

ellis

European Laboratory for Learning and Intelligent Systems

NLP FOR HEALTH

From clinical notes to mental health and large language models

Shaoxiong Ji | ELLIS Institute Finland & University of Turku | October 17, 2026

olaresearch.org/jis

OUTLINE

01 Introduction

02 Medical Text Understanding

03 Mental Health

04 Large Language Models

05 Lessons and Outlook

01

INTRODUCTION

Why language matters in health



WHERE HEALTH LANGUAGE LIVES

Clinical records

Discharge summaries, radiology and pathology reports, operative notes. Long, terse, full of abbreviations: *“a [xx] y/o M w/ Hep C, HTN”*.

Patient voice

Social media, forums and patient reviews: drug side effects, symptoms, distress. Informal, noisy, self-labelled.

Conversations

Counselling sessions, doctor–patient dialogue, interactions with EHR systems. Multi-turn, sensitive, rarely recorded.

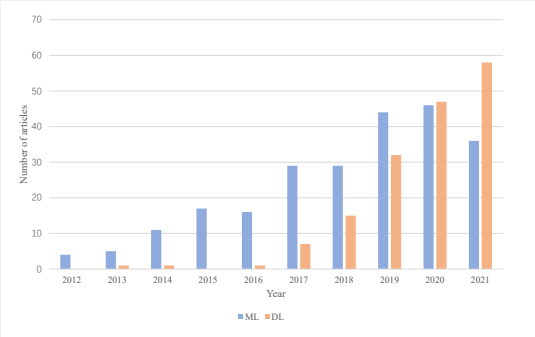
Takeaway. Most clinically relevant information is written in free text, so structured data alone misses much of the picture.

TASKS COVERED IN THIS LECTURE

Task	**Input → output**	**Formulation**
Clinical coding	discharge summary → set of ICD codes	multi-label classification
Outcome prediction	clinical notes → readmission, diagnosis, length of stay	multitask learning
Clinical NER	clinical sentences → entity spans	sequence labelling
Entity linking	mention → UMLS / SNOMED CT concept	similarity ranking
ADE detection	tweet / forum post → adverse drug event?	document classification
Mental health detection	posts → depression, suicide risk, . . .	(ordinal) classification
Supportive dialogue	conversation → defenses, therapist responses	classification, generation

HOW THE METHODS EVOLVED

1. **Rules and lexicons** (since ∼1970): regex, dictionaries, sentiment lexicons. Precise, low recall.
2. **Feature-based ML**: SVM, logistic regression on n-grams, LIWC, topics.
3. **Deep learning** (∼2017+): CNN/RNN encoders with attention.
4. **Domain pretrained LMs**: ClinicalBERT, BioBERT, MentalBERT.
5. **Large language models**: prompting, instruction tuning, agents.



Mental illness detection papers per year: deep learning overtakes classical ML around 2020.

Zhang, T., Schoene, A. M., Ji, S., and Ananiadou, S. (2022). "Natural language processing applied to mental illness detection: a narrative review". *NPJ digital medicine* 5.1, p. 46

WHY HEALTHCARE NLP IS HARD

Data

- Long documents with jargon, abbreviations and copy-paste “note bloat”
- Privacy (HIPAA, GDPR): little shareable data, few annotations
- Labels are noisy: under-coding, self-reported diagnoses, subreddit membership

Stakes

- Huge, long-tailed label spaces (ICD-10-CM: $\sim$68,000 codes)
- Errors have a human cost: a missed suicide risk is not a missed movie recommendation
- Clinicians need explanations they can check

HEALTH TEXT IS PERSONAL DATA

GDPR in practice

- Text with medical information is personal data, even when pseudonymised and even when posted publicly
- **Pseudonymisation**: replace each user by an ID; keeps longitudinal analysis possible
- **Anonymisation**: remove all links between texts of the same user; harder for text than for numbers
- Collect only what you need, do not re-share, delete at the end of the project, get ethics approval

Example: peer support forums (iPOF)

- People post distress on mental health forums hoping for help; research must not erode that sense of safety
- Ethics framework built with legal, clinical, academic and lived-experience input; approved by the UK Health Research Authority
- Data protection impact assessment with the NHS host trust

Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617; Lobban, F. et al. (2025). "Evaluating peer online forums to support health: ethical and practical challenges". *J Med Internet Res* 27.1, e73427

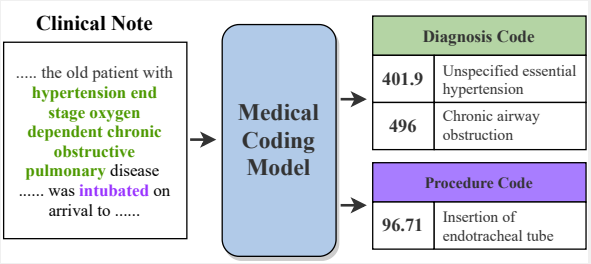
02

MEDICAL TEXT UNDERSTANDING

Coding, outcome prediction, entity recognition, drug safety



WHAT IS CLINICAL CODING?



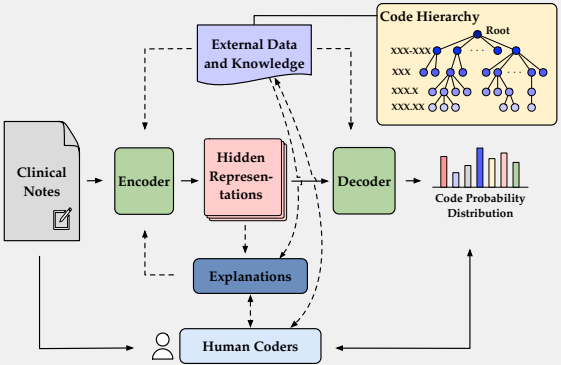
Manual coding today

- ∼60 cases per coder per day (NHS Scotland)
- Backlogs of months to over a year
- UK accuracy ∼83% on average (range 50–98%)

Map a free-text record to codes from a classification system (ICD-9/10/11, OPCS, CCS) used for statistics, planning, research and billing.

Dong, H. et al. (2022). “Automated clinical coding: what, why, and where we are?” *NPJ digital medicine* 5.1, p. 159; Ji, S. et al. (2024). “A unified review of deep learning for automated medical coding”. *ACM computing surveys* 56.12, pp. 1–41

A UNIFIED VIEW OF CODERS



Encoder CNN, RNN, Transformer, graph or hierarchical

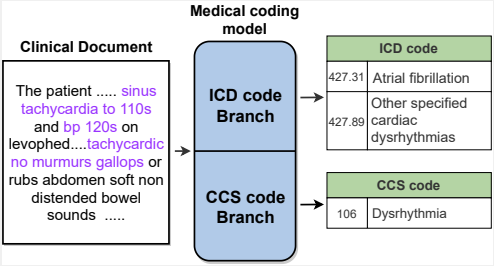
Deep connections stacking, residual, embedding injection

Decoder label-wise attention (dominant), hierarchical, multitask, zero-shot, generative

Auxiliary knowledge code descriptions, ICD hierarchy, UMLS, Wikipedia, human coders

Ji, S. et al. (2024). "A unified review of deep learning for automated medical coding". *ACM computing surveys* 56.12, pp. 1–41

MULTITASK CODING: ICD + CCS



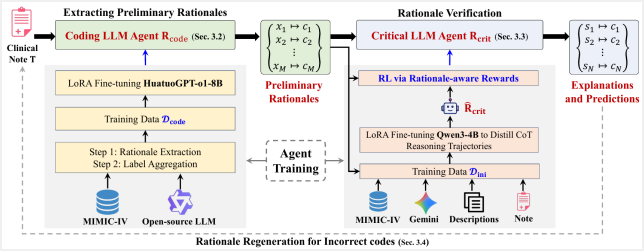
CCS groups encode ICD relations for free (427.31, 427.89 $\rightarrow$ CCS 106).

- **Recalibrated aggregation**: context-aware reweighting of noisy features
- **Joint loss** $0.7\,\mathcal{L}_{\text{ICD}} + 0.3\,\mathcal{L}_{\text{CCS}}$
- **Focal loss** up-weights rare codes

MIMIC-III full	Macro-F1	Micro-F1
MARN	**11.6**	**58.4**
w/o MTL, RAM, focal	6.8	51.5
MultiResCNN	8.5	55.2

Sun, W., Ji, S., Cambria, E., and Marttinen, P. (2021). "Multitask recalibrated aggregation network for medical code prediction". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 367–383; Sun, W., Ji, S., Cambria, E., and Marttinen, P. (2022). "Multitask balanced and recalibrated network for medical code prediction". *ACM Transactions on Intelligent Systems and Technology* 14.1, pp. 1–20

CODING WITH LLM AGENTS



- **Coding agent** (8B, LoRA): a rationale per code
- **Critical agent** (4B): verifies each rationale; RL reward is lower for a right code with fabricated evidence
- Runs on two local GPUs: notes stay in the hospital

MIMIC-IV, ICD-10

F1 (1,000 notes)	Macro	Micro
ICDAgent	**24.3**	46.7
GPT-5-Chat	21.5	39.6
Gemini-2.5-Pro	19.8	35.6
CoRelation (classifier)	15.2	**50.0**

- Frontier APIs code cautiously: low recall
- Classifiers win on frequent codes (micro), the agent on rare codes (macro), and it explains itself

Yin, Z., Cao, Y., Wang, T., Chen, J., and Ma, F. (2026). "ICDAGENT: Empowering Agentic Large Language Models for Explainable Medical Coding". *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14131–14149

WHERE CLINICAL CODING STANDS

34%
GPT-4 exact match, ICD-10-CM lookup

24%
best LLM macro-F1, MIMIC-IV ICD-10

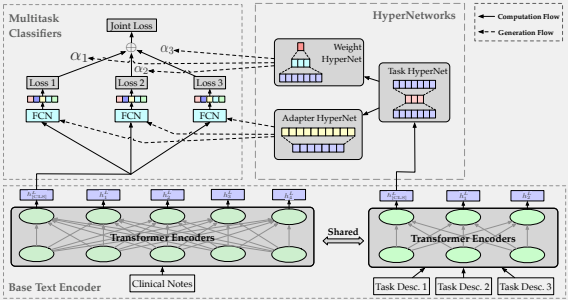
47%
best LLM micro-F1; classifier 50%

2×
macro-F1 once evaluation is fixed

- **Rare codes, not long notes**: with corrected evaluation, note length barely matters; rare codes cause most errors
- **General LLMs code cautiously**: low recall, near-miss codes; small open models trained on coding rationales do better
- **Faithful explanations**: keywords do not justify a code, and generated rationales can cite evidence absent from the note
- **Data**: MIMIC is US ICU data, possibly under-coded, with rules on which models may see it; ICD-11 needs new benchmarks

Edin, J. et al. (2023). "Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study". *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 2572–2582; Soroush, A. et al. (2024). "Large language models are poor medical coders—benchmarking of medical code querying". *NEJM AI* 1.5, AIdbp2300040; Yin, Z., Cao, Y., Wang, T., Chen, J., and Ma, F. (2026). "ICDAGENT: Empowering Agentic Large Language Models for Explainable Medical Coding". *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14131–14149

OUTCOME PREDICTION



Tasks from MIMIC-III notes: 30-day readmission, diagnosis (∼2.6k classes), length of stay. Hypernetworks generate **task heads from label semantics** and learn the task loss weights.

Zero-shot diagnosis

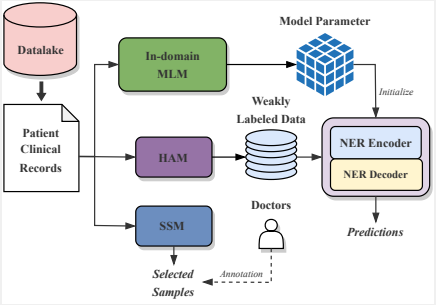
415 diagnoses unseen in training (MIMIC-III discharge notes):

Model	AUC
MT-Hyper	**68.3**
MT-RAM	10.4

Encoding what labels *mean* lets a model handle classes it never saw.

Ji, S. and Marttinen, P. (2023). "Patient outcome and zero-shot diagnosis prediction with hypernetwork-guided multitask learning". *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 589–598

NER WITHOUT LABELLED DATA



- Finnish hospital text with no annotated NER data
- **Weak labels** from FinMeSH dictionaries + lemmatiser, matched by edit distance
- **Continued pretraining** on hospital text
- **Sample selection** by clustering: doctors annotate a representative test set

F1	Surgery	Radiology
Weak labels + pretraining	**13.7**	**14.2**
w/o continued pretraining	7.8	8.3

Pretraining nearly doubles F1.

Sun, W. et al. (2023). "Weak supervision and clustering-based sample selection for clinical named entity recognition". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 444–459

FINE-TUNE OR PROMPT FOR NER?

- **GPT-NER** recasts tagging as generation: the LLM rewrites the sentence with entities marked; few-shot
- Which examples go in the prompt matters
- LLMs over-confidently label non-entities (hallucination)
- Below a fine-tuned model overall, but clearly better when labelled data is extremely scarce

Supervised fine-tuning	In-context learning
Lots of labelled data	No or very little labelled data
Restricted label set	Huge or open label set
Short entities	Long, fuzzily defined entities
Many common entities	Mainly long-tail entities

Takeaway. Weak labels (previous slide) and few-shot prompting are two answers to the same problem: no annotated in-domain data.

Wang, S. et al. (2025). “Gpt-ner: Named entity recognition via large language models”. *Findings of the association for computational linguistics: NAACL 2025*, pp. 4257–4275; Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617

ENTITY LINKING

Why it is hard

- Many forms, one concept: *MI*, *heart attack* → myocardial infarction (C0027051)
- Similar forms, different concepts: *Candida* (a yeast) vs. *Cardia* (part of the stomach)
- UMLS: >3.3M concepts, mostly English terms

SapBERT

Contrastive learning on UMLS synonyms; link each mention to its nearest concept.

Dutch cancer forum

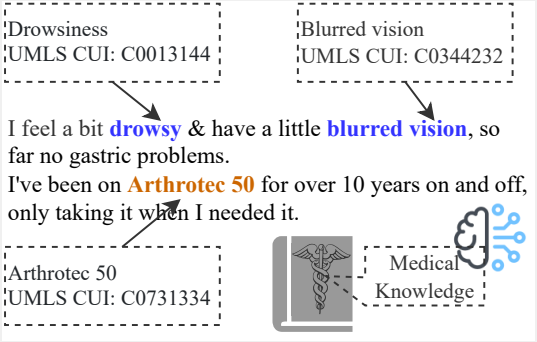
368,840 mentions from MedRoBERTa.nl NER, linked with a Dutch SapBERT; 100 checked by hand:

Outcome	Mentions
NER error	42
Wrong concept	20
Related concept	20
Correct	18

Errors accumulate across pipeline stages.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021). "Self-alignment pretraining for biomedical entity representations". *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4228–4238; Hartendorp, F., Seinen, T., Mulligen, E. van, and Verberne, S. (2024). "Biomedical entity linking for dutch: Fine-tuning a self-alignment bert model on an automatically generated wikipedia corpus". *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pp. 253–263; Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617

ADVERSE DRUG EVENTS (ADE)



- Pharmacovigilance relies on spontaneous reports, which are incomplete and biased
- Patients describe side effects on social media and forums
- Task: does a post report an ADE? Needs to know *which strings are drugs and which are reactions* (e.g. UMLS concepts)

Gao, Y., Ji, S., and Marttinen, P. (2024). "Knowledge-augmented graph neural networks with concept-aware attention for adverse drug event detection". *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9787–9798

ADES FROM PATIENT FORUMS

1. Filter potentially relevant posts
2. Get or create NER training data
3. Tag drug names and side effects (NER)
4. Normalise side effects to an ontology (e.g. SNOMED CT)
5. Relate drugs and side effects in the same post
6. Compare with a knowledge base to flag *new* relations

Needs: drug and side-effect vocabularies, pretrained biomedical encoders, labelled data for training and evaluation.

Rare cancer forum (GIST)

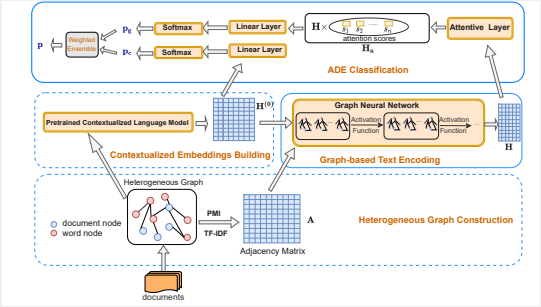
124,103 posts (2009–2020), donated by the patient organisation for research.

ADE extraction	Score
Precision / recall	70% / 74%
F1	0.72
Human pairwise F1	0.80

Whole forum: 13,376 ADE mentions for imatinib, 2,335 for sunitinib. Patient text complements pharmacovigilance for rare cancers.

Dirkson, A., Verberne, S., Kraaij, W., Oortmerssen, G. van, and Gelderblom, H. (2022). "Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers". *Scientific Reports* 12.1, p. 10317; Magge, A. et al. (2021). "DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter". *Journal of the American Medical Informatics Association* 28.10, pp. 2184–2192; Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617

GRAPHS + BERT: CGEM



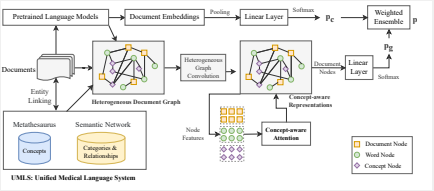
- Corpus graph of documents and words (TF-IDF, PMI edges)
- Document nodes start from BERT embeddings; GCN/GAT on top
- Interpolate graph and BERT predictions

F1	SMM4H	TwiMed-Pub
CGEM	**89.9**	**86.7**
Strongest baseline	84.3	85.3

Takeaway. General-domain RoBERTa beats clinical BERTs on tweets: choose the pretrained model by text register.

Gao, Y., Ji, S., Zhang, T., Tiwari, P., and Marttinen, P. (2022). "Contextualized graph embeddings for adverse drug event detection". *joint European conference on machine learning and knowledge discovery in databases*. Springer, pp. 605–620

ADDING KNOWLEDGE: KNOWCAGE



- UMLS concepts become graph nodes
- Concept-aware attention: separate queries per node-type pair

F1	SMM4H	CADEC	TwiMed-Pub
KnowCAGE	**91.8**	**90.4**	**87.3**
CGEM	89.9	–	86.7
GPT-4 zero-shot	75.9	75.1	87.0

Takeaway. GPT-4 over-predicts ADEs (SMM4H precision 62.4, recall 96.7): look at precision and recall, not only F1.

Gao, Y., Ji, S., and Marttinen, P. (2024). "Knowledge-augmented graph neural networks with concept-aware attention for adverse drug event detection". *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9787–9798

03

MENTAL HEALTH

Content note: this part discusses depression and suicide

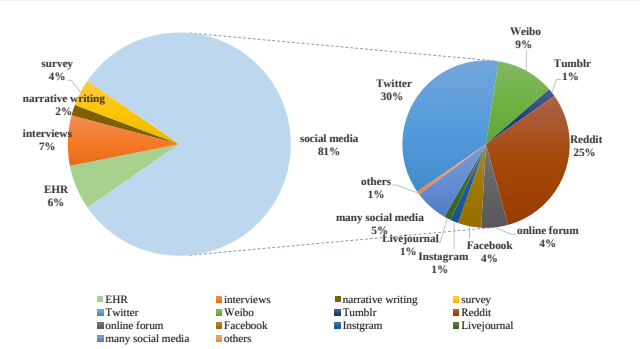


A DECADE OF MENTAL HEALTH NLP

Narrative review of 2012–2021:

10,467 records → 7,536 deduplicated → 611 screened in → **399 papers**

- 81% social media, 7% interviews, 6% EHRs
- 81% English
- Depression 45%, suicide 20%



Zhang, T., Schoene, A. M., Ji, S., and Ananiadou, S. (2022). "Natural language processing applied to mental illness detection: a narrative review". *NPJ digital medicine* 5.1, p. 46

DOMAIN PRETRAINING: MENTALBERT

- Continued masked-LM pretraining of BERT and RoBERTa on mental-health subreddits
- Released publicly (with access requests for responsible use)
- A strong, cheap baseline for later LLM studies

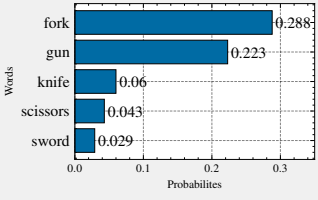
Weighted F1	MentalBERT	MentalRoBERTa
DR (depression)	94.6	94.2
CLPsych15	62.6	69.7
Dreaddit (stress)	80.0	81.8
T-SID (suicide)	88.6	89.0
SAD (stress causes)	67.3	68.4
CAMS (causes)	39.7	47.6

Numbers as reported in the EMNLP 2023 benchmark.

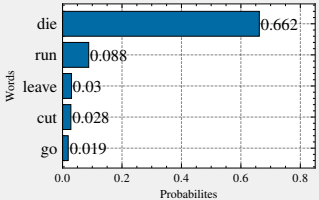
Ji, S. et al. (2022). "Mentalbert: Publicly available pretrained language models for mental healthcare". *proceedings of the thirteenth language resources and evaluation conference*, pp. 7184–7190; Yang, K. et al. (2023). "Towards interpretable mental health analysis with large language models". *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6056–6077

CLASSIFYING ≠ UNDERSTANDING

Fill-mask: *"This life is not worth living. I am going to buy a knife and [MASK]."*



BERT

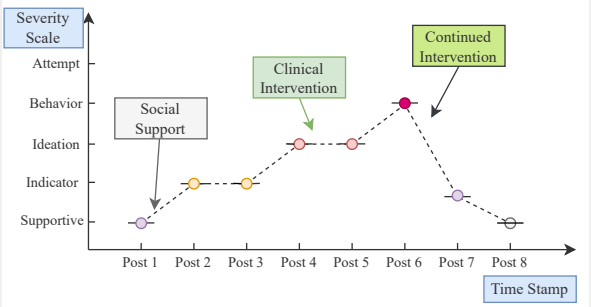


MentalRoBERTa

- High F1 on keyword labels ≠ understanding suicidal *intent*
- Domain pretraining shifts behaviour; GPT-2 continuations were violent, not empathetic
- Needs fine-grained (C-SSRS) labels, recall, uncertainty and human oversight

Ji, S. (2022). "Towards intention understanding in suicidal risk assessment with natural language processing". *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4028–4038

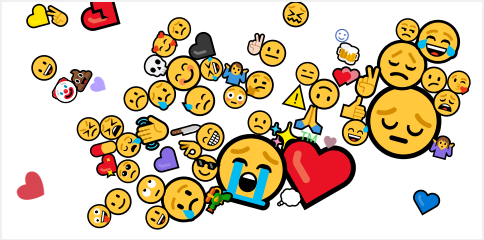
FROM ONE POST TO A TIMELINE



- Model severity across a user's posts, not one-shot classification
- Time interventions: social support, then clinical help
- Open issues: most suicide datasets need data usage agreements; ∼82.5% English; no multilingual intent datasets

Ji, S. (2022). "Towards intention understanding in suicidal risk assessment with natural language processing". *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4028–4038

DO NOT THROW AWAY THE EMOJIS



Emojis in suicide-related Reddit posts

Suicide class	Recall	F1
BERT, emojis removed	71.4	78.7
BERT, original	78.3	84.4
BERT, emoji → text	**88.0**	**87.1**
ChatGPT, zero-shot	72.6	80.3

Takeaway. Preprocessing choices are modelling choices.

25,051 posts with emojis (2,329 suicide-related); suicide-related posts are longer and use fewer, more negative emojis.

Zhang, T. et al. (2024). "SuicidEmoji: derived emoji dataset and tasks for suicide-related social content". *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1136–1141

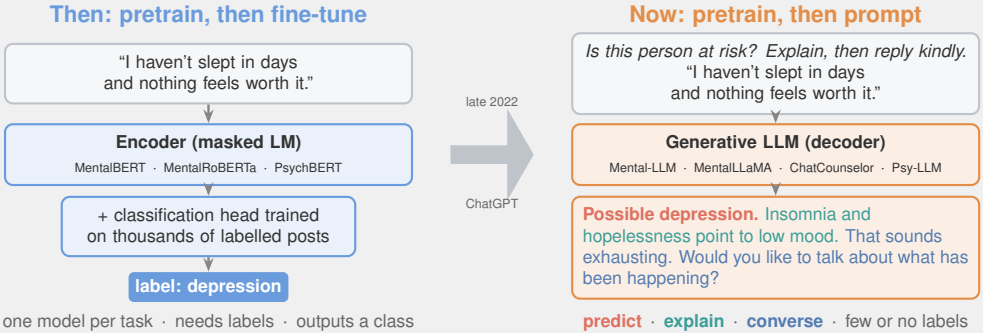
04

LARGE LANGUAGE MODELS

Promise, limits and evaluation



A PARADIGM SHIFT



From models that *classify* to models that *predict, explain and converse*. Is the shift justified?

Ji, S., Zhang, T., Yang, K., Ananiadou, S., and Cambria, E. (2023). *Rethinking large language models in mental health applications.*

CHATGPT FOR MENTAL HEALTH?

Weighted F1	ChatGPT zero-shot	MentalRoBERTa fine-tuned
DR	82.4	**94.2**
Dreaddit	71.8	**81.8**
T-SID	33.3	**89.0**
CAMS	33.9	**47.6**

Emotion-enhanced chain-of-thought plus expert few-shot examples narrow the gap (Dreaddit 75.4) but do not close it.

- **Prompt sensitivity**: changing one adjective swings Dreaddit F1 from 53.1 to 71.8
- **Explanations**: fluent and mostly reliable when the prediction is right; unreliable when it is wrong
- **Automatic metrics** (BARTScore best) only moderately match human ratings

Yang, K. et al. (2023). "Towards interpretable mental health analysis with large language models". *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6056–6077

ADAPTING LLMS TO MEDICINE

Three stages

Pretraining next-token prediction on broad text: language and facts

SFT instruction–response pairs, cross-entropy on the response: task behaviour

RL preference or reward signals: alignment and reasoning

Why adapt? Is *PT* physical therapy or prothrombin time?

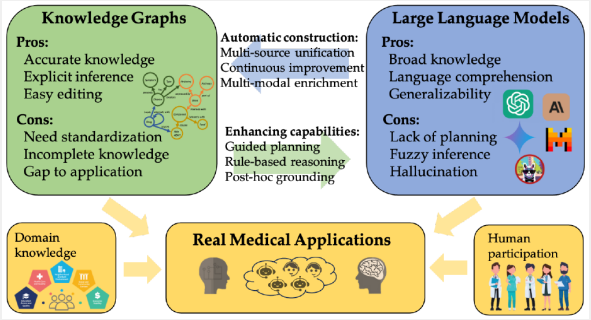
Continued pretraining vs. SFT

	Continued pretraining	SFT
Data	unlabelled notes, PubMed	instruction–response pairs
Objective	same as pretraining	next token on the response
Learns	domain language	task behaviour
Example	BioBERT, MentalBERT	*"Summarise this discharge note"* → summary

Cost grows from prompt tuning to instruction tuning to full continual pretraining.

Sahoo, S. S. and Wang, Y. (2025). *A Hands-on Tutorial on Supervised Fine-Tuning and Reinforcement Learning of Large Language Models for Clinical Applications.* Tutorial at AMIA 2025 Annual Symposium

KNOWLEDGE GRAPHS MEET LLMS



LLMs build KGs unify sources, extract new triples from the literature, add multimodal data

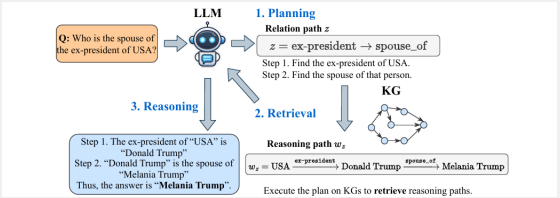
KGs guide LLMs planning, structured reasoning, post-hoc error checking

Co-modelling decision support, drug discovery, knowledge management

The same idea as code descriptions in coding and UMLS in KnowCAGE: explicit knowledge where models are weak.

Xu, R. et al. (2025). "A survey on unifying large language models and knowledge graphs for biomedicine and healthcare". *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6195–6205

REASONING ON GRAPHS



RoG: plan a relation path z, retrieve matching paths from the KG, reason over them:

$$P_\theta(a \mid q, \mathcal{G}) = \sum_z P_\theta(a \mid q, z, \mathcal{G})\, P_\theta(z \mid q)$$

Asked *"Will this patient with type 2 diabetes develop cardiovascular comorbidities?"*, a doctor first plans sub-questions: obesity? metformin? smoking? blood pressure? LLMs plan poorly without such knowledge.

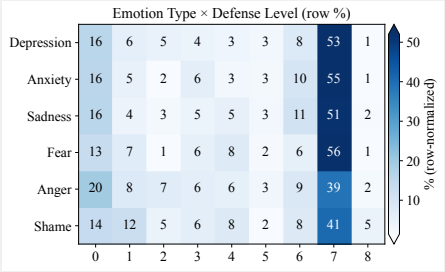
- **GCR**: a KG-Trie constrains decoding so every reasoning path exists in the KG
- **Reflection**: QA pairs generated from KG triples audit what an LLM knows and whether its reasoning is grounded

Luo, L., Li, Y.-F., Haffari, R., and Pan, S. (2024). "Reasoning on graphs: Faithful and interpretable large language model reasoning". *International Conference on Learning Representations*. Vol. 2024, pp. 14400–14423; Luo, L. et al. (2025). "Graph-constrained reasoning: faithful reasoning on knowledge graphs with large language models". *Proceedings of the 42nd International Conference on Machine Learning*, pp. 41540–41565; Xu, R. et al. (2025). "A survey on unifying large language models and knowledge graphs for biomedicine and healthcare". *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6195–6205

PSYCHOLOGICAL DEFENSES

- **PsyDefConv**: 200 emotional-support dialogues, 2,336 help-seeker turns labelled with DMRS defense levels ($\kappa = 0.639$)
- **LLM co-pilot** speeds up expert annotation by 22%

	Acc.	Macro-F1
Ministral-8B fine-tuned	**64.8**	**31.5**
Gemini 2.5 Pro zero-shot	56.4	26.0
GPT-5 zero-shot	52.8	19.5



Defense level by emotion (row %)

Models over-predict the majority level: report macro metrics.

Na, H. et al. (2026). "You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations". *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14428–14448

SYNTHETIC CLINICAL TEXT

- Targets: prescriptions, clinical letters, short episode descriptions, full records
- From a dictionary or knowledge base to dense labelled text is enough to train NER; clinically sound *letters* are much harder
- Synthetic4Health (annotated letters), MLM4SynMed (records generated by masked language modelling)

Evaluate three things

Similarity: ROUGE, embeddings, length distributions, can a classifier tell real from synthetic? **Utility**: train on synthetic, test on real. **Diversity**.

Utility check: ICPC coding

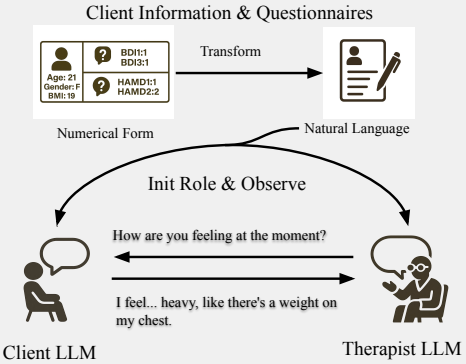
2M Dutch primary-care episodes (average 2.3 words, ∼1,300 ICPC classes); GPT-2 generates episodes from code labels.

Training data	Acc.	F1
Real	**91.8**	**89.8**
Synthetic	81.4	72.2

Usable, but less diverse: F1 drops more than accuracy.

Ren, L., Belkadi, S., Han, L., Del-Pinto, W., and Nenadic, G. (2025). “Synthetic4Health: generating annotated synthetic clinical letters”. *Frontiers in digital health* 7, p. 1497130; Belkadi, S., Ren, L., Micheletti, N., Han, L., and Nenadic, G. (2025). “MLM4SynMed: Masked Language Modelling for Synthetic Free-text Medical Records Generation”. *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*. IEEE, pp. 531–542; Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617

SYNTHETIC COUNSELLING DATA



- Real therapy transcripts are scarce and protected
- Questionnaires (HAM-D, HAM-A, BDI) of 2,090 people ground a **client agent** talking to a CBT **therapist agent**
- Only **open-weight models run locally**: clinical data never leaves the institution
- Fine-tuned Llama3-8B beats a Psych8k-tuned baseline on CounselingBench (macro-F1 0.484 vs. 0.455)
- LLM judges correlate only $r \approx 0.4$ with experts: human evaluation stays essential

Vu, D. N. L. et al. (Oct. 2026). "Roleplaying with Structure: Synthetic Therapist-Client Conversation Generation from Questionnaires". *Proceedings of the Fifth Workshop on NLP for Positive Impact (NLP4PI)*. ed. by Atwell, K. et al. Budapest, Hungary: Association for Computational Linguistics, pp. 395–417

05

LESSONS AND OUT-LOOK

What carries over across tasks



RECURRING LESSONS

Modelling

- Domain pretraining is a cheap, large win
- Inject knowledge: code descriptions, UMLS, hierarchies
- Use label semantics and related tasks (multitask, zero-shot)

Evaluation

- Report macro and per-class metrics; check confusion matrices
- Precision vs. recall matters (missed risk is costly)
- Prompt sensitivity and LLM judges need auditing

Data

- Labels are often proxies (subreddits, keywords, billing codes)
- Weak supervision and synthetic data help when annotation is scarce
- Spend expert time on representative test sets

RESPONSIBLE USE

- **Decision support, not diagnosis**: keep clinicians in the loop
- **Privacy**: public data still needs ethics approval; paraphrase examples; prefer local models for clinical data
- **Bias**: English, US, single-platform data do not generalise
- **Faithful explanations**: attention and self-explanations can mislead
- **Uncertainty**: almost absent in suicide-risk work
- **Monitoring**: models drift as ontologies, guidelines and populations change

OPEN PROBLEMS



Discussion: which of the tasks in this lecture would you trust an LLM to do today, and what evidence would change your mind?

FURTHER LEARNING: TUTORIALS

Venue	Tutorial	Hands-on
LREC 2026	HealthcareNLP: where are we and what is next? (Han, Rayson, Verberne, Moore, Nenadic)	annotation and NER exercises
KDD 2025	MedKLM: knowledge graphs and LLMs for medicine (Xu et al.)	survey and outline
AMIA 2025	Supervised fine-tuning and RL of LLMs for clinical applications (Sahoo, Wang)	Colab: LoRA SFT, GRPO, DPO

Code: `https://github.com/4dpicture/HealthNLP`, `https://github.com/HECTA-UoM/InsightBuddy-AI`, `https://github.com/CogStack/MedCATtutorials`

Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617; Xu, R. et al. (2025). "A survey on unifying large language models and knowledge graphs for biomedicine and healthcare". *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6195–6205; Sahoo, S. S. and Wang, Y. (2025). *A Hands-on Tutorial on Supervised Fine-Tuning and Reinforcement Learning of Large Language Models for Clinical Applications.* Tutorial at AMIA 2025 Annual Symposium

REFERENCES I

Belkadi, S., Ren, L., Micheletti, N., Han, L., and Nenadic, G. (2025). “MLM4SynMed: Masked Language Modelling for Synthetic Free-text Medical Records Generation”. *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*. IEEE, pp. 531–542.

Dirkson, A., Verberne, S., Kraaij, W., Oortmerssen, G. van, and Gelderblom, H. (2022). “Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers”. *Scientific Reports* 12.1, p. 10317.

Dong, H. et al. (2022). “Automated clinical coding: what, why, and where we are?” *NPJ digital medicine* 5.1, p. 159.

Edin, J. et al. (2023). “Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study”. *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 2572–2582.

Gao, Y., Ji, S., and Marttinen, P. (2024). “Knowledge-augmented graph neural networks with concept-aware attention for adverse drug event detection”. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9787–9798.

Gao, Y., Ji, S., Zhang, T., Tiwari, P., and Marttinen, P. (2022). “Contextualized graph embeddings for adverse drug event detection”. *joint European conference on machine learning and knowledge discovery in databases*. Springer, pp. 605–620.

Han, L., Rayson, P., Verberne, S., Moore, A., and Nenadic, G. (2026). *HealthcareNLP: where are we and what is next?* Tutorial at LREC 2026, arXiv:2512.08617.

Hartendorp, F., Seinen, T., Mulligen, E. van, and Verberne, S. (2024). “Biomedical entity linking for dutch: Fine-tuning a self-alignment bert model on an automatically generated wikipedia corpus”. *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pp. 253–263.

Ji, S. (2022). “Towards intention understanding in suicidal risk assessment with natural language processing”. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4028–4038.

Ji, S. and Marttinen, P. (2023). “Patient outcome and zero-shot diagnosis prediction with hypernetwork-guided multitask learning”. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 589–598.

Ji, S., Zhang, T., Yang, K., Ananiadou, S., and Cambria, E. (2023). *Rethinking large language models in mental health applications*.

Ji, S. et al. (2022). “Mentalbert: Publicly available pretrained language models for mental healthcare”. *proceedings of the thirteenth language resources and evaluation conference*, pp. 7184–7190.

Ji, S. et al. (2024). “A unified review of deep learning for automated medical coding”. *ACM computing surveys* 56.12, pp. 1–41.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021). “Self-alignment pretraining for biomedical entity representations”. *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4228–4238.

REFERENCES II

Lobban, F. et al. (2025). “Evaluating peer online forums to support health: ethical and practical challenges”. *J Med Internet Res* 27.1, e73427.

Luo, L., Li, Y.-F., Haffari, R., and Pan, S. (2024). “Reasoning on graphs: Faithful and interpretable large language model reasoning”. *International Conference on Learning Representations*. Vol. 2024, pp. 14400–14423.

Luo, L. et al. (2025). “Graph-constrained reasoning: faithful reasoning on knowledge graphs with large language models”. *Proceedings of the 42nd International Conference on Machine Learning*, pp. 41540–41565.

Magge, A. et al. (2021). “DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter”. *Journal of the American Medical Informatics Association* 28.10, pp. 2184–2192.

Na, H. et al. (2026). “You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations”. *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14428–14448.

Ren, L., Belkadi, S., Han, L., Del-Pinto, W., and Nenadic, G. (2025). “Synthetic4Health: generating annotated synthetic clinical letters”. *Frontiers in digital health* 7, p. 1497130.

Sahoo, S. S. and Wang, Y. (2025). *A Hands-on Tutorial on Supervised Fine-Tuning and Reinforcement Learning of Large Language Models for Clinical Applications*. Tutorial at AMIA 2025 Annual Symposium.

Soroush, A. et al. (2024). “Large language models are poor medical coders—benchmarking of medical code querying”. *NEJM AI* 1.5, AIdbp2300040.

Sun, W., Ji, S., Cambria, E., and Marttinen, P. (2021). “Multitask recalibrated aggregation network for medical code prediction”. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 367–383.

Sun, W., Ji, S., Cambria, E., and Marttinen, P. (2022). “Multitask balanced and recalibrated network for medical code prediction”. *ACM Transactions on Intelligent Systems and Technology* 14.1, pp. 1–20.

Sun, W. et al. (2023). “Weak supervision and clustering-based sample selection for clinical named entity recognition”. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 444–459.

Vu, D. N. L. et al. (Oct. 2026). “Roleplaying with Structure: Synthetic Therapist-Client Conversation Generation from Questionnaires”. *Proceedings of the Fifth Workshop on NLP for Positive Impact (NLP4PI)*. Ed. by Atwell, K. et al. Budapest, Hungary: Association for Computational Linguistics, pp. 395–417.

Wang, S. et al. (2025). “Gpt-ner: Named entity recognition via large language models”. *Findings of the association for computational linguistics: NAACL 2025*, pp. 4257–4275.

REFERENCES III

Xu, R. et al. (2025). “A survey on unifying large language models and knowledge graphs for biomedicine and healthcare”. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6195–6205.

Yang, K. et al. (2023). “Towards interpretable mental health analysis with large language models”. *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6056–6077.

Yin, Z., Cao, Y., Wang, T., Chen, J., and Ma, F. (2026). “ICDAGENT: Empowering Agentic Large Language Models for Explainable Medical Coding”. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14131–14149.

Zhang, T., Schoene, A. M., Ji, S., and Ananiadou, S. (2022). “Natural language processing applied to mental illness detection: a narrative review”. *NPJ digital medicine* 5.1, p. 46.

Zhang, T. et al. (2024). “SuicidEmoji: derived emoji dataset and tasks for suicide-related social content”. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1136–1141.

ellis

THANK YOU

shaoxiong.ji@utu.fi
olaresearch.org/jis