

Rețele neuronale și deep learning

Cursul 12: Modele lingvistice mari în finanțe și formularea instrucțiunilor (prompt engineering)



Prof. dr. Daniel Traian PELE

Academia de Studii Economice din București (ASE), Departamentul de Statistică și Econometrie
Academia Română, Institutul de Prognoză Economică (IPE)
IDA Institute Digital Assets ▪ AI4EFin ▪ MSCA Digital Finance

Obiectivele cursului

- Să scrieți obiectivul unui model de limbaj, perplexitatea și temperatura de eșantionare și să le măsurați pe discursuri ale unei bănci centrale.
- Să explicați preantrenarea, ajustarea pe instrucțiuni și alinierea la preferințele umane, cu funcțiile lor de pierdere.
- Să transformați un prompt într-un clasificator și să evaluați prompturile zero-shot, few-shot și chain-of-thought pe texte financiare etichetate, cu intervale de încredere.
- Să construiți și să evaluați generarea augmentată prin regăsire (RAG) pe discursuri reale și să măsurați halucinațiile.
- Să testați retrospectiv un model de limbaj folosit pentru valoarea expusă la risc (VaR) și să cunoașteți riscurile: scurgerea de date, costul, calibrarea.

Abrevieri

AI	inteligență artificială
API	interfață de programare a aplicațiilor
ASE	Academia de Studii Economice din București
BCE	Banca Centrală Europeană
BERT	reprezentări bidirecționale ale codificatorului transformer
BM25	funcția de ordonare Best Matching 25
BPE	codificare prin perechi de octeți (byte-pair encoding)
DPO	optimizarea directă a preferințelor
ECE	eroarea de calibrare așteptată
ES	pierderea așteptată în coada distribuției (expected shortfall)
FinBERT	model BERT adaptat textelor financiare
GARCH	heteroscedasticitate condiționată autoregresivă generalizată
GPT	transformer generativ preantrenat
HS	simulare istorică
IPE	Institutul de Prognoză Economică, Academia Română
KL	divergența Kullback–Leibler
LLM	model lingvistic mare
LoRA	adaptare de rang mic (low-rank adaptation)
LR	raportul de verosimilitate
MRR	rangul reciproc mediu
RAG	generare augmentată prin regăsire (retrieval-augmented generation)
RLHF	reinforcement learning from human feedback
SFT	ajustare fină supervizată
TF-IDF	frecvența termenului–frecvența inversă în documente
UE	Uniunea Europeană
VaR	valoarea expusă la risc (Value-at-Risk)

Cuprins

- Prezicerea următorului token și modelele lingvistice mari
- Preantrenare, ajustare pe instrucțiuni și aliniere
- Evaluarea formulării instrucțiunilor pe date etichetate
- Generare augmentată prin regăsire pe discursurile BCE
- Modelele de limbaj ca modele de risc: LLM-VaR
- Riscuri și guvernanță

Datele de tip text în instituțiile financiare

- **Piețe**
  - Știrile și rețelele sociale mișcă prețurile în cîteva minute; sentimentul este un semnal
- **Bănci centrale**
  - Discursurile și minutele ghidează așteptările privind dobînzile; în datele noastre, Banca Centrală Europeană (BCE) are 400 de discursuri din 2022
- **Credit și risc**
  - Dosare de credit, rapoarte anuale, rapoarte de risc: citirea lor costă ore de analiză
- **Acest curs**
  - Modele lingvistice mari (LLM) deschise, rulate pe un laptop, evaluate pe date etichetate, niciodată acceptate fără verificare

Modelul de limbaj: probabilitatea următorului token, prin entropie încrucișată

- **Pe scurt:** tot ce face un LLM, de la răspunsuri la cod, este prezicerea repetată a următorului token.

$$p_\theta(x_1, \ldots, x_T) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t})$$

$$\mathcal{L}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}), \quad \mathrm{PPL} = e^{\mathcal{L}}$$

- **Discursuri BCE** după iulie 2025: perplexitatea 14,71 (0,5B), 10,63 (1,5B), 9,43 (3B).

- **unde**
  - x_t = tokenul t
  - p_θ = softmax peste vocabular (transformer cauzal, cursul 7)
  - Perplexitatea = numărul efectiv de tokeni următori echiprobabili.

- **În practică:** o perplexitate mai mică pe documentele proprii înseamnă că modelul „vorbește” limbajul domeniului.

(Bishop & Bishop, 2024, ch. 12)

Exercițiu practic: notebook §1

Tokenii ca fragmente de cuvinte: codificarea byte-pair

- **Pe scurt:** cuvintele frecvente sînt un singur token, iar cele rare și numerele se împart; modelul nu vede litere, doar indici de tokeni.

$$(a, b)^* = \arg\max_{(a,b)} \ \mathrm{count}(a\,b) \ \Rightarrow \ \text{unește } a\,b \to ab$$

- **unde**
  - Se repetă pînă cînd vocabularul este complet: codificarea byte-pair (BPE), 151 665 de tokeni în Qwen2.5
  - WordPiece, 30 522 de tokeni într-un model de reprezentări bidirecționale ale codificatorului transformer (BERT) adaptat finanțelor (FinBERT).

- **Exemplu** (20 tokeni): - | Cash | flow | from | operating | activities | before | investments | was | EUR | | 0 | . | 8 | - | 1 | . | 2 | million | ..
- **Financial PhraseBank**: 1,35 tokeni Qwen și 1,24 tokeni WordPiece pe cuvînt.

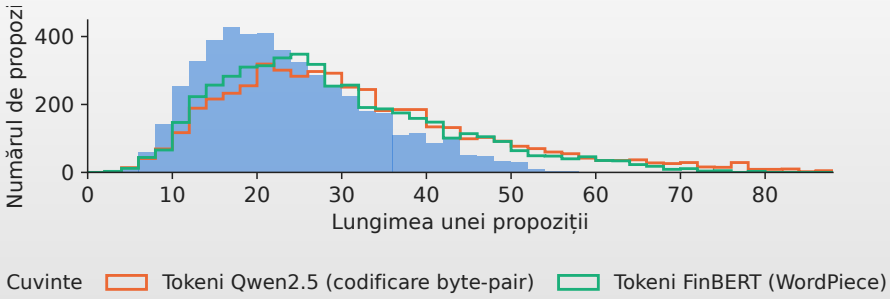
- **În practică:** costul și limita contextului se măsoară în tokeni, nu în cuvinte: socotiți aproximativ 1,3 tokeni pe cuvînt în engleză.

(Sennrich, Haddow & Birch, 2016)

Exercițiu practic: notebook §1

Lungimea propozițiilor în cuvinte și în tokeni: două metode de tokenizare

- **Pe scurt:** aceleași 4 838 propoziții sînt puțin mai lungi în tokeni decît în cuvinte.



Sursa: Financial PhraseBank (Malo et al., 2014). Date: 4.838 de propoziții.

- Media: 23,1 cuvinte, 31,1 tokeni Qwen, 28,6 tokeni WordPiece
  - Un discurs BCE are în medie 4 143 de tokeni (cel mai lung: 22 897).

NN_L12_tokens_per_sentence

Temperatura schimbă forma distribuției următorului token

- **Pe scurt:** modelul produce scoruri (logits); temperatura decide cît de îndrăznește eșantionăm din ele.

$$p_T(v \mid x_{<t}) = \frac{\exp(z_v/T)}{\sum_u \exp(z_u/T)}$$

$$H_T = -\sum_v p_T(v) \log p_T(v)$$

- **unde**
  - z_v = scorul (logit) tokenului v
  - $T \to 0$: mereu tokenul cel mai probabil (lacom)
  - $T > 1$: distribuție mai plată, mai diversă, cu mai multe erori
  - H_T = entropia, în nats.

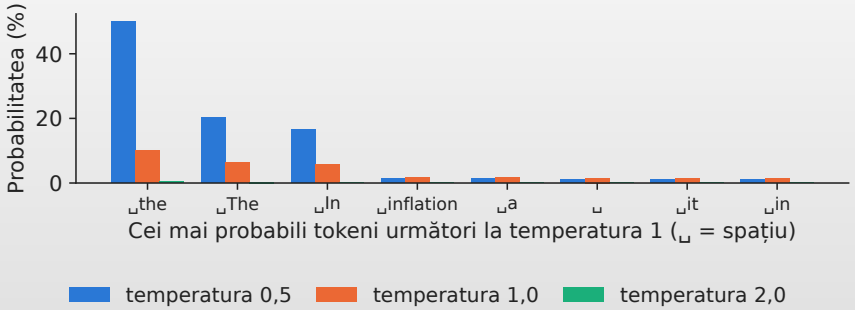
- **Prompt** din discursul BCE din 30 octombrie 2026: „The inflation impact is limited: . . . ”
- **Primul token** „the”: 49,9 % la $T = 0{,}5$, 10,1 % la 1, 0,4 % la 2
  - Entropia 1,74, 6,20, 10,29.

- **În practică:** pentru clasificare și extragere folosiți temperatura 0
  - Temperaturile mari sînt pentru generarea de idei.

Exercițiu practic: notebook §2

Probabilitățile următorului token: un model real, o frază reală

- **Pe scurt:** la temperatură mică aproape toată masa de probabilitate stă pe o continuare; la temperatură mare, tokenii improbabili devin plauzibili.



Sursa: Banca Centrală Europeană, setul de discursuri. Date: începutul unei fraze din discursul BCE din 30 octombrie 2026; modelul Qwen2.5-1.5B-Instruct.

NN_L12_next_token

Legile de scalare: pierderea ca funcție putere de dimensiunea modelului

- **Pe scurt:** pierderea scade ca o putere a numărului de parametri și de tokeni de antrenare.

$$\mathcal{L}(P, D) = E + \frac{A}{P^{\alpha}} + \frac{B}{D^{\beta}}$$

estimare proprie: $\log \mathcal{L} = a + b \log P, \quad \hat{b} = -0{,}100$

- **Trei puncte nu fac o lege**: panta descrie eșantionul nostru, nu este o prognoză.
- **Optimul de calcul**: aproximativ 20 de tokeni de antrenare pe parametru.

- **unde**
  - $P =$ numărul de parametri
  - $D =$ tokenii de antrenare
  - $E =$ pierderea ireductibilă
  - $\alpha, \beta \approx 0{,}3$ (Hoffmann et al., 2022)
  - $\hat{b}$ din cele trei dimensiuni Qwen2.5, pe discursurile BCE de după iulie 2025.

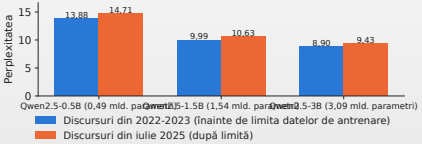
- **În practică:** un model de 3B pe laptop nu costă nimic pe apel
  - Un model de vîrf costă pe token: măsurați întîi cîștigul.

(Kaplan et al., 2020; Hoffmann et al., 2022)

Exercițiu practic: notebook §1

Limita datelor de antrenare și ușurința aparentă a discursurilor mai vechi

- **Pe scurt:** textul pe care modelul l-ar fi putut vedea la preantrenare are o perplexitate mai mică decît textul scris după aceea.



Sursa: Banca Centrală Europeană, setul de discursuri. Date: 40 + 40 de discursuri, primii 1.024 de tokeni din fiecare.

- 3B: 8,90 pe discursurile din 2022–2023, față de 9,43 după iulie 2025
  - Mai mică pentru discursurile vechi la toate dimensiunile.
- Și temele s-au schimbat
  - Diferența este un simptom, nu o dovadă, a memorării.

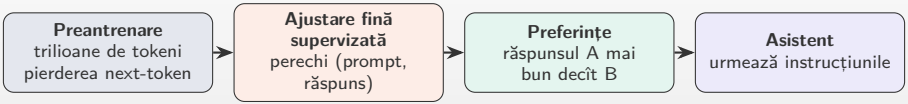
- **În practică:** evaluați un LLM pe documente scrise după limita datelor sale de antrenare
  - Altfel scorul poate fi umflat prin scurgerea datelor.

NN_L12_perplexity_size

Cuprins

- Prezicerea următorului token și modelele lingvistice mari
- Preantrenare, ajustare pe instrucțiuni și aliniere
- Evaluarea formulării instrucțiunilor pe date etichetate
- Generare augmentată prin regăsire pe discursurile BCE
- Modelele de limbaj ca modele de risc: LLM-VaR
- Riscuri și guvernanță

Cele trei etape de antrenare ale unui asistent bazat pe un model de limbaj



- **Transformerul generativ preantrenat (GPT)**: doar etapa 1 dă un model care continuă textul, nu unul care răspunde.
- **Ajustarea fină supervizată (SFT)** învață formatul
  - **reinforcement learning from human feedback (RLHF)** sau **optimizarea directă a preferințelor (DPO)** învață ce răspunsuri preferă oamenii.
- **Qwen2.5-Instruct** (Qwen Team, 2024): 18 trilioane de tokeni la preantrenare, apoi SFT și optimizarea preferințelor.

(Ouyang et al., 2022; Qwen Team, 2024)

Ajustarea pe instrucțiuni este pierderea next-token calculată doar pe răspuns

- **Pe scurt:** modelul vede promptul și răspunsul, dar este evaluat doar pe tokenii răspunsului.

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \sum_{t=1}^{|\mathbf{y}|} \log p_\theta\left(y_t \mid \mathbf{x}, \mathbf{y}_{<t}\right)$$

- **unde**
  - $\mathbf{x}$ = instrucțiunea (cu un șablon de conversație: rolurile sistem, utilizator, asistent)
  - $\mathbf{y}$ = răspunsul de referință
  - $\mathcal{D}$ = de la zeci de mii la milioane de perechi.

- **Adaptarea de rang mic (LoRA)**: se antrenează $\mathbf{W} = \mathbf{W}_0 + \mathbf{BA}$, cu $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$, $r \ll d$
  - $\mathbf{W}_0$ rămîne fix.
- **Un strat** 4096×4096**,** $r = 8$: 65.536 de ponderi antrenabile în loc de 16,8 milioane.

- **În practică:** o bancă își adaptează un model deschis la propriile documente cu LoRA, pe o singură placă grafică, fără ca datele să iasă din instituție.

(Hu et al., 2022)

Alinierea: modele de recompensă și optimizarea directă a preferințelor

- **Pe scurt:** oamenii aleg răspunsul mai bun dintr-o pereche; modelul este împins spre răspunsurile preferate.

$$\Pr(y_w \succ y_l \mid x) = \sigma\big(r_\phi(x, y_w) - r_\phi(x, y_l)\big)$$

$$\max_{\theta} \ \mathbb{E}\big[r_\phi(x, y)\big] - \beta\,\mathrm{KL}\big(p_\theta \,\|\, p_{\mathrm{SFT}}\big)$$

$$\mathcal{L}_{\mathrm{DPO}} = -\log\sigma\left(\beta\log\frac{p_\theta(y_w|x)}{p_{\mathrm{SFT}}(y_w|x)} - \beta\log\frac{p_\theta(y_l|x)}{p_{\mathrm{SFT}}(y_l|x)}\right)$$

- **unde**
  - y_w, y_l = răspunsul preferat și cel respins
  - r_ϕ = modelul de recompensă (Bradley–Terry)
  - σ = funcția logistică
  - Penalizarea Kullback–Leibler (KL), cu ponderea β, ține modelul aproape de modelul SFT.

- **RLHF** estimează r_ϕ, apoi optimizează politica
  - **DPO** renunță la modelul de recompensă.
- **Efect secundar**: modelele aliniate sună sigure
  - Siguranța nu înseamnă precizie.

- **În practică:** alinierea stabilește tonul și siguranța, nu cunoștințele financiare: evaluați pe sarcina proprie.

(Ouyang et al., 2022; Rafailov et al., 2023)

Modelele deschise folosite în acest curs

Model	Parametri	Rolul în acest curs
Qwen2.5-0.5B-Instruct	0,49 mld.	perplexitate, clasificare zero-shot
Qwen2.5-1.5B-Instruct	1,54 mld.	probabilitățile următorului token, few-shot, embedding
Qwen2.5-3B-Instruct	3,09 mld.	cele mai bune prompturi, chain-of-thought, răspunsuri la întrebări, LLM-VaR
BAAI bge-small-en-v1.5	33 mil.	embedding pentru regăsire densă (384 de dimensiuni)
Frecvența termenului–frecvența inversă în documente (TF-IDF) + regresie logistică	–	reperul clasic, antrenat pe datele etichetate

- Toate modelele rulează local, pe un laptop; niciun text nu părăsește calculatorul și nu se plătește nicio interfață de programare a aplicațiilor (API).
- TF-IDF: numărul de apariții ale cuvintelor și perechilor de cuvinte, ponderat cu raritatea lor în documente.

(Qwen Team, 2024; Xiao et al., 2023)

Cuprins

- Prezicerea următorului token și modelele lingvistice mari
- Preantrenare, ajustare pe instrucțiuni și aliniere
- Evaluarea formulării instrucțiunilor pe date etichetate
- Generare augmentată prin regăsire pe discursurile BCE
- Modelele de limbaj ca modele de risc: LLM-VaR
- Riscuri și guvernanță

Structura unui prompt

- **Sistem**
  - Rolul și regulile: „You are a careful financial analyst."
- **Instrucțiune**
  - Sarcina, mulțimea etichetelor, formatul: „Answer with one word: positive, negative or neutral."
- **Exemple** (opțional)
  - k perechi etichetate intrare–răspuns: few-shot
- **Intrare**
  - Propoziția, tweet-ul sau documentul de prelucrat

(Brown et al., 2020; Wei et al., 2022)

- **Variantele testate**
  - Zero-shot: instrucțiune + intrare
  - Few-shot: între 3 și 24 de exemple
  - Chain-of-thought: „explică, apoi răspunde"
  - Cinci reformulări ale instrucțiunii
- **Regula acestui curs**
  - Fiecare prompt este un model: se evaluează pe un set de test separat

Promptul ca clasificator: probabilitățile cuvintelor-etichetă

- **Pe scurt:** citim probabilitatea fiecărei etichete în poziția răspunsului: fără generare de text, fără interpretarea lui și cu un scor de tip probabilitate.

$$\hat{p}(c \mid x) = \frac{\sum_{v \in \mathcal{V}_c} p_\theta(v \mid \text{prompt}(x))}{\sum_{c'} \sum_{v \in \mathcal{V}_{c'}} p_\theta(v \mid \text{prompt}(x))}$$

$$\hat{c} = \arg\max_c \hat{p}(c \mid x)$$

- **unde**
  - $\mathcal{V}_c$ = primii tokeni ai cuvîntului-etichetă c în toate formele lui („positive", „ Positive", . . .)
  - Promptul se termină acolo unde începe răspunsul asistentului.

- **Rapid**: o singură trecere înainte pe element
  - Instrucțiunea și exemplele comune se codifică o singură dată (memoria cheie–valoare).
- **Comparabil**: același macro-F1 și aceleași intervale de încredere ca pentru orice clasificator.

- **În practică:** restrîngeți spațiul răspunsurilor
  - Răspunsurile libere trebuie interpretate și pot eșua fără să observați.

(Zhao et al., 2021)

Exercițiu practic: notebook §3

Două seturi de date etichetate: știri financiare și tweet-uri financiare

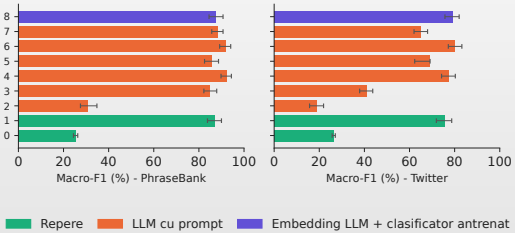
Set de date	Etichete	Antrenare	Test	Ponderea neutrelor (test)
Financial PhraseBank, 100 % acord	pozitiv, negativ, neutru	1581	678	61,4 %
Twitter financial news	optimist (bullish), pesimist (bearish), neutru	9 543	900	65,6 %

- PhraseBank: 4 838 de propoziții din știri despre companii listate în Finlanda, etichetate de 5–8 adnotatori; folosim cele 2 259 cu etichete unanime, împărțite 70/30 stratificat (valoarea generatorului aleator: 20261108).
- Twitter: împărțirea publicată; testul = eșantion stratificat de 900 din cele 2 388 de tweet-uri de validare.
- Exemplele pentru prompturile few-shot se extrag doar din partea de antrenare.

(Malo et al., 2014)

LLM-uri cu prompt comparate cu un reper antrenat

- **Pe scurt:** nicio strategie nu cîștigă peste tot; clasificatorii antrenați folosesc datele etichetate, prompturile nu.



Sursa: Financial PhraseBank (Malo et al., 2014) și Twitter financial news (zeroshot). Date: seturi de test: 678 de propoziții (acord 100 %) și 900 de tweet-uri; intervale bootstrap de 95 %.

- Cel mai bun pe PhraseBank: zero-shot 3B (92,2 %)
  - Pe Twitter: few-shot 3B (79,8 %)
  - Zero-shot 3B: 92,2 % și 77,4 %.

NN_L12_prompt_results

Acuratețea și macro-F1, cu intervale bootstrap de 95 %

Metoda	Acur.	F1	Int. 95 %	Acur.	F1	Int. 95 %
	PhraseBank			Twitter		
clasa majoritară	61,4	25,4	[24,4; 26,3]	65,6	26,4	[25,7; 27,1]
TF-IDF + logistică	90,3	86,9	[83,8; 90,0]	81,4	75,5	[71,9; 78,7]
zero-shot 0,5B	62,8	31,0	[27,5; 34,8]	23,1	18,7	[15,7; 21,9]
zero-shot 1,5B	85,3	85,1	[82,2; 87,9]	37,8	40,9	[37,9; 43,7]
zero-shot 3B	93,1	92,2	[89,8; 94,4]	81,2	77,4	[74,2; 80,3]
few-shot 1,5B	87,8	85,9	[82,5; 88,7]	73,7	68,8	[62,3; 69,1]
few-shot 3B	93,4	92,1	[89,2; 94,1]	84,3	79,8	[77,2; 83,2]
chain-of-thought 3B	88,5	88,4	[85,7; 90,7]	64,7	65,1	[62,0; 68,0]
embedding 1,5B + logistică	91,3	87,5	[84,5; 90,7]	83,0	79,0	[75,7; 82,0]

- Macro-F1 = media celor trei scoruri F1 pe clase, astfel încît clasa mică negativă contează cît cea neutră; intervalele provin din 1.000 de reeșantionări bootstrap ale setului de test; few-shot = media a 3 rulări.
- Int. = intervalul bootstrap de 95 % al macro-F1.

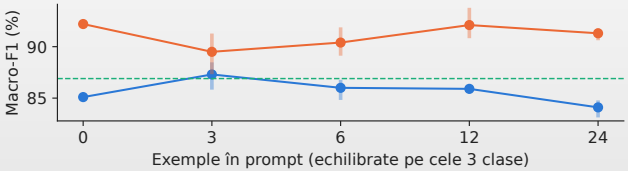
(calcul propriu)

Exercițiu practic: notebook §3

Efectul numărului de exemple din prompt

- **Pe scurt:** învățarea în context: exemplele schimbă răspunsul fără să modifice nicio pondere.

$$\hat{c} = \arg\max_{c} \; p_\theta\big(c \mid (x_1, c_1), \ldots, (x_k, c_k), x\big)$$



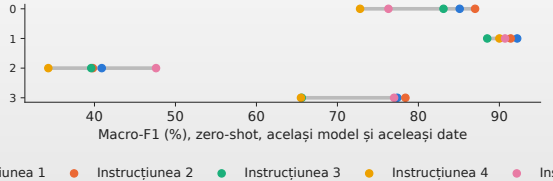
Sursa: Financial PhraseBank (Malo et al., 2014). Date: setul de test de 678 de propoziții; barele = intervalul pe 3 valori ale generatorului aleator.

- 1,5B: 85,1 % fără exemple, 84,1 % cu 24
  - 3B: 92,2 % și 91,3 %
  - Cu 12 exemple, rulările mută 3B între 91,0 și 93,6 %.

NN_L12_fewshot_curve

Sensibilitatea la prompt: efectul reformulării instrucțiunii

- **Pe scurt:** cinci instrucțiuni cu același sens, același model și aceleași date: diferențele sînt o sursă de risc de model.



Sursa: Financial PhraseBank și Twitter financial news. Date: seturile de test de mai sus; cinci reformulări ale instrucțiunii.

- Amplitudinea macro-F1 pe cele cinci formulări: PhraseBank 3B 3,6 puncte, Twitter 3B 12,9 puncte, Twitter 1,5B 13,3 puncte.

- **În practică:** alegeți promptul pe un set de validare și fixați-l, ca pe orice hiperparametru
  - Niciodată pe setul de test.

NN_L12_prompt_sensitivity

Chain-of-thought: raționamentul scris înaintea răspunsului

- **Pe scurt:** tokenii de raționament generați devin context suplimentar pentru eticheta finală.

$$p(c \mid x) = \sum_{r} p(r \mid x)\, p(c \mid x, r) \;\approx\; p(c \mid x, \hat{r})$$

- **unde**
  - r = textul raționamentului, $\hat{r}$ = cel lacom (cel mult două propoziții, în medie 35,9 cuvinte)
  - Eticheta se extrage din „Label: . . . ”
  - Extragerea reușește pentru 100,0 % din răspunsurile PhraseBank.

- **Macro-F1**: PhraseBank 88,4 % (zero-shot 3B 92,2 %)
  - Twitter 65,1 % (77,4 %).
- **Costul**: 389 ms pe propoziție, față de 48,5 ms pentru zero-shot.

- **În practică:** chain-of-thought ajută la raționamente în mai mulți pași
  - Pentru etichete de un cuvînt pe texte scurte, testați înainte să plătiți pentru el.

(Wei et al., 2022)

Exercițiu practic: notebook §5

Un exemplu de răspuns chain-of-thought

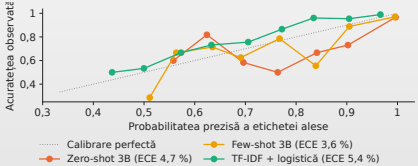
- **Intrare** (setul de test PhraseBank)
  - "CapMan said the deal 's effect on its cash flow for 2009 totals EUR3 .4 m , but the transaction would not affect its financial results for 2009 as it was executed at fair value ."
- **Ieșirea** modelului Qwen2.5-3B-Instruct, temperatura 0
  - "Reasoning: The sentence indicates that the deal has a positive impact on CapMan's cash flow, which is generally seen as beneficial for investors. However, it also clarifies that the transaction does not affect their financial results, suggesting no immediate change to profitability. Given these points, the overall sentiment leans positive as it highlights a direct benefit. Label: positive"
- **Cum se citește**
  - Raționamentul este o poveste plauzibilă, nu o urmă fidelă a calculului
  - Verificați eticheta pe date, nu explicația pe intuiție

(calcul propriu)

Calibrarea probabilităților etichetelor pe date reale

- **Pe scurt:** o probabilitate de 0,9 trebuie să fie corectă de nouă ori din zece; punctele sub diagonală înseamnă încredere excesivă.

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{n} \left|\mathrm{acc}(S_b) - \mathrm{conf}(S_b)\right|$$



Sursa: Financial PhraseBank (Malo et al., 2014). Date: setul de test de 678 de propoziții, 10 intervale.

- Eroarea de calibrare așteptată (ECE): zero-shot 3B 4,7 %, few-shot 3B 3,6 %, TF-IDF + logistică 5,4 %.

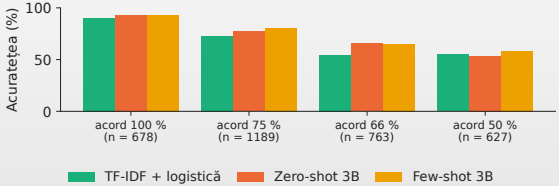
- **În practică:** recalibrați pe un set de validare (scalarea temperaturii) înainte de a folosi probabilitățile LLM ca scoruri de risc.

(Guo et al., 2017)

NN_L12_calibration

Acuratețea în funcție de gradul de acord al adnotatorilor

- **Pe scurt:** propozițiile care îi împart pe adnotatori sînt cu adevărat ambigue; niciun model nu atinge pe ele acuratețea din cazurile unanime.



Sursa: Financial PhraseBank (Malo et al., 2014). Date: propoziții nefolosite nici la antrenare, nici ca exemple.

- Zero-shot 3B: 93,1 % la acord de 100 %, 53,7 % la 50 %
  - TF-IDF: 90,3 % și 54,9 %.

- **În practică:** raportați calitatea etichetelor din setul de test
  - Un model nu poate fi mai sigur decît oamenii care l-au etichetat.

NN_L12_agreement

Matricea de confuzie a celui mai bun prompt

Real \ prezis	pozitiv	negativ	neutru
pozitiv	156	2	13
negativ	1	82	8
neutru	20	3	393

- **Modelul**: zero-shot 3B, setul de test PhraseBank.
- **Neutru citit ca pozitiv**: 20 propoziții
  - **pozitiv citit ca neutru**: 13.
- **Negativ și pozitiv** se confundă rar: $1 + 2$.

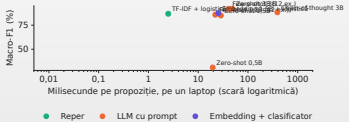
- Erorile stau la granița cu neutrul: „profitul a crescut, dar mai puțin decît se aștepta” este cazul dificil, nu știrile bune față de cele rele.

(calcul propriu)

Exercițiu practic: notebook §6

Costul față de precizie: milisecunde pe propoziție, pe un laptop

- **Pe scurt:** cel mai ieftin model este de mii de ori mai rapid decît cel mai lent; diferența de precizie este mică.



Sursa: Financial PhraseBank (Malo et al., 2014). Date: setul de test de 678 de propoziții; timp măsurat pe procesorul grafic al unui laptop.

- TF-IDF: 2,50 ms pe propoziție
  - Zero-shot 3B: 48,5 ms
  - Chain-of-thought 3B: 389 ms.

- **În practică:** dacă aveți etichete, antrenați un model mic
  - Fără etichete, folosiți un prompt
  - Pentru milioane de documente, costul decide.

NN_L12_cost_accuracy

Embedding-uri LLM cu un clasificator mic, ca alternativă ieftină la ajustarea fină

- **Pe scurt:** fixăm modelul de limbaj, facem media ultimelor stări ascunse și estimăm deasupra o regresie logistică.

$$\mathbf{e}(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} \mathbf{h}_t^{(L)}, \qquad \hat{p}(c \mid x) = \operatorname{softmax}\left(\mathbf{W}\mathbf{e}(x) + \mathbf{b}\right)_c$$

- **Macro-F1**: PhraseBank 87,5 %, Twitter 79,0 %.
- **Folosește etichetele** ca TF-IDF, dar cu cunoștințele lingvistice ale LLM.

- **unde**
  - $\mathbf{h}_t^{(L)} \in \mathbb{R}^{1536}$ = ultima stare ascunsă a Qwen2.5-1.5B
  - $\mathbf{W}$ estimat cu penalizare L2 aleasă prin validare încrucișată în 5 pași.

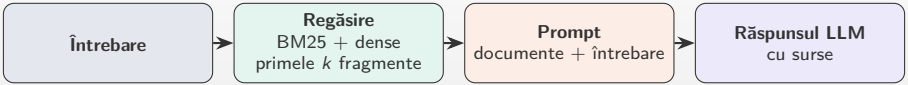
- **În practică:** înainte de o ajustare fină completă, încercați un clasificator pe embedding: cîteva minute pe laptop, fără gradient prin LLM.

Exercițiu practic: notebook §6

Cuprins

- Prezicerea următorului token și modelele lingvistice mari
- Preantrenare, ajustare pe instrucțiuni și aliniere
- Evaluarea formulării instrucțiunilor pe date etichetate
- Generare augmentată prin regăsire pe discursurile BCE
- Modelele de limbaj ca modele de risc: LLM-VaR
- Riscuri și guvernanță

RAG: regăsirea documentelor, urmată de generarea răspunsului



$$p(y \mid q) \approx \sum_{z \in \mathrm{top}_k(q)} p_\eta(z \mid q)\, p_\theta(y \mid q, z)$$

- **Generarea augmentată prin regăsire (RAG)**: z = fragmentul regăsit, p_η = modulul de regăsire, p_θ = generatorul.
- **Corpusul nostru**: 400 de discursuri BCE, 8 400 de fragmente de 200 de cuvinte (suprapunere 50).

(Lewis et al., 2020)

Trei metode de regăsire: cuvinte comune, termeni ponderați, embedding

- **Pe scurt:** metodele lexicale potrivesc cuvinte, cele dense sensul, iar metoda hibridă le combină; interogarea este titlul unui discurs.

$$\mathrm{BM25}(q,z)=\sum_{w\in q}\mathrm{idf}(w)\frac{f(w,z)(k_1+1)}{f(w,z)+k_1\left(1-b+b\frac{|z|}{|z|}\right)}$$

$$s_{\text{dense}}=\cos\big(\mathbf{e}(q),\mathbf{e}(z)\big),\qquad s_{\text{RRF}}(z)=\sum_{m}\frac{1}{60+\mathrm{rank}_m(z)}$$

- **Rangul reciproc mediu (MRR)**: BM25 0,515, dense 0,567, hibrid 0,622.

- **unde**
  - $f(w,z)=$ aparițiile cuvîntului w în fragmentul z
  - $k_1=1{,}5$, $b=0{,}75$
  - Best Matching 25 (BM25)
  - $\mathbf{e}=$ embedding bge-small
  - RRF $=$ fuziunea rangurilor.

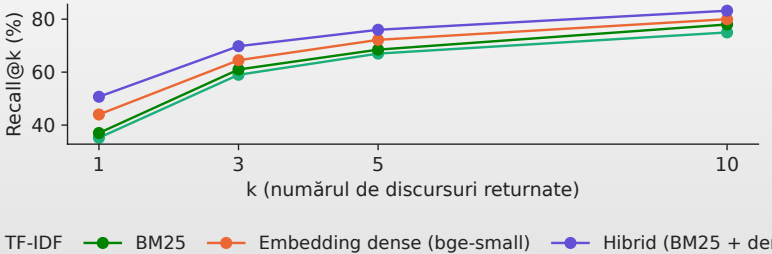
- **În practică:** răspunsul nu poate fi mai bun decît ce s-a regăsit: evaluați întîi separat modulul de regăsire.

(Robertson & Zaragoza, 2009; Xiao et al., 2023)

Exercițiu practic: notebook §7

Regăsirea discursului corect după titlu

- **Pe scurt:** recall@k = ponderea celor 400 de titluri pentru care discursul se află printre primele k rezultate.



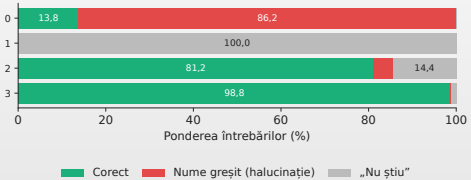
Sursa: Banca Centrală Europeană, setul de discursuri. Date: 400 de discursuri, 8400 de fragmente; interogarea = titlul discursului.

- Recall@1: TF-IDF 35,2 %, BM25 37,0 %, dense 44,0 %, hibrid 50,7 %
  - Recall@10 hibrid 83,2 %.

NN_L12_rag_recall

Răspunsurile cu și fără documente regăsite

- **Pe scurt:** întrebarea: ce membru al conducerii a ținut discursul cu acest titlu? 160 de discursuri din 15 ianuarie 2025, ulterioare datelor de antrenare ale modelului.



Sursa: Banca Centrală Europeană, setul de discursuri. Date: 160 de discursuri cu un singur vorbitor, 15 ianuarie 2025 - 30 octombrie 2026; modelul Qwen2.5-3B-Instruct.

- Răspuns obligatoriu fără documente: 86,2 % nume greșite, Philip Lane de 137 de ori
  - RAG 81,2 % corecte, oracol 98,8 %
  - Doar vorbitorul cel mai frecvent: 35,6 %.

- **În practică:** un nume greșit spus cu siguranță este eșecul costisitor: cereți surse în fiecare răspuns și verificați-le.

NN_L12_rag_qa

Deciziile de proiectare ale unui sistem RAG pentru o bancă

- **Fragmentarea**
  - Aici 200 de cuvinte cu suprapunere de 50; tabelele și notele de subsol au nevoie de reguli proprii
- **Metadatele**
  - Vorbitorul, data și titlul însoțesc fiecare fragment; răspunsul le citează
- **Regăsirea**
  - Hibrid BM25 + dense: recall@3 = 73,1 % pe discursurile folosite la întrebări
- **Abținerea**
  - Permiteți „nu știu”: fără documente, modelul se abține atunci la 100,0 % din întrebări
- **Evaluarea**
  - Un set fix de întrebări cu răspunsuri cunoscute, rulat din nou după fiecare modificare

Cuprins

- Prezicerea următorului token și modelele lingvistice mari
- Preantrenare, ajustare pe instrucțiuni și aliniere
- Evaluarea formulării instrucțiunilor pe date etichetate
- Generare augmentată prin regăsire pe discursurile BCE
- Modelele de limbaj ca modele de risc: LLM-VaR
- Riscuri și guvernanță

LLM-VaR: prognoza cuantilei din istoricul randamentelor inclus în prompt

- **Pe scurt:** fără estimare și fără ajustare fină: modelul citește 250 de randamente și scrie valoarea expusă la risc (VaR).

$$\widehat{\mathrm{VaR}}^{\alpha}_{t+1} = \mathrm{LLM}\big(\mathrm{prompt}(r_{t-249}, \ldots, r_t; \alpha)\big)$$

$$\mathrm{LR}_{uc} = -2 \ln \frac{(1-\alpha)^{n_0} \alpha^{n_1}}{(1-\hat{\pi})^{n_0} \hat{\pi}^{n_1}} \sim \chi^2_1$$

- **unde**
  - r_t = randamentul logaritmic zilnic al S&P 500 (%)
  - $\alpha \in \{1\,\%, 5\,\%\}$
  - Temperatura 0
  - n_1 = depășirile
  - Testul raportului de verosimilitate (LR) Kupiec.

- **Replicarea noastră**: Qwen2.5-3B-Instruct, 459 de zile, din 2 ianuarie 2025 pînă la 30 octombrie 2026.
- **Repere**: simularea istorică (HS), 250 de zile
  - Heteroscedasticitate condiționată autoregresivă generalizată (GARCH), GARCH(1,1)-t.

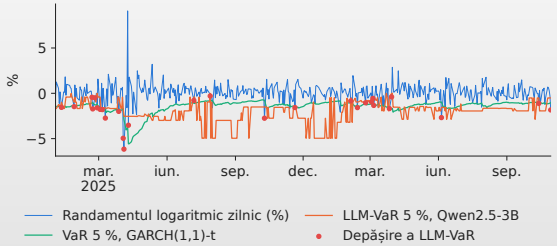
- **În practică:** un model deschis de 3B pe laptop este cel mai ieftin test al ideii
  - Lucrarea a folosit modele comerciale mari.

(Pele et al., 2026; Kupiec, 1995)

Exercițiu practic: notebook §9

LLM-VaR cu un model deschis de 3B pe S&P 500

- **Pe scurt:** VaR de 5 % a modelului de limbaj comparată cu GARCH; punctele roșii marchează zilele în care LLM-VaR a fost depășită.



Sursa: Yahoo Finance, cotații zilnice. Date: 2 ianuarie 2025 - 30 octombrie 2026, S&P 500, zile de tranzacționare.

- 5 %: LLM 28 depășiri, GARCH 35, HS 23, așteptate 23,0.
- VaR medie de 5 %: LLM −1,99 %, GARCH −1,42 %
  - Corelația lor 0,05: modelul mic acoperă riscul fiind prudent, nu urmărind volatilitatea.

NN_L12_llm_var

Testarea retrospectivă a LLM-VaR față de reperele clasice

Model	Dep. 1 %	Rata (%)	p_{uc}	Dep. 5 %	Rata (%)	p_{uc}	VaR 5 % medie
LLM-VaR (Qwen2.5-3B)	4	0,87	0,777	28	6,10	0,295	−1,99
HS (250 de zile)	6	1,31	0,528	23	5,01	0,991	−1,56
GARCH(1,1)-t	9	1,96	0,067	35	7,63	0,016	−1,42

- Depășiri așteptate: 4,6 la 1 % și 23,0 la 5 %; valoare p verde: acoperirea nu este respinsă la 10 %.
- Răspunsuri citite ca număr: 100,0 % la 5 %, 100,0 % la 1 %; un răspuns ilizibil păstrează valoarea din ziua precedentă.
- Dimensiunea modelului contează: studiul publicat a evaluat modele mult mai mari; rezultatele de aici descriu doar un model de 3B.

(calcul propriu; Pele et al., 2026)

Exercițiu practic: notebook §9

Cuprins

- Prezicerea următorului token și modelele lingvistice mari
- Preantrenare, ajustare pe instrucțiuni și aliniere
- Evaluarea formulării instrucțiunilor pe date etichetate
- Generare augmentată prin regăsire pe discursurile BCE
- Modelele de limbaj ca modele de risc: LLM-VaR
- Riscuri și guvernanță

Șase riscuri ale LLM într-o instituție financiară și testul pentru fiecare

Risc	Observat în acest curs	Control
Halucinația	86,2 % nume greșite fără documente	RAG cu surse; abținerea permisă
Scurgerea datelor	perplexitate mai mică pe discursurile vechi	testare pe texte ulterioare limitei datelor
Sensibilitatea la prompt	12,9 puncte macro-F1 între formulări	selecție pe validare, fixare, versionare
Calibrarea	ECE 4,7 % zero-shot, 5,4 % TF-IDF	scalarea temperaturii, praguri stabilite pe validare
Costul și latența	389 ms față de 2,50 ms pe propoziție	cel mai mic model care trece testul
Protecția datelor	documente ale clienților în prompturi	modele locale; obligațiile din Regulamentul Uniunii Europene (UE) privind inteligența artificială (AI) pentru utilizările cu risc ridicat, precum scoringul de credit

⊡ Regulamentul (UE) 2024/1689 privind inteligența artificială.

Test 1: temperatura

- Clasificați 10.000 de știri cu un LLM și obțineți etichete diferite la două rulări pentru aceeași știre
  - Ce schimbați mai întîi?

Răspuns

- Setați temperatura la 0 (sau citiți probabilitățile etichetelor): eșantionarea la $T > 0$ este aleatoare prin construcție.

Test 2: evaluarea

- Un furnizor raportează o acuratețe de 95 % a LLM-ului său pe Financial PhraseBank
  - Ce două întrebări puneți?

Răspuns

- Pe ce subset de acord (cel unanim este cel mai ușor) și a fost PhraseBank în datele de antrenare (publicat în 2014, copiat pe scară largă)? Cereți și macro-F1: 61,4 % dintre propoziții sînt neutre.

Test 3: RAG

- Asistentul RAG răspunde greșit, deși documentul corect există în corpus
  - Unde căutați mai întîi?

Răspuns

- La modulul de regăsire: verificați dacă documentul este printre primele k (recall@k)
  - Abia apoi la prompt și la model.

Recapitulare

- Un LLM este un model al următorului token
  - Perplexitatea, temperatura și tokenii sînt mărimile sale de bază.
- Un prompt este un model: evaluați-l pe etichete separate, cu intervale, rulări repetate și reformulări.
- Cu etichete, modelele mici antrenate și embedding-urile LLM sînt greu de depășit
  - Fără etichete, prompturile oferă un început solid.
- RAG transformă răspunsurile inventate în răspunsuri cu surse
  - Evaluați întîi modulul de regăsire.
- Urmează: modelele fundație pentru serii de timp din energie și finanțe (cursul 13).

Cercetările grupului nostru

- **LLM-VaR și, pentru pierderea așteptată în coada distribuției (ES), LLM-ES**: LLM-uri generale estimează, fără reantrenare, ambele măsuri de risc dintr-un prompt structurat cu randamentele trecute
  - Testate retrospectiv față de modelele clasice de risc (Pele, Bolovăneanu, Lin, Ren, Ginavar, Spilak, Andrei, Toma, Lessmann & Härdle, 2026, Expert Systems with Applications).
- **Prognoza acțiunilor cu LLM**: LLaMA 3.3 combină în prompt istoricul prețurilor cu sentimentul știrilor și depășește LLaMA 3.1 și un model autoregresiv integrat cu medie mobilă (Tak & Pele, 2025, PICBE).
- **Narațiunile băncii centrale**: un indice de sentiment construit cu un model transformer din minutele Băncii Naționale a României îmbunătățește prognoza inflației (Amza, Bolovăneanu & Pele, 2026, PICBE) — același tip de text ca discursurile BCE din acest curs.
- Rezultatele sînt preluate din rezumatele lucrărilor; cifrele de pe celelalte slide-uri provin din datele noastre.

(Academia de Studii Economice din București, ASE; Academia Română, Institutul de Prognoză Economică, IPE)

Explorați vizual

- **Transformer Explainer**: temperatura din acest curs pe un GPT-2 care rulează în browser: la temperatură mică o continuare ia aproape toată probabilitatea, la temperatură mare tokenii improbabili devin plauzibili.
- **Mixture of experts (M. Grootendorst)**: un router trimite fiecare token la cîțiva experți, astfel că modelul poate avea mulți parametri, dar folosește doar o parte dintre ei pentru fiecare token.
- **LLM Inference Handbook**: ce se întîmplă între prompt și răspuns: procesarea promptului (prefill), decodarea token cu token, memoria cache pentru keys și values și timpul pînă la primul token.
- **LLM Visualization (B. Bycroft)**: aceeași arhitectură la mai multe dimensiuni, de la un GPT mic la GPT-2 și GPT-3.

Lecturi și resurse suplimentare

- **Hugging Face**, curs despre LLM: huggingface.co/learn/llm-course.
- **Qwen2.5**, modele și raport: huggingface.co/Qwen.
- **Ghid de formulare a prompturilor**: promptingguide.ai.
- **Setul de discursuri BCE**: ecb.europa.eu.
- **Quantinar**: The Next Word Problem: quantinar.com.
- **Acest curs**: notebook, seminar și Quantlets — github.com/danpele/NEURAL_BIZ.

Bibliografie (1/2)

- Amza, A., Bolovăneanu, V. & Pele, D. T. (2026). Inflation forecasting with monetary policy narratives: Evidence from an emerging economy. *Proceedings of the International Conference on Business Excellence*, 20(1), 75–83. doi:10.2478/picbe-2026-0012
- Brown, T. B. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33. arXiv:2005.14165
- Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330. arXiv:1706.04599
- Hoffmann, J. et al. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35. arXiv:2203.15556
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. arXiv:2106.09685
- Kaplan, J. et al. (2020). Scaling laws for neural language models. arXiv:2001.08361
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2), 73–84. doi:10.3905/jod.1995.407942
- Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33. arXiv:2005.11401
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J. & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782–796. doi:10.1002/asi.23062
- Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35. arXiv:2203.02155

Bibliografie (2/2)

- Pele, D. T., Bolovăneanu, V., Lin, M.-B., Ren, R., Ginavar, A. T., Spilak, B., Andrei, A.-V., Toma, F.-M., Lessmann, S. & Härdle, W. K. (2026). In the beginning was the Word: LLM-VaR and LLM-ES. *Expert Systems with Applications*, 295, 128676. doi:10.1016/j.eswa.2025.128676
- Qwen Team (2024). Qwen2.5 technical report. arXiv:2412.15115
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D. & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36. arXiv:2305.18290
- Robertson, S. & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. doi:10.1561/1500000019
- Sennrich, R., Haddow, B. & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the ACL*, 1715–1725. doi:10.18653/v1/P16-1162
- Tak, R. & Pele, D. T. (2025). LLM-driven stock prediction: Capturing market trends with LLaMA. *Proceedings of the International Conference on Business Excellence*, 19(1), 529–540. doi:10.2478/picbe-2025-0043
- Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35. arXiv:2201.11903
- Xiao, S., Liu, Z., Zhang, P. & Muennighoff, N. (2023). C-Pack: Packaged resources to advance general Chinese embedding (BAAI general embedding). arXiv:2309.07597
- Zhao, Z., Wallace, E., Feng, S., Klein, D. & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *Proceedings of the 38th International Conference on Machine Learning*, 12697–12706. arXiv:2102.09690

Anexa A: promptul folosit pentru clasificarea zero-shot

- **Sistem**
  - "You are a careful financial analyst."
- **Utilizator**
  - "Classify the sentiment of this sentence from financial news, from the point of view of an investor in the company. Answer with one word: positive, negative or neutral. Sentence: ... Answer:"
- **Few-shot**
  - La fel, cu perechi „Sentence: ... Answer: etichetă" înaintea intrării; exemple echilibrate pe clase și amestecate
- **Chain-of-thought**
  - "... First explain your reasoning in at most two short sentences, then give the final answer on a new line as 'Label: <positive|negative|neutral>'."

Anexa B: detalii de implementare

- Modele în float16, pe procesorul grafic al laptopului (PyTorch)
  - Biblioteca Hugging Face transformers
  - Scikit-learn pentru TF-IDF, regresia logistică și bootstrap.
- Valorile generatorului aleator: 20261108 pentru împărțiri și eșantioane
  - Exemplele few-shot cu valorile 20261108, $+1$ și $+2$.
- Întreaga analiză a durat 2,2 minute la ultima rulare (ieșirile modelelor stocate)
  - Prima rulare, aproximativ 40 de minute.
- Predicțiile pentru fiecare element sînt publicate împreună cu Quantlet-urile, astfel încît fiecare cifră poate fi recalculată fără placă grafică.