A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks

Yanis Labrak^{1,3} Mickael Rouvier¹ Richard Dufour²

(1) LIA, Avignon Université (2) LS2N, Nantes Université (3) Zenidoc











Diversified Benchmark

Task	Dataset	Eval	Metric	Reference
	HoC	Test	F1	Baker et al. (2016)
CLS	LitCovid	Test	F1	Chen et al. (2021)
	PubHealth	Test	Accuracy	Neema and Toni (2020)
	N2C2 2006 Smokers	Test	Accuracy	Uzuner et al. (2008)
	BioASQ 7b	Test	Accuracy	Tsatsaronis et al. (2015)
04	MedMCQA	Dev	Accuracy	Pal et al. (2022)
QA	SciQ	Test	Accuracy	Welbl et al. (2017)
	Evidence Inference 2.0	Test	Accuracy	DeYoung et al. (2020)
RE	GAD	Test	Accuracy	Bravo et al. (2015)
NI I	SciTail	Test	Accuracy	Khot et al. (2018)
INLI	MedNLI	Test	Accuracy	Shivade (2017)
NER	BC5CDR	Test	F1	Li et al. (2016)
	NCBI-disease	Test	F1	Dogan et al. (2014)



Generative:









VS.





Flan-T5

- Tk-Instruct
- GPT 3.5

Alpaca







Instruction Format

Instruction: You are given a science question (easy level) and four answer options (associated with "A", "B", "C", "D"). Your task is to find the correct answer based on scientific facts, knowledge and reasoning. Don't generate anything other than one of the following characters: 'A B C D'.

Input: Heavy forces on periodontal ligament causes: (A) Hyalinization (B) Osteoclastic activity around tooth (C) Osteoblastic activity around tooth (D) Crest bone resorption

Constraints: The answer must be one or more of the following letters: 'A', 'B', 'C', 'D'. You must generate one and only one letter for each question. All questions have an answer. No justification is required.

Output:

11

Instruction Format - Few-shot

1

Vector Space	Instruction: You are a highly intelligent and accurate healthcare domain Named-entity recognition (NER) system. You are tasked to do Named-entity recognition (NER) for 'disease' and 'none' only, please generate the appropriate label. Constraints: You can choose only one label from: none or disease.
	Examples: Example 1: <example-1> Example 2: <example-2> Example 3: <example-3> Example 4: <example-4> Example 5: <example-5> Considering the sentence : Clustering of missense mutations in the ataxia - telangiectasia gene in a sporadic T - cell leukaemia . And considering your precedents predictions: Clustering none of none missense none mutations none in none the none Input : The label of « leukaemia » at the position 17 of the sentence is ?</example-5></example-4></example-3></example-2></example-1>

Avignon Université

Instruction Format - Few-shot

Instruction: You are given a science question (easy level) and four answer options (associated with "A", "B", "C", "D"). Your task is to find the correct answer based on scientific facts, knowledge and reasoning. Don't generate anything other than one of the following characters: 'A B C D'.

Constraints: The answer must be one or more of the following letters: 'A', 'B', 'C', 'D'. You must generate one and only one letter for each question. All questions have an answer. No justification is required.

Examples:

Example 1: Hyalinisation of the periodontal Ligament, due to excessive orthodontic forces results in (A) Frontal resorption (B) Undermining resorption (C) Cementum remaining intact (D) Dentine remaining intact Output: B

Example 2: The earliest response of pulpitis is: (A) Cyst formation (B) Calcification (C) Hyalinization (D) Formation of dental granuloma Output: C

Example 3: Among the secondary changes in tooth the most useful one for age determination is: (A) Attrition (B) Secondary dentine deposition (C) Root resorption (D) Root transparency Output: D

Example 4: Feature of aging periodontium is (A) Lacunae in bone and cementum (B) Increased cell size (C) Increased cell number (D) Scalloping of cementum & alveolar bone surface Output: D

Example 5: Bacteria found in gingivitis are localized in (A) Connective tissue fibres (B) Gingival sulcus (C) Alveolar bone (D) Periodontal ligament Output: B

Input: Heavy forces on periodontal ligament causes: (A) Hyalinization (B) Osteoclastic activity around tooth (C) Osteoblastic activity around tooth (D) Crest bone resorption Output:



Recursive Chain-of-Thought

Non-working example using normal prompting:

Instruction: You are a highly intelligent and accurate healthcare domain Named-entity recognition (NER) system. You are tasked to do Named-entity recognition (NER) for 'disease' and 'none' only, please generate the appropriate labels for each words using pipe separators.

Considering the sentence : Clustering of missense mutations in the ataxia - telangiectasia gene in a sporadic T - cell leukaemia .

Output:

Clustering|none|of|none|missense|disease|mutations|disease|in|none|the|none|ataxia|disease|-|none|telangiectasia|disease|gene|none|in|none|a|none|sporadic|none|T|non e|-|none|cell|none|leukaemia|none|.|none|

Labelize all the specialized terms as "disease" labels.



Recursive Chain-of-Thought

nstruction: You are a highly intelligent and accurate healthcare domain Named-entity recognition (NER) system. You are tasked to do Named-entity recognition (NER) for 'disease' and	
none' only, please generate the appropriate label.	

Constraints: You can choose only one label from: none or disease.

Examples:

Example 1 : <example-1>

[...]

Considering the sentence : Clustering of missense mutations in the ataxia - telangiectasia gene in a sporadic T - cell leukaemia .

And considering your precedents predictions: Clustering|none||of|none||missense|none||mutations|none||in|none

Input : The label of « the » at the position 6 of the sentence is ?

Output: None



Recursive Chain-of-Thought

Instruction: You are a highly intelligent and accurate healthcare domain Named-entity recognition (NER) system	. You are tasked to do Named-entity recognition (NER) for 'disease' and
'none' only, please generate the appropriate label.	

Constraints: You can choose only one label from: none or disease.

Examples:

Example 1 : <example-1>

[...]

Considering the sentence : Clustering of missense mutations in the ataxia - telangiectasia gene in a sporadic T - cell leukaemia .

And considering your precedents predictions: Clustering|none||of|none||missense|none||mutations|none||in|none||the|none

Input : The label of « ataxia » at the position 7 of the sentence is ?

Output: Disease



Protocol

Common issue with free text generations:

Instruction: You are given a science question (easy level) and four answer options (associated with "A", "B", "C", "D"). Your task is to find the correct answer based on scientific facts, knowledge and reasoning. Don't generate anything other than one of the following characters: 'A B C D'.

Input: Heavy forces on periodontal ligament causes: (A) Hyalinization (B) Osteoclastic activity around tooth (C) Osteoblastic activity around tooth (D) Crest bone resorption

Constraints: The answer must be one or more of the following letters: 'A','B','C','D'. You must generate one and only one letter for each question. All questions have an answer. No justification is required.

Output: Heavy forces on the periodontal ligament can lead to increased osteoclastic activity around the tooth. This is because excessive mechanical stress on the ligament triggers a response from the body to remodel the surrounding bone tissue, resulting in osteoclastic activity, which breaks down bone tissue. Therefore, the correct answer is (**B**) Osteoclastic activity around tooth.

Noisy output which needs manual evaluation or Regex.

Goal is to compute a comparable metric with BERT-based models.



Zero-Shot Results



Avignon Université

Tack	Datasat	ChatG	iPT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	
Idan	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIMEUDERI
	HoC	62.24		56.36		50.77		1.21		82.75
019	LitCovid	67.20		51.48		36.42		1.58		90.60
OLS	PubHealth	63.20		72.46		53.70		52.80		75.39
	N2C2 2006 Smokers	NaN		22.12		16.35		10.57		60.58
	BioASQ 7b	89.24		90.97	_	88.09		79.05		73.39
04	MedMCQA	48.91		41.05		33.85		24.91		38.15
QA	SciQ	90.10		87.00		55.30		24.90		74.20
	Evidence Inference 2.0	59.98		66.45		41.33		32.49		65.47
RE	GAD	47.75		49.81	_	48.88		51.12		79.78
NLI	SciTail	73.57		93.51	_	57.53		39.60		93.51
	MedNLI	NaN		77.00		33.19		33.47		83.76
NER	BC5CDR	92.12		68.26	_	84.54		82.11		97.65
	NCBI-disease	90.97		90.75		87.91		11.58		98.72



Tack	Datasat	ChatG	iPT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	
Idan	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIMEUDERI
	HoC	62.24		56.36		50.77		1.21		82.75
019	LitCovid	67.20		51.48		36.42		1.58		90.60
OLS	PubHealth	63.20		72.46		53.70		52.80		75.39
	N2C2 2006 Smokers	NaN		22.12		16.35		10.57		60.58
	BioASQ 7b	89.24		90.97	_	88.09		79.05		73.39
04	MedMCQA	48.91		41.05		33.85		24.91		38.15
	SciQ	90.10		87.00		55.30		24.90		74.20
	Evidence Inference 2.0	59.98		66.45		41.33		32.49		65.47
RE	GAD	47.75		49.81	_	48.88		51.12		79.78
	SciTail	73.57		93.51	_	57.53		39.60		93.51
	MedNLI	NaN		77.00		33.19		33.47		83.76
NEP	BC5CDR	92.12		68.26	_	84.54		82.11		97.65
	NCBI-disease	90.97		90.75	_	87.91		11.58		98.72



Tack	Datacat	ChatG	iPT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	
Task	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIWIEUDENI
	HoC	62.24		56.36		50.77		1.21		82.75
019	LitCovid	67.20		51.48		36.42		1.58		90.60
	PubHealth	63.20		72.46		53.70		52.80		75.39
	N2C2 2006 Smokers	NaN		22.12		16.35		10.57		60.58
	BioASQ 7b	89.24		90.97	_	88.09		79.05		73.39
04	MedMCQA	48.91		41.05		33.85		24.91		38.15
	SciQ	90.10		87.00		55.30		24.90		74.20
	Evidence Inference 2.0	59.98		66.45		41.33		32.49		65.47
RE	GAD	47.75		49.81	_	48.88		51.12		79.78
NILI	SciTail	73.57		93.51	_	57.53		39.60	. –	93.51
	MedNLI	NaN		77.00		33.19		33.47		83.76
NER	BC5CDR	92.12		68.26	_	84.54		82.11	_	97.65
	NCBI-disease	90.97		90.75		87.91		11.58		98.72



Tack	Datasat	ChatG	PT	Flan-U	JL2	Tk-Inst	ruct	Alpa	ca	PubMedBERT
Idan	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIMEUDENI
	HoC	62.24		56.36		50.77		1.21		82.75
019	LitCovid	67.20		51.48		36.42		1.58		90.60
OLS	PubHealth	63.20		72.46		53.70		52.80		75.39
	N2C2 2006 Smokers	NaN		22.12		16.35		10.57		60.58
	BioASQ 7b	89.24		90.97	_	88.09		79.05		73.39
04	MedMCQA	48.91		41.05		33.85		24.91		38.15
	SciQ	90.10		87.00		55.30		24.90		74.20
	Evidence Inference 2.0	59.98		66.45		41.33		32.49		65.47
RE	GAD	47.75		49.81	_	48.88		51.12		79.78
NLI	SciTail	73.57		93.51	_	57.53		39.60		93.51
	MedNLI	NaN		77.00		33.19		33.47		83.76
NER	BC5CDR	92.12		68.26	_	84.54		82.11		97.65
	NCBI-disease	90.97	_	90.75		87.91		11.58		98.72



Tack	Datasat	ChatG	PT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	
Idan	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIMEUDERI
	HoC	62.24		56.36		50.77		1.21		82.75
019	LitCovid	67.20		51.48		36.42		1.58		90.60
OLS	PubHealth	63.20		72.46		53.70		52.80		75.39
	N2C2 2006 Smokers	NaN		22.12		16.35		10.57		60.58
	BioASQ 7b	89.24		90.97	_	88.09		79.05		73.39
04	MedMCQA	48.91		41.05		33.85		24.91		38.15
	SciQ	90.10		87.00		55.30		24.90		74.20
	Evidence Inference 2.0	59.98		66.45		41.33		32.49		65.47
RE	GAD	47.75		49.81	_	48.88		51.12		79.78
	SciTail	73.57		93.51	_	57.53		39.60		93.51
	MedNLI	NaN		77.00		33.19		33.47		83.76
NER	BC5CDR	92.12		68.26	_	84.54		82.11		97.65
	NCBI-disease	90.97		90.75		87.91		11.58		98.72



Tack	Datasat	ChatG	PT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	
Iask	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIMEUDERI
	HoC	62.24		56.36		50.77		1.21		82.75
019	LitCovid	67.20		51.48		36.42		1.58		90.60
OLS	PubHealth	63.20		72.46		53.70		52.80		75.39
	N2C2 2006 Smokers	NaN		22.12		16.35		10.57		60.58
	BioASQ 7b	89.24		90.97	_	88.09		79.05		73.39
04	MedMCQA	48.91		41.05		33.85		24.91		38.15
QA	SciQ	90.10		87.00		55.30		24.90		74.20
	Evidence Inference 2.0	59.98		66.45		41.33		32.49		65.47
RE	GAD	47.75		49.81	_	48.88		51.12		79.78
	SciTail	73.57		93.51	_	57.53		39.60		93.51
	MedNLI	NaN		77.00		33.19		33.47		83.76
NER	BC5CDR	92.12		68.26	_	84.54		82.11		97.65
	NCBI-disease	90.97	_	90.75		87.91		11.58		98.72







Avignon Université

Tack	Dataset	ChatG	iPT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	
Idan	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDMEUDERI
	HoC	62.24	38.34	56.36	54.86	50.77	25.48	1.21	38.78	82.75
019	LitCovid	67.20	72.77	51.48	46.95	36.42	57.49	1.58	64.09	90.60
	PubHealth	63.20	66.29	72.46	50.53	53.70	66.04	52.80	55.64	75.39
	N2C2 2006 Smokers	NaN	NaN	22.12	42.31	16.35	37.50	10.57	31.73	60.58
	BioASQ 7b	89.24	92.03	90.97	91.64	88.09	86.36	79.05	79.82	73.39
	MedMCQA	48.91	56.37	41.05	43.34	33.85	33.18	24.91	29.50	38.15
	SciQ	90.10	93.50	87.00	88.40	55.30	47.00	24.90	36.80	74.20
	Evidence Inference 2.0	59.98	63.83	66.45	65.06	41.33	38.79	32.49	94.18	65.47
RE	GAD	47.75	52.25	49.81	53.37	48.88	57.87	51.12	57.68	79.78
NILI	SciTail	73.57	65.62	93.51	92.66	57.53	71.31	39.60	40.26	93.51
	MedNLI	NaN	NaN	77.00	<u>79.18</u>	33.19	34.81	33.47	34.45	83.76
NED	BC5CDR	92.12	93.12	68.26	83.32	84.54	83.23	82.11	84.07	97.65
	NCBI-disease	90.97	92.27	90.75	87.65	87.91	87.50	11.58	92.27	98.72



Tack	Datacat	ChatG	PT	Flan-L	JL2	Tk-Inst	ruct	Alpa	ca	PubMedBERT
Idan	Dalasel	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDWEUDERT
	HoC	62.24	38.34	56.36	54.86	50.77	25.48	1.21	38.78	82.75
212	LitCovid	67.20	72.77	51.48	46.95	36.42	57.49	1.58	64.09	90.60
	PubHealth	63.20	66.29	72.46	50.53	53.70	66.04	52.80	55.64	75.39
	N2C2 2006 Smokers	NaN	NaN	22.12	42.31	16.35	37.50	10.57	31.73	60.58
	BioASQ 7b	89.24	92.03	90.97	91.64	88.09	86.36	79.05	79.82	73.39
	MedMCQA	48.91	56.37	41.05	43.34	33.85	33.18	24.91	29.50	38.15
	SciQ	90.10	93.50	87.00	88.40	55.30	47.00	24.90	36.80	74.20
	Evidence Inference 2.0	59.98	63.83	66.45	65.06	41.33	38.79	32.49	94.18	65.47
RE	GAD	47.75	52.25	49.81	53.37	48.88	57.87	51.12	57.68	79.78
NU	SciTail	73.57	65.62	93.51	92.66	57.53	71.31	39.60	40.26	93.51
	MedNLI	NaN	NaN	77.00	<u>79.18</u>	33.19	34.81	33.47	34.45	83.76
NER	BC5CDR	92.12	93.12	68.26	83.32	84.54	83.23	82.11	84.07	97.65
NER	NCBI-disease	90.97	92.27	90.75	87.65	87.91	87.50	11.58	<u>92.27</u>	98.72



Task	Dataset	ChatGPT		Flan-UL2		Tk-Instruct		Alpaca		
		zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIVICUDENT
CLS	HoC	62.24	38.34	56.36	54.86	50.77	25.48	1.21	38.78	82.75
	LitCovid	67.20	72.77	51.48	46.95	36.42	57.49	1.58	64.09	90.60
	PubHealth	63.20	66.29	72.46	50.53	53.70	66.04	52.80	55.64	75.39
	N2C2 2006 Smokers	NaN	NaN	22.12	42.31	16.35	37.50	10.57	31.73	60.58
QA	BioASQ 7b	89.24	92.03	90.97	91.64	88.09	86.36	79.05	79.82	73.39
	MedMCQA	48.91	56.37	41.05	43.34	33.85	33.18	24.91	29.50	38.15
	SciQ	<u>90.10</u>	93.50	87.00	88.40	55.30	47.00	24.90	36.80	74.20
	Evidence Inference 2.0	59.98	63.83	66.45	65.06	41.33	38.79	32.49	94.18	65.47
RE	GAD	47.75	52.25	49.81	53.37	48.88	57.87	51.12	57.68	79.78
NLI	SciTail	73.57	65.62	93.51	92.66	57.53	71.31	39.60	40.26	93.51
	MedNLI	NaN	NaN	77.00	79.18	33.19	34.81	33.47	34.45	83.76
NER	BC5CDR	92.12	93.12	68.26	83.32	84.54	83.23	82.11	84.07	97.65
	NCBI-disease	90.97	92.27	90.75	87.65	87.91	87.50	11.58	92.27	98.72



Task	Dataset	ChatGPT		Flan-UL2		Tk-Instruct		Alpaca		
		zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	zero-shot	5-shot	FUDIVICUDENT
CLS	HoC	62.24	38.34	56.36	54.86	50.77	25.48	1.21	38.78	82.75
	LitCovid	67.20	72.77	51.48	46.95	36.42	57.49	1.58	64.09	90.60
	PubHealth	63.20	66.29	72.46	50.53	53.70	66.04	52.80	55.64	75.39
	N2C2 2006 Smokers	NaN	NaN	22.12	42.31	16.35	37.50	10.57	31.73	60.58
QA	BioASQ 7b	89.24	92.03	90.97	<u>91.64</u>	88.09	86.36	79.05	79.82	73.39
	MedMCQA	48.91	56.37	41.05	43.34	33.85	33.18	24.91	29.50	38.15
	SciQ	90.10	93.50	87.00	88.40	55.30	47.00	24.90	36.80	74.20
	Evidence Inference 2.0	59.98	63.83	66.45	65.06	41.33	38.79	32.49	94.18	65.47
RE	GAD	47.75	52.25	49.81	53.37	48.88	57.87	51.12	57.68	79.78
NLI	SciTail	73.57	65.62	93.51	92.66	57.53	71.31	39.60	40.26	93.51
	MedNLI	NaN	NaN	77.00	<u>79.18</u>	33.19	34.81	33.47	34.45	83.76
NER	BC5CDR	92.12	93.12	68.26	83.32	84.54	83.23	82.11	84.07	97.65
	NCBI-disease	90.97	92.27	90.75	87.65	87.91	87.50	11.58	92.27	98.72



