance mean, variance, and confidence intervals of the primary metrics. The results are obtained using bootstrapping with a subsample size of 30 and repeated the process 100 times at a 95% confidence interval.

S2.4 Statistic test results

The **quantitative_evaluation_statistic_test.xlsx** file provides the detailed statistic test results to quantify the statistical difference between the LLM performance on the 12 datasets.

S2.5. Dynamic K-nearest few-shot results

Table S3. Detailed dynamic K-nearest few-shot results.

		GPT-3.5			GPT-4		
		K = 1 shot	2 shot	5 shot	1 shot	2 shot	5 shot
Named entity recognition							
BC5CDR-chemical	Entity F1	0.6728	0.6738	0.6787	0.7904	0.7959	0.7950
NCBI Disease	Entity F1	0.4154	0.4093	0.3769	0.5262	0.4980	0.4881
Relation extraction							
ChemProt	Macro F1	0.1419	0.1455	0.1463	0.2205	0.3353	0.3435
	Micro F1	0.2408	0.1824	0.1820	0.4396	0.5092	0.5275
DDI2013	Macro F1	0.1229	0.2419	0.2078	0.2965	0.2936	0.3175
	Micro F1	0.3281	0.3508	0.3822	0.5206	0.5020	0.5248
Multi-label document classification							
HoC	Macro F1	0.6931	0.6885	0.7242	0.7177	0.7035	0.7322
	Micro F1	0.6866	0.6885	0.7290	0.7207	0.7105	0.7330
LitCovid	Macro F1	0.6364	0.6357	0.6484	0.6500	0.6633	0.7055
	Micro F1	0.7048	0.7056	0.7229	0.7359	0.7546	0.7927
Question answering							
MedQA (5-Option)	Accuracy	0.5192	0.5153	0.5295	0.7376	0.7596	0.7753
	Macro F1	0.4311	0.4280	0.4397	0.6143	0.6328	0.6468
PubMedQA	Accuracy	0.5740	0.5740	0.6500	0.7200	0.7420	0.7560
	Macro F1	0.5073	0.4964	0.5242	0.5728	0.5851	0.5836
Text summarization							
PubMed	Rouge-L	0.2472	0.2482	0.2438	0.2524	0.2538	0.2411
	BERT score	0.7267	0.7278	0.7251	0.7273	0.7271	0.7200
	BART score	-4.7751	-4.7609	-4.7920	-4.8142	-4.7919	-4.8697

MS^2	Rouge-L	0.0974	0.0946	0.0946	0.1230	0.1234	0.1226
	BERT score	0.6681	0.6660	0.6654	0.6879	0.6877	0.6868
	BART score	-5.1329	-5.1411	-5.1428	-5.2626	-5.2633	-5.2838
Text simplification							
Cochrane	Rouge-L	0.2528	0.2538	0.2536	0.2567	0.2587	0.2587
	FKG	12.3898	12.4502	12.5908	11.8360	11.6763	11.8388
	DCR	9.7913	9.8791	9.9214	9.5561	9.4289	9.4811
PLOS	Rouge-L	0.2451	0.2466	0.2482	0.2430	0.2416	0.2430
	FKG	13.2760	13.2652	13.3785	12.1148	12.0537	12.0698
	DCR	10.8511	10.7889	10.7330	10.1539	10.0404	10.0848

Table S3 shows the detailed dynamic K-nearest few-shot results.

S3. Qualitative evaluation on the PubMed Text Summarization Benchmark

S3.1 Annotation guideline

For each model response, please rate the following. Keep in mind that the order of the model outputs may be randomly shuffled, meaning that what's Model 1 in sample 1 could become Model 2 in sample 2 as an example.

Rating ranges from 1 (bad) 2,3,4 to 5 (good):

1. Accuracy of generated summaries: Does the summary contain correct information from the original article?

1 (bad): The summary includes false or misleading information that is significantly different from the original article.

2: Some elements of the summary contain correct information however the overall summary is inaccurate.

3: The main point of the summary is correct, however it may include some inaccurate information from the original article.

4: The summary mostly avoids inaccuracies, but may include minor inaccurate information from the original article.

5 (good): The summary is accurate based on the original article.

Additional annotations: if the response contains false or misleading information, please identify them.

2. Completeness of generated summaries: Does the summary capture the key information from the original article?

1 (bad): The summary is incomplete (missing key information), or leaves out crucial details.

2: The summary is somewhat complete, but it lacks key information that impact its comprehensiveness.

3: The summary is moderately complete, but certain details are missing, requiring enhancement.

4: The summary is largely comprehensive, but a few minor details could be refined for better alignment with the original article.

5 (good): The summary text is comprehensive and includes all relevant information.

3. Readability of generated summaries: Is the summary easy to read?

1 (bad): The text is highly difficult to read, full of grammatical errors, and lacks coherence and clarity.

2: The text is somewhat difficult to read, and there are occasional grammatical errors. The coherence and clarity could be improved.

3: The text is moderately easy to read, but there are noticeable grammatical errors and some parts lack coherence and clarity.

4: The text is fairly easy to read, with only a few minor grammatical errors. Overall coherence and clarity are good, but there is room for improvement.

5 (good): The text is easy to read, well-structured, and flows naturally.

S3.2 Statistic test results

The **qualitative_evaluation_summarization_statistic_test.xlsx** file provides the detailed statistic test results.

S3.3 Examples

Example 1

1.1 Gold standard:

purposethis study was aimed to determine the effect of two chemically distinct denture cleansers and water on the surface hardness of acrylic and silicone based soft denture liners at various time intervals.materials and methodstwo commonly used commercial resilient liner material were selected based on their chemical composition (silicone- and acrylic - based soft liners) for this investigation .
120 cylindrical specimens were made of 15 mm 10 mm dimensions (according to astm : d-2240 - 64 t) in a custom made metal mold .
all specimens were stored in artificial saliva throughout the study .
forty specimens were cleansed daily in 0.5% sodium hypochlorite solution ; forty were cleansed in

sodium perborate and remaining forty specimens were daily rinsed in water .

testing was done at 1 week , 1 month , 3 months and 6 months for surface hardness using a shore a durometer .

a mean of 3 reading for each sample was subjected to one - way anova , post hoc test and pair - t test for statistical analysis .

p values of less than 0.05 were taken as statistically significant.resultssurface hardness of all the samples was significantly higher after a period of 6 months irrespective of the cleansing treatment .

minor changes were observed between control , sodium hypochlorite and sodium perborate groups with time .

greater change was observed in surface hardness of acrylic - based soft denture liners as compared to silicone - based soft liners for all groups , as time progressed.conclusionsilicone-based soft denture liners performed significantly better in all cleansing treatments than acrylic - based soft denture liners .

1.2 Fine-tuned BART:

Annotations: Accuracy: 4.5; Completeness: 3; Readability: 3

purposethe aim of this study was to determine the effect of two chemically distinct denture cleansers and water on the surface hardness of acrylic and silicone based soft denture liners at various time intervals.materials and methodsthe study was conducted in maharishi markandeshwar college of dental sciences and research , mullana , ambala , haryana , india .

two commonly used commercial resilient liner materials based on their chemical composition (silicone- and acrylic - based soft liners) were selected for the study .

a total of hundred and twenty cylindrical specimens were made of the dimensions 15 mm 10 mm (according to astm : d-2240 - 64t) .

soft denture liners were manipulated according to manufacturer 's instructions and expressed into the mold .

the mold was then covered from the top by a cellophane sheet and another glass slab was pressed tightly against the mold to remove excess material and to shape the specimens according to the dimensions of the mold .

group a (control) , group b (sodium hypochlorite) and group c (sodium perborate) comprising of 40 specimens each . each group was divided further into 4 subgroups : subgroup i , ii , iii and iv (consisted of 10 specimens each) to be tested at a time interval of 1 week , 1 month , 3 months and 6 months respectively .

each subgroup was again divided into two minor subgroups : minor subgroup a - consisted samples made of acrylic based soft denture liner and minor subgroup b - consisted of samples made of silicone based soft denture liner .

specimens in group a (control) were cleansed daily by rinsing with water and then were stored in artificial saliva for the entire period of the study .

specimens in group b were immersed in 0.5% sodium hypochlorite solution for ten minutes daily , rinsed in water and stored in artificial saliva at room temperature .

specimens in group c were cleansed in a solution of sodium perborate denture cleansing tablets (dissolved in 250 ml water as recommended by the manufacturer) .

specimens of all 3 groups were tested at 1 week , 1 month and 3 months time intervals at central institute of plastics engineering & technology (cipet) , panipat , haryana.resultsthe specimens were tested using a shore a durometer which

1.3 GPT-3.5 zero-shot

Annotations: Accuracy: 4.5; Completeness: 3.5; Readability: 5

soft denture lining materials, such as acrylic and silicone types, are useful for treating patients with ridge atrophy or resorption, bony undercuts, bruxing tendencies, congenital or acquired oral defects, xerostomia, and dentures opposing natural dentition. however, these materials present problems such as weakening of bond, loss of resiliency, color alterations, and porosity. rougher surfaces of the lining materials can enhance adhesion of microorganisms and fungal growth, so regular cleansing procedures are necessary to prevent infection. mechanical cleansing is not advisable for soft denture liners, so chemical cleansing using denture cleansers is the preferred method. alkaline peroxide and sodium hypochlorite cleansers are commonly used, with sodium perborate being the most effective. in a study comparing acrylic and silicone-based soft denture liners, it was found that silicone-based liners performed better in terms of surface hardness over a period of 6 months when cleansed daily. chemical denture cleansers can be used daily without affecting the surface hardness of soft denture liners, and silicone-based liners showed better compatibility with cleansing solutions and maintained their resiliency better.

1.4 GPT-4 zero-shot

Annotations: Accuracy: 5; Completeness: 3.5; Readability: 5

the study investigates the effect of two chemically distinct denture cleansers on the surface hardness of acrylic and silicone based soft denture liners. soft denture liners are used to make wearing dentures more comfortable for patients, but can present problems such as weakening of bond between lining and denture, loss of resiliency, and increased surface roughness. denture cleansers are essential for maintaining oral hygiene, with peroxide cleansers being the most commonly used. the study found that after one month, significant changes were noted in surface hardness for acrylic samples when compared to those of sodium hypochlorite group and sodium perborate group. at six months, silicone-based soft denture liners showed better compatibility with cleansing solutions and maintained their resiliency better, proving to be more promising for long term usage. the study concludes that chemical denture cleansers can be used daily to cleanse soft denture liners without adversely affecting their surface hardness.

1.5 LLaMA 2 13B zero-shot

Annotations: Accuracy: 3.5; Completeness: 2; Readability: 1.5

acrylic - based soft denture liners at all time intervals . for silicone - based soft denture liners highly significant increase in surface hardness was seen at 1 month and 3 months . in sodium perborate group , highly significant increase in surface hardness was seen for acrylic - based soft denture liners at all time intervals . for silicone - based soft denture liners highly significant increase in surface hardness was seen at 1 month and 3 months . the results of the present study indicate that silicone - based soft denture liners are more resistant to the effects of denture cleansers than acrylic - based soft denture liners . the results of the present study are in agreement with the findings of mese and guzel5 who concluded that silicone - based soft denture liners are more resistant to the effects of denture cleansers than acrylic - based soft denture liners . the results of the present study are also in agreement with the findings of sara et al.14 who concluded that sodium perborate denture cleanser proved to be most effective . the results of the present study are also in agreement with the findings of de freitas fernandes et al.15 and ferreira et al.16 who concluded that the best results were found for the treatment with 0.5% naocl for 10 minutes in comparison to other cleansers . the results of the present study are also in agreement with the findings of mese and guzel5 who concluded that after 6 months hardness values of all resilient liners evaluated were higher with increased duration of immersion . the hardness values of acrylic resin - based liners showed greater change than those of silicone products . the results of the present study are also in agreement with the findings of mese and guzel5 who concluded that the hardness values of acrylic resin - based liners showed greater change than those of silicene products . the results of the present study are also in agreement with the findings of mese and guzel5 who concluded that the hardness values of acrylic resin - based liners showed greater change than those of silicene products . the results of the present study are also in agreement with the findings of mese and guzel5 who concluded that

Example 2

2.1 Gold standard

an inflammatory myofibroblastic tumor (imt) is a distinctive neoplasm composed of myofibroblastic and fibroblastic spindle cells , accompanied by inflammatory infiltration of plasma cells , lymphocytes , and eosinophils .

imts rarely occur in the urinary bladder .

it is important to distinguish this tumor from other malignant spindle cell tumors .

herein , we report a patient with an imt showing muscle invasion , who underwent a transurethral resection of the bladder tumor and , at a later date , partial cystectomy . the resected tumor specimen revealed a proliferation of spindle - shaped cells on a background of plasma cells and lymphocytes .

immunohistochemical staining showed the tumor to be positive for anaplastic lymphoma kinase (alk) , smooth muscle actin , and vascular endothelial growth factor (vegf) .

such histopathological findings were indicative of an imt , suggesting the use of inhibitors of alk and vegf as pharmacotherapy .

2.2 Fine-tuned BART

Annotations: Accuracy: 5; Completeness: 4.5; Readability: 4.5

an inflammatory myofibroblastic tumor (imt) is a distinctive neoplasm composed of myofibroblastic and fibroblastic spindle cells accompanied by inflammatory infiltration of plasma cells , lymphocytes , and eosinophils .

it is important to distinguish this tumor from other malignant spindle cell tumors , such as the sarcomatoid variant of urothelial carcinoma and leiomyosarcoma .

we report a case of imt of the urinary bladder in a 52-year - old male diagnosed by transurethral resection of the bladder tumor (turbt) treated by partial cystectomy .

histopathology results were consistent with an imt .

the bladder tissue , including muscle , was widely infiltrated by spindle cells in a myxoid stroma accompanied by infiltration of inflammatory cells .

immunohistochemical staining was positive for vimentin and smooth muscle actin and negative for ae1/ae3 , desmin , myogenin , s-100 , cd34 , c - kit , cd68 , and bcl-2 .

a pathological examination of the excised specimen revealed the proliferation of spindle cells expanding into the bladder musculature , accompanied by inflammatory cell infiltration .

for the assessment of therapeutic options , we measured the expressions of anaplastic lymphoma kinase (alk) , vascular endothelial growth factor (vegf) , and cyclooxygenase 2 (cox2) , known as growth factors involved in tumor proliferation .

2.3 GPT-3.5 zero-shot

Annotations: Accuracy: 5; Completeness: 3; Readability: 4.5

this case report describes a 52-year-old male with an inflammatory myofibroblastic tumor (imt) of the urinary bladder. the patient had a history of rheumatoid arthritis and was treated with prednisolone. imaging studies revealed a bladder tumor infiltrating the muscle layer. the tumor was resected by partial cystectomy, and histopathology confirmed the diagnosis of imt. immunohistochemical staining showed positive expression of vimentin and smooth muscle actin, and negative expression of other markers. the expression of growth factors involved in tumor proliferation, such as anaplastic lymphoma kinase (alk), vascular endothelial growth factor (vegf), and cyclooxygenase 2 (cox2), was also measured. surgical resection is the main treatment for imt, but pharmacotherapy may be considered for nonresectable tumors. close follow-up is recommended for imt cases.

2.4 GPT-4 zero-shot

Annotations: Accuracy: 5; Completeness: 3.5; Readability: 4.5

an inflammatory myofibroblastic tumor (imt) is a unique neoplasm composed of myofibroblastic and fibroblastic spindle cells, often accompanied by inflammatory infiltration. a case of imt in the urinary bladder of a 52-year-old male was diagnosed and treated by partial cystectomy. the tumor was characterized by atypical spindle cell proliferation and inflammatory cell infiltrates, primarily involving lymphocytes and plasma cells. imts are classified as intermediate tumors and are often found in the mesentery, omentum, retroperitoneum, pelvis, and abdominal soft tissues. however, occurrence in the urinary bladder is unusual. in terms of treatment, surgical resection is often used, but pharmacotherapy may become a viable therapeutic option in cases of recurrence when the tumor is not resectable.

2.5 LLaMA 13B zero-shot

Annotations: Accuracy: 5; Completeness: 3; Readability: 3

an inflammatory myofibroblastic tumor (imt) is a distinctive neoplasm composed of myofibroblastic and fibroblastic spindle cells accompanied by inflammatory infiltration of plasma cells , lymphocytes , and eosinophils . it is important to distinguish this tumor from other malignant spindle cell tumors , such as the sarcomatoid variant of urothelial carcinoma and leiomyosarcoma . we report a case of imt of the urinary bladder in a 52 - year - old male diagnosed by transurethral resection of the bladder tumor (turbt) treated by partial cystectomy . his medical history revealed a diagnosis of rheumatoid arthritis 3 years ago , and prednisolone 10 mg / day was subsequently prescribed . cystoscopy revealed a solitary nonpapillary tumor with surrounding edema at the dome of the bladder . enhanced computed tomography and magnetic resonance imaging revealed an early enhancing bladder tumor infiltrating the outside of the bladder muscle layer . intraoperative findings showed a large solid bladder mass , measuring approximately 3 cm from the dome to the anterior wall of the bladder . we resected the tumor to the depth of the muscle layer , but normal muscle tissue was not seen . histopathology results were consistent with an imt . the bladder tissue , including muscle , was widely infiltrated by spindle cells in a myxoid stroma accompanied by infiltration of inflammatory cells . immunohistochemical staining was positive for vimentin and smooth muscle actin and negative for ae1/ae3 , desmin , myogenin , s - 100 , cd34 , c - kit , cd68 , and bcl - 2 . the excised specimen is shown in figure 2d . a pathological examination of the excised specimen revealed the proliferation of spindle cells expanding into the bladder musculature , accompanied by inflammatory cell infiltration . for

Example 3

3.1 Gold standard

backgroundit is recognized that candida dubliniensis commonly colonizes oral and subgingival sites in immunocompetent subjects with periodontal disease.objectivesince there are few data available on genetic characterization of c. dubliniensis in periodontal pockets and other oral sites , the aim of this study was to characterize subgingival and mucosal c. dubliniensis isolates recovered from immunocompetent subjects and to assay the genetic similarity of such isolates from both niches in the same patient by random amplified polymorphic dna (rapd).design

c. dubliniensis recovered from subgingival plaque and from buccal cavity samples were studied in 240 immunocompetent non - smoking individuals . arbitrary amplification was carried out by rapd - polymerase chain reaction (pcr).resultsrapd analysis showed identical genotypes of c. dubliniensis in different sampling sites (buccal cavity and subgingival areas) in eight of 10 patients except for those derived from two participants who presented presumably unrelated isolates.conclusionson the basis of the findings presented , the origin of the colonization of c. dubliniensis in subgingival biofilm seems to be the buccal cavity in a single patient .

consequently , it may be assumed that most of c. dubliniensis in these sites arise from the endogenous commensal strains .

3.2 Fine-tuned BART

Annotations: Accuracy: 5; Completeness: 4; Readability: 4.5

background : candida dubliniensis is a yeast species that is characterized by its in vitro resistance to the antifungal azole group in hiv patients , for their capacity to adhere in vitro to human buccal epithelial cells and other microorganisms in the oral microbiota , as well as the high proteinase activity.objectives:the aim of this study was to determine the prevalence of c. dubliniensis in the buccal cavity and subgingival biofilm in immunocompetent individuals.materials and methods : a total of 240 individuals were included in the study .

clinical periodontal evaluations included clinical examination and radiographs with clinical measurements : clinical attachment level (cal) , pocket depth (pd) , plaque index (29) , gingival index (30) , and bleeding on probing (bop) .

yeasts were isolated by conventional mycological methods : colony color on the chromogenic medium , micromorphology in agar milk with 1% tween 80 , and carbohydrate assimilation tests using the commercially available kit api i d 32c .

dendrogram was generated by the unweighted pair group method using arithmetic averages (upgma).results : out of the 203 yeasts , c. dubliniensis was isolated in 4.2% (10/240) and 4.6% (11/240) of patients in both niches , respectively .

eight genetic clusters and five main genotypes were obtained at an sc of 90% , genotypes i , ii , iii , iv , and v.conclusions:c. dubliniensis can colonize subgingival biofilm in both healthy and immunocompetent individuals .

3.3 GPT-3.5 zero-shot

Annotations: Accuracy: 5; Completeness: 4.5; Readability: 5

a study was conducted to investigate the presence and distribution of candida dubliniensis in the buccal cavity and subgingival biofilm of individuals with periodontal disease. a total of 21 isolates of c. dubliniensis were collected from 240 non-smoking individuals. clinical evaluations were performed, and measurements of periodontal parameters were taken. the study found that c. dubliniensis was present in both the buccal cavity and subgingival biofilm, with a prevalence of 4.2% and 4.6% respectively. the species distribution of yeast isolates varied according to the periodontal health status. rapd analysis showed that c. dubliniensis isolates had similar genotypes in different sampling sites, suggesting a common clonal origin. these findings provide baseline information on the colonization of c. dubliniensis in the oral cavity.

3.4 GPT-4 zero-shot

Annotations: Accuracy: 5; Completeness: 4; Readability: 4.5

the study collected 21 isolates of c. dubliniensis from 240 non-smoking individuals at the university of buenos aires dental clinic. the samples were taken from the buccal cavity and subgingival biofilm, and were cultured on a differential chromogenic medium. the isolated yeasts were identified using conventional mycological methods and further characterized using specific pcr with primers from actin gene-associated intron sequences of c. dubliniensis. the study also used rapd-pcr analysis for genetic characterization of the isolates. the results showed that c. dubliniensis was isolated in 4.2% and 4.6% of patients in the buccal cavity and subgingival biofilm respectively. the study concluded that the origin of the colonization of c. dubliniensis in subgingival biofilm seems to be the buccal cavity.

3.5 LLaMA 2 13B zero-shot

Annotations: Accuracy: 1.5; Completeness: 1; Readability: 1

. 1) . the other case was patient 10 , who presented two genotypes in the buccal cavity and three in the subgingival biofilm . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dubliniensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and subgingival biofilm were identical in 80% of the patients . the genotypes of c. dublininensis in the buccal cavity and sub