

Mentalizáció és nagy nyelvi modellek

Összefoglalás: A mentalizáció képessége lehetővé teszi mások és saját mentális állapotaink – például a gondolatok, érzelmek, vágyak és szándékok – megértését és értelmezését. Ezért a tudatelmélet (Theory of Mind, ToM), a kísérleti, fejlődés és klinikai pszichológia kiemelten vizsgált területe. Az emberi interakciókban segít előre jelezni és felfogni mások viselkedését, továbbá kulcsszerepet játszik az empátia és a társas kapcsolatok kialakításában, fenntartásában. Ezek a képességek a gyermekkorban fokozatosan alakulnak ki, de a felnőttkorban is további finomodásuk figyelhető meg. A képesség hiánya vagy zavara olyan pszichológiai problémákhoz, mint a borderline személyiségzavar vagy autizmus, vezethet. A tudatelmélet működésének vizsgálatához számos mérőeljárást, feladatt és kérdőívet fejlesztettek ki. A legújabb mesterséges intelligencia rendszerek, a nagy nyelvi modellek (LLMs), bizonyos méret felett bizonyos területeken néhány, a humán képességet megközelítő vagy fölülmúló teljesítményekre képesek. Érdekes kérdés, hogy a szociális kogníció területén, hogyan teljesítenek. Az előadás összefoglalja a mentalizáció alapjait és a legfontosabb mérési módszereket, valamint áttekinti a nyelvi modellek e feladatokban elért eredményeit és korlátait a szakirodalom alapján. **Kulcsszavak:** Mentalizáció, kognitív fejlődés, LLMs, mesterséges intelligencia.

Abstract: The ability to mentalize allows us to understand and interpret the mental states of others and ourselves, such as thoughts, emotions, desires, and intentions. Therefore, Theory of Mind (ToM) is an intensively studied field of experimental, developmental, and clinical psychology. It helps us predict and understand the behavior of others in human interactions, and it plays a key role in the development and maintenance of empathy and social relationships. These abilities gradually develop in childhood, but their further refinement can be observed in adulthood. The lack or damage of this ability can lead to psychological problems such as borderline personality dis-

* *Dunaiújvárosi Egyetem,
Tanárképző Központ
E-mail: juhaszle@uniduna.hu*

[1] Gál Z. (2015): *A tudatelmélet életkori különbségei, kapcsolata a munkamemória kapacitással és a társas pozícióval*. PhD-értékezés. Szeged.

[2] Csibra G.–Gergely G. (2002): A naiv tudatelmélet az evolúciós lélektan szempontjából. *Magyar Tudomány*, 1., pp. 56–63.

[3] Kovács, Á. M.–Téglás, E. (2011): A „mentalizáció” eredete. *Korunk*, 22., (9.), pp. 107–115.

[4] Gál Z. (2015): A tudatelmélet életkori változásainak és szerepének áttekintése óvodáskortól fiatal felnőttkorig. *Iskolakultúra*, 25., (5–6.), pp. 59–73.

[5] Shamay-Tsoory, S. G.–Shur, S.–Barzcai-Goodman, L.–Medlovich, S.–Harari, H.–Levkovitz, Y. (2007): Dissociation of cognitive from affective components of theory of mind in schizotypy. *Psychiatry Research*, 149., pp. 11–23. <https://doi.org/10.1016/j.psychres.2005.10.018>

order or autism. A number of measurement procedures, tasks, and questionnaires have been developed to examine the functioning of theory of mind.

The latest artificial intelligence systems, large language models (LLMs), are capable of some performances approaching or exceeding human capabilities in certain areas above a certain size. It is an interesting question how they perform in the field of social cognition. The presentation summarizes the basics of mentalization and the most important measurement methods, and reviews the achievements and limitations of language models in these tasks based on the literature.

Keywords: Mentalization, cognitive development, LLMs, artificial intelligence.

Humán mentalizáció

A MENTALIZÁCIÓ KÉPESSÉGE

A mentalizáció mások és magunk megértésének az alapja. Olyan – legalább is fejlettebb formájában – kizárólagos emberi képesség, amely alapja a szociális kogníciónak, a társas kommunikációnak az emberi interakcióknak, a társas tudásnak és egyéb szociális készségeknek. A magyar fejlődéslélektani és klinikai pszichológiai szakirodalomban a mentalizációnak számos alternatív elnevezésével találkozhatunk: tudatelmélet (Theory of Mind, ToM), elmeteória, elmeolvasás, és naiv pszichológia [1]. Általában a társas életben eligazodást segítő kognitív kompetenciákat értik rajta, szűkebb értelemben azt a képességet, hogy másoknak – esetleg magunknak – mentális állapotokat, azaz tudást, szándékot, vágyakat és vélekedéseket tulajdonítunk [2, 3].

A mentalizáció lehetővé teszi, hogy a társas interakciókban a többi fél viselkedését előre jelezzük, értelmezzük és megmagyarázzuk. Ennek függvényében a saját viselkedésünket tudjuk módosítani, a helyzethez adaptálni. Segíti a kommunikációkat, szerepe van a kooperációban, a viselkedés irányításában, befolyásolásában, továbbá az interakció kezdeményezésében és fenntartásában [4].

Shamay-Tsoory [5] a mentalizáció két komponensét különbözteti meg. A kognitív tudatelmélet a mások mentális állapotairól való racionális gondolkodást, következtetéseket jelent. Az affektív tudatelmélet a mások érzéseiről azaz az emocionális állapotokról hozott következtetéseket jelent. A szerző empátiafejlődési modelljében a kognitív tudatelmélet kialakulása az affektív következtetések alapja, míg ez utóbbi

az érzelmi empátia képességének fontos összetevője.

A mentalizáció elméleteinek két karakterisztikus csoportja van, amelyek a nyelvészethez hasonlóan a kompetencia-performancia kettősség alapján határozzák meg önmagukat (Kovács – Téglás, 2011). A kompetenciaelmélet (Perner 1991) szerint a gyerekek mentalizációs teljesítményében minőségi változás, ún. kompetenciaugrás figyelhető meg az óvodás kor kezdetén. A 4 éves gyerekek egyes, a későbbiekben áttekintett tudatelméleti feladatokat (pl. váratlan áthelyezés) helyesen oldják meg, míg a fiatalabbak rendre elbuknak ezen. A kompetenciaelmélet szerint a mentalizáció tudatos, szándékos erőfeszítés eredménye. A performancia-elmélet velünk született, automatikus készségnek tartja a tudatelméletet. Az ezen nézetet képviselő kutatók szerint a tudatelméleti feladatokon azért sikertelenek a 4 évesnél fiatalabbak, mert az eredményes végrehajtáshoz más kognitív képességek és a nyelv megfelelő fejlettségére is szükség van. Így a feladatokon nyújtott teljesítmény nem a mentalizáció szintjét tükrözi. Az integratív „kettős folyamat” elmélet szerint – hasonlóan Kahneman [6] gyors és lassú kognitív működéseéhez – a két rendszer függetlenül működik és fejlődik. A *Type I* folyamatok korán kifejlődnek, tudattalanul, implicit módon működnek, viszonylag elemi mentalizációs képességet jelentenek. A *Type II* folyamatok már fejlett kifinomult mentalizációt jelentenek. A nyelv és a munkamemória, a végrehajtó kognitív funkciók fejlettségétől szorosan függnek, folyamataikban azokra támaszkodnak.

[6] Kahneman, D. (2013): *Gyors és lassú gondolkodás*. Budapest: HVG Könyvek..

[7] Baron-Cohen, S.–Leslie, A.–Frith, U. (1985): Does the autistic child have a „theory of mind”? *Cognition*, 21., (1.), pp. 37–46.

ToM TESZTEK

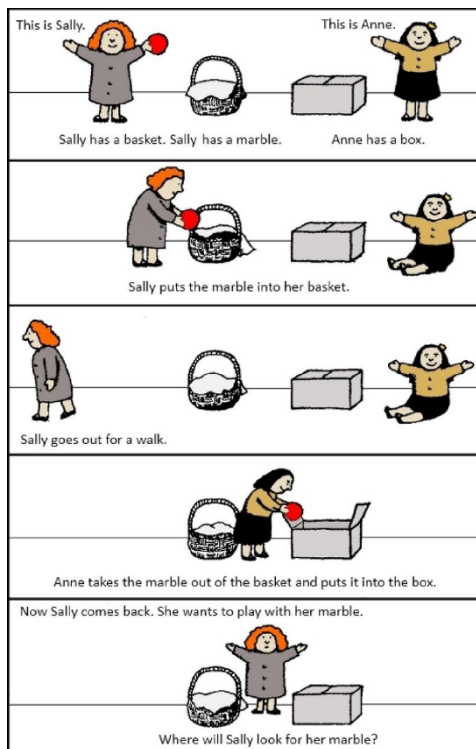
Téves vélekedés teszt (elsőfokú téves vélekedés)

Ennek megértéséhez az szükséges, hogy másoknak lehetnek olyan elképzelései a világról, amelyek nem helyesek.

Tipikus példája a Sally–Anne teszt, a váratlan áthelyezés feladat [7]. Ebben a vizsgálatban részt vevő (gyerek) egy történetet lát rajzon, vagy esetleg elbábozva, illetve a vizsgálatvezető által kommentálva. A történet résztvevői Sally és Anne egy kosárral illetve egy dobozzal. Sally egy üveggolyót tesz a kosárba, majd elhagyja a helyszínt. Ezután Anne áttesz a golyót a dobozba. A zárórészben Sally visszatér, és a vizsgált személyt megkérdezik, hogy Sally hol fogja keresni az üveggolyót. A tesztet sikeresen teljesítők az eredeti helyet, a kosarat kell válaszolnia.

[4] Gál Z. (2015): A tudatelmélet életkori változásainak és szerepének áttekintése óvodáskortól fiatal felnőttkorig. *Iskolakultúra*, 25., (5–6.), pp. 59–73.

1. ábra. Sally-Anne test



Másodfokú, verbális hamis vélekedés-tulajdonítási feladat

A beágyazott mentális feladatok megértésén alapszik. A vizsgálatban a megoldáshoz másoknak egy harmadik személyre vonatkozó gondolatait kell kikövetkeztetnie. Egy példa [4]:

- Elmesélek egy történetet. Egy faluban játszódik. Nézd, ezek itt a házak, ez itt a templom. A falunak van egy parkja is, ez itt. Két gyerek játszik benne. Ez Jancsi, ez pedig itt Mari.

- Itt vannak mind a ketten a parkban. Amott jön a fagyaltárus. Jancsi szeretne fagyit venni, de nincs nála pénz. Nagyon szomorú emiatt. – Ne aggódj – mondja neki a fagyaltárus, – menj csak haza, hozzál pénzt, és később is vehetsz fagyit. Egész délután itt leszek a parkban... – Nagyon jó – mondja Jancsi, – délután visszajövök fagyit venni.
- Jancsi tehát hazamegy. Itt lakik, ebben a házban. A fagyaltárus közben azt mondja magában: – Inkább átmegyek a templomhoz, talán ott több fagyit tudok eladni.
- A fagyaltárus tehát átmegy a templom elé a kocsijával. Útközben elmegy Jancsiék háza előtt. Jancsi észreveszi őt, kimegy hozzá, és azt kérdezi: – Hová mégy? A fagyaltárus azt mondja neki: – Megyek a templom elé, ott fogok fagyit árulni. El is megy a templomhoz. Egy kicsit később pedig Jancsi is utána megy.
- Közben Mari is hazamegy. Itt lakik, ebben a házban. Aztán elindul megkeresni Jancsit, és elmegy Jancsiék házához. Bekopog az ajtón, kijön Jancsi anyukája, és Mari megkérdezi: – Itthon van Jancsi? – Nincsen, elment fagyit venni.

Tesztkérdés: Mit gondol Mari, hová ment Jancsi fagyit venni?

Indoklás: Miért gondolja ezt Mari?

Irónia feladat

Az irónia megértéséhez másodfokú szándéktulajdonításra van szükség. Egy jellemző irónia feladat [8].

Kati és Éva jó barátnők. Gyakran járnak el együtt. Amikor ma találkoztak, esni kezdett az eső. Ezért moziba mentek. Ahogy vége lett a filmnek és felálltak, Kati így szólt: Na, mehetünk vissza a napsütésbe!

Kati tényleg úgy értette, hogy a napsütésbe mehetnek vissza? Amikor Kati ezt mondta, boldog volt? Miért mondta ezt Kati? Valóban sütött a nap, mikor Kati ezt mondta?

Faux pas tesztek

Ennek megértésekor a vizsgálati személynek fel kell ismernie, hogy az elszólás véletlenül történt, valamint azt is, hogy az elszólás megbánthatott egy másik személyt. Mindkét személy mentális állapotait figyelembe kell vennie [4].

[4] Gál Z. (2015): A tudatelmélet életkori változásainak és szerepének áttekintése óvodáskortól fiatal felnőttkorig. *Iskolakultúra*, 25., (5–6.), pp. 59–73.

[8] Györi M. (2009): A tudatelméleti képesség változatossága autizmusban – és implikációi az atipikus megismerésre és tanulásra nézve. *Gyógypedagógiai Szemle*, 37., (2–3.), pp. 96–111.

[9] Ivády R. E.–Takács B.–Pléh Cs. (2007): Tudatelmélet és idegen nyelvelsajátítás – valódi kapcsolat vagy városi legenda? In: G. Kampis–K. Mund: *Tudat és elme*. Budapest: Typotex, pp. 59–74.

[10] Baron-Cohen, S.–Wheelwright, S.–Hill, J.–Raste, Y.–Plumb, I. (2001): The Reading the Mind in the Eyes Test revised Version: A study with normal adults, and adult with Asperger Syndrome or high functioning autism. *Journal of Child Psychology and Psychiatry*, 42., (2.), pp. 241–251.

[11] Rakoczy, H. (2022): Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, 1., pp. 223–235.

Az egész osztály benevezett az iskolai mesemondó versenyre. Emma nagyon szeretett volna nyerni. Amíg ő nem volt ott az iskolában, addig kihirdették a verseny eredményét: Aliz nyert. Másnap, amikor Aliz meglátta Emmát, így szólt: „Sajnálom, hogy nem díjazták a mesédet”. „Mire gondolsz?” – kérdezte Emma. „Jaj, semmi” – mondta Aliz.

Kérdések: Mondott bárki valami olyat, amit nem kellett volna, vagy valami olyat, ami nem volt odaillő?

Szemekből Olvasás Teszt

Magyarul [9], eredetileg Reading the Mind in the Eyes Test [10]. A vizsgált személy egy szempár alapján próbálja azonosítani, hogy a szempár tulajdonosa mit érezhetett vagy gondolt.

2.ábra. Szemekből olvasás teszt

irigy

pánikba esett



Forrás: [9].

A TUDATELMÉLET FEJLŐDÉSE

Rakoczy [11] a mentalizáció fejlődésében két „forradalmat” állapít meg. 9 hónapos korban az alap mentalizációs képességek jelennek meg, mint mások céljának azonosítása, és annak figyelembevétele, hogy mit észlel a másik (perception-goal psychology). A második forradalom 4 év körül következik be, ez már a finomabb metareprezentációs mentalizáció megjelenése (belief-desire psychology). Ennek folyamányaként a gyerekek már a téves vélekedés tesztet helyesen oldják meg.

A fejlődés alapvető stádiumai [4, 12]:

- 3–4 éves gyerekek képesek megérteni a *hamis vélekedést*, vagyis azt, hogy egy személynek a valóságtól eltérő gondolatai vannak a világról (mint például a Sally–Anne tesztben).
- 6 éves kor körül a gyerekek megértik egy személy gondolatait egy másik személy gondolatairól, vagyis képesek megérteni a *beágyazott mentális állapotokat*.
- 8 éves kor körül képesek megérteni a *metaforát és iróniát*, ez pedig azt a képességek mozgósítja, hogy megértse egy kijelentés nem szó szerinti jelentését, ezentúl képesek elkülöníteni a viccet a hazugságtól.
- A 9–11 éves gyerekek megértik a *faux pas*-t. A faux pas megértése már egy fejlettebb tudatelméletet igényel, hiszen két mentális állapot egyidejű megértésére van szükség: a faux pas-t elkövető személy nézőpontját, aki nincs tudatában az elszólásnak és a másik személy perspektíváját, akit megbántott vagy megbánthatott az elszólás.

A MENTALIZÁCIÓ TÁRSAS/KOGNITÍV KORRELÁTUMAI

Gál [1] összefoglalójában számos tényező mentalizációval való együttjárását ismertette, habár ezeknél az oki hatások természete nem mindig tisztázott [12].

Fejlettebb nyelvi szinten lévő gyerekek mentalizációja gyorsabban fejlődik: magasabb a teljesítményük a hamis vélekedés-teszteken.

Minél gyakoribb a gondozóval, testvérekkel való társalgásban mentális állapotokra való utalás, később (7 hónap múlva) magasabb a ToM-teljesítmény (~40 hónap). A család anyagi helyzete, szülők iskolázottsága, testvérek száma és a gyermek általános kognitív képességei pozitívan korrelálnak a ToM-mal.

A fegyelmező, túlságosan hatalmat gyakorló szülői nevelési stílus (kiabál, utasít) esetén viszont negatív a kapcsolat. Elutasított, ellentmondásos/ambivalens szociometria státuszú 8–9 éveseknél rosszabb mentalizáció figyelhető meg. Az 5 éves korban rosszabb ToM-teljesítményű gyerekek gyakrabban vesznek részt bullyingban serdülőkorban. De a zaklatásban „hangadók” ToM-értékei magasabbak, mint az áldozatoké vagy az őt követőké [13].

[1] Gál Z. (2015): *A tudatelmélet életkori különbségei, kapcsolata a munkamemória kapacitással és a társas pozícióval*. PhD-értekezés. Szeged.

[4] Gál Z. (2015): A tudatelmélet életkori változásainak és szerepének áttekintése óvodáskortól fiatal felnőttkorig. *Iskolakultúra*, 25., (5–6.), pp. 59–73.

[12] Fodor A. (2020): A mentalizációról kicsit mélyebben. In: *Békés Iskolák*. <https://www.bekesiskolak.hu/a-mentalizacio-rol-kicsit-melyebben/>

[13] Fricz B.–Gál Z.–Kasik L. (2017): A kortársbántalmazás kapcsolata a tudatelmélettel. *Iskolakultúra*, 27., (1–2.), pp. 98–113. doi:<https://doi.org/10.17543/ISKKULT.2017.1-12.98>

- [14] Wiesmann, C. G.–Friederici, A. D.–Singer, T.–Steinbeis, N. (2020): Two systems for thinking about others' thoughts in the developing brain. *Proc. Natl Acad. Sci USA*, 117., pp. 6928–6935.
- [15] Frith U. (1991): *Autizmus. A rejtély nyomában. Budapest: Kapocs Autizmus Alapítvány.*
- [16] Frith, C. D. (1992): *The Cognitive Neuropsychology of Schizophrenia*. Hove: LEA. Taylor & Francis.
- [17] Egyed K.–Juhász, L. Zs. (2001): A szkizofrénia tudatelméleti kutatása. *Magyar Pszichológiai Szemle*, 55., (4.), pp. 517–526. DOI:10.1556/MPSzle.55.2000.4.10
- [18] Sprong, M.–Schothorst, P.–Vos, E.–Van Engeland, H. (2007): The theory mind in schizophrenia. *British Journal of Psychiatry*, 191., (1.), pp. 5–13. doi:10.1192/bjpp.107.035899
- [19] Song, Z.–Corcoran, R.–Daly, N.–Abu-Akel, A.–Gillespie, S. M. (2023): Psychopathic traits and theory of mind task performance: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, 151., 105231. <https://doi.org/10.1016%2Fj.neubiorev.2023.105231>
- [20] Uekermann, J.–Daum, I. (2008): Social cognition in alcoholism: a link to prefrontal cortex dysfunction? *Addiction*, 103., (5.), pp. 726–735.
- [21] Lee, L.–Harkness, K. L.–Sabbagh, M. A.–Jacobson, J. A. (2005): Mental state decoding abilities in clinical depression. *Journal of Affective Disorders*, 86., (2–3.), pp. 247–58.
- [22] Preißler, S.–Dziobek, I.–Ritter, K.–Heekeren, H. R.–Roepke, S. (2010): Social Cognition in Borderline Personality Disorder: Evidence for Disturbed Recognition of the Emotions, Thoughts, and Intentions of others. *Frontiers in Behavioral Neuroscience*, 4., <https://doi.org/10.3389/fnbeh.2010.00182>
- [23] Fónagy, P.–Bateman, A. W. (2006): Mechanisms of change in mentalization-based treatment of BPD. *Journal of Clinical Psychology*, 62., pp. 411–430. DOI: 10.1002/jclp.20241

NEUROBIOLÓGIA ÉS PSZICHOPATOLÓGIA

A szociális kognícióban számos agyterület szerepet kap. A prefrontális, parietális és temporális területek egyes részei, amelyek a tudatelméleti kognitív működésekben jelentősebben részt vesznek. Az elülső prefrontális és a hátrébb elhelyezkedő lebenyek kapcsolatainak kialakulását feltételezik a 4. életévben megjelenő mentalizációs teljesítményerősödés mögött [14]. A mentalizáció különböző fokú sérülését, zavarát mutatták ki számos mentális probléma, illetve atipikus neurális állapot esetén. Az autizmus spektrumhoz tartozó zavarok tüneteinek egy részét már korábban is az állapottal élők különböző fokú „elmevakására” vezették vissza [15]. Schizophrenia. C. Frith a schizophrenia számos tünete alapján a mentalizációs képesség zavarát feltételezte [16, 17]. Számos későbbi vizsgálat tapasztalata megerősítette Frith hipotézisét [18]. Vannak adatok, habár nem egyértelműek, arra vonatkozóan is, hogy a szocio-, és pszichopátiában személyek teljesítménye is rosszabb egyes mentalizációs feladatokban [19]. Alkoholisták esetén feltehetően a prefrontális területet érintő kognitív funkciók sérülése miatt a tudatelméleti feladatokban is csökken a teljesítmény [20]. A normálistól eltérő mentalizációs mintázatot figyeltek meg a depresszió akut fázisában is [21]. A mentalizáció problémái a borderline személyiségzavarban szenvedők tüneteinek kialakulásában is központi szerepet tölthetnek [22]. A dezorganizált kötődésű személy mentalizációja rosszabb, míg a biztonságosabban kötődőké komplexebb és fejlettebb. Peter Fónagy és munkatársai ennek alapján a borderline állapot kezelésére mentalizáció alapú terápiát fejlesztett ki [23],

ami a pszichodinamikus szemlélet, a kognitív viselkedésterápia, az ökológia és a rendszerelmélet elemeit ötvözi. Az eljárás célja a mentalizációs kapacitás növelése, ezáltal az affektív szabályozás javítása, a szuicid és önsértési késztetések csökkentése, az interperszonális kapcsolatok megerősítése [12]. Egyéni és csoportterápiás formában is alkalmazzák.

Nagy nyelvi modellek

NYELVI MODELLEK

2024-ben a legfejlettebb mesterségesintelligencia-architektúrák (Artificial Intelligence, AI) a Nagy Nyelvi modellek (Large Language Models, LLMs). Ezek bonyolult, komplex mesterséges neurális hálók, amelyek funkciója a nyelvi modellezés. A nyelvi modellek tulajdonképpen olyan függvények (komputációs modellek), amely szósorozatokhoz valószínűségi értéket rendelnek hozzá; valószínűség-eloszlások szövegek felett. Mennyire valószínű egy szósorozat egy adott nyelvben? [24]

Például alkalmasak annak megállapítására, hogy:

- $P(„a quick brown dog”) > P(„dog quick a brown”)?$, vagy
- $P(„barks”|„a quick brown dog”) > P(„meows”|„a quick brown dog”)?$

A nyelvi modellek segítségével a szósorozat következő eleme megjósolható, generálható. Két fő alap feladattípusban használható:

- *Generatív*: megtalálni a következő szót.
- *Klasszifikáció*: megtalálni a legmegfelelőbb címkét vagy választ egy szósorozatra.

Természetes nyelvfeldolgozásnak (Natural Language Processing, NLP) nevezik a mesterséges intelligencia azon részterületét, amely elsősorban az emberi nyelv feldolgozásának és generálásának számítógépes lehetőségeivel és alkalmazásaival foglalkozik. Az NLP története alatt számos nyelvmodell-architektúrát dolgoztak ki, az informatika általános fejlődésének megfelelően. Kezdetben elsősorban a szósorozatok statisztikai előfordulásán alapuló ún. „bag of words”, „n-gram” modellek majd a neurális hálózatok újrafelfedezésével az ún. rekurrens hálózatok különböző típusai segítségével valósítottak meg nyelvi modelleket [24].

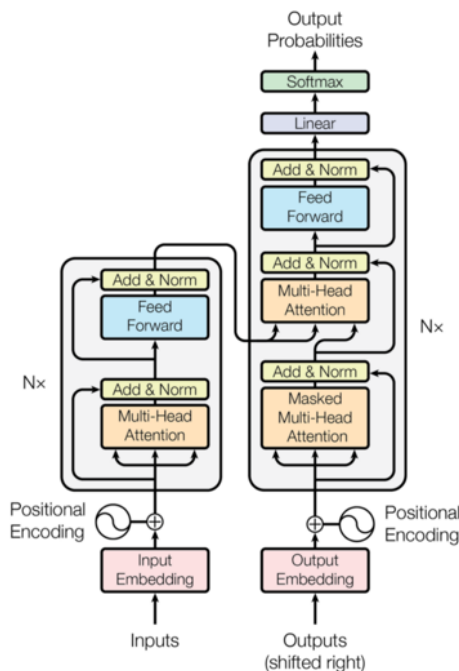
[12] Fodor A. (2020): A mentalizációról kicsit mélyebben. In: *Békés Iskolák*. <https://www.bekesiskolak.hu/a-mentalizacio-rol-kicsit-melyebben/>

[24] Jurafsky, D.–Martin, J. H. (2023): *Speech and Language Processing* (3rd ed. draft). Forrás: <https://web.stanford.edu/~jurafsky/slp3/>

[25] Waswani, A.–Shazeer, N.–Parmar, N.–Uszkoreit, J.–Jones, L.–Gomez, A. N.–Polosukhin, I. (2017): *Attention Is All You Need* (arXiv: 1706.03762. kötet).
Letöltés dátuma: 2023. szeptember 17.
forrás: <https://arxiv.org/abs/1706.03762>

A Transformer (transzformátor) hálózati architektúra 2019-es kifejlesztésével [25] az újonnan létrehozott nyelvi modellek mindegyike ezt az alapstruktúrát használja ki.

3. ábra. Transformer architektúra



Forrás: [25].

TRANSFORMER ALAPÚ NYELVI MODELLEK

A Transformer alapmodell egy többszörösen összetett encoder-decoder hálózat. Az encoder (kódoló) rész kimenete a bemeneti tokenek (szavak, részsavak) kontextusfüggő reprezentációja. A reprezentáció létrehozásának központi eleme az ún. „self-attention” mechanizmus.

Ez az input különböző elemei közötti kapcsolatokat és dependenciákat fedezi fel és képezi le. Transformer-modellek esetén a szavak (tokenek) reprezentációja függ a szavak kontextusától, azaz attól, hogy milyen nyelvi környezetben fordultak elő, így az egyes itemeknek többszörös reprezentációja fog létrejönni. A decoder (visszakódoló rész) bemenete a sajátmaga által korábban generált elemek és az encoder kontextus-vektorai. A decoder rész is a self-attention mechanizmust használja. Az encoder bemenetét pedig az ún. „cross-attention” mechanizmus segítségével használja fel. A transformer-modellek több rétegből és több párhuzamos feldolgozó fejből állnak. Nagyon sok paraméterük van, már az alapmodellnek is 65 milliárd. 8 figyelmi fejet és 6 figyelmi réteget tartalmazott, a tokeneket reprezentáló vektorok hossza 512 volt [25].

A transformer-alapú nagy nyelvi modellek felépítése alapján 3 fő típust lehet megkülönböztetni [26]:

- *Csak encoder.* Ezek csak encoder modult tartalmaznak. Tréningjük auto-enkódolás, vagyis nem felügyelt tanulásról van szó. A tréning alatt maszkolt bemeneteket kapnak, a maszk helyén álló elemet kell a modellnek megtanulnia. Ezen modellek különösen alkalmasak a bemenet megértésére, arról nyelvi leképezés kialakítására. Szövegek, szavak kategorizálására, információkeresésre különösen alkalmasak. Jellegzetes képviselőjük a BERT (Bidirectional Encoder Representation from Transformer). Maszkolt Nyelvi model. A másik tréning feladata: következő mondat predikció. Jellemzői: 12 fej, 24 réteg, 340 millió paraméter.
- *Csak decoder.* Autoregresszív modellek: egyik bemenetük az előző generált token. Az aktuális porciók mögötti bemenetek maszkolva vannak. Elsősorban szöveg, kérdés-válasz NLP feladatokra használják őket. Jellegzetes képviselőik a GPT (Generative Pre-trained Transformer) család tagjai, mint a GPT-4 vagy Chat-GPT. Ezen modellek paraméterezése nem nyilvános, de a GPT-4 feltehetően 192 fejet, 120 réteget és összesen 1760 milliárd paramétert tartalmaz.
- *Encoder-decoder.* Az eredeti Transformer-felépítésnek megfelelően mindkét modult magukban foglalják. Ezért egyszerre auto-enkódolók és auto-regresszívek. Ún. Seq2seq architektúrájuk van; azaz általában bemeneti szövegből új szöveget állítanak elő. Ilyen az automatikus gépi fordítás vagy a szöveg kivonatolás. Ilyen felépítésű modell a BART (Bidirectional and Auto-Regressive Transformers), ami a BERT és GPT-2 ötvöze. 16 fejet, 24 réteget, 400 millió paramétert tartalmaz. Egy másik ismert reprezentánsa a T5 (Text-to-Text transformer). Ennek a tréningje maszkolt nyelvmodellezés. 128 fej, 48 réteg, 11 milliárd paraméter jellemzi.

[25] Waswani, A.–Shazeer, N.–Parmar, N.–Uszkoreit, J.–Jones, L.–Gomez, A. N.–Polosukhin, I. (2017): *Attention Is All You Need* (arXiv: 1706.03762. kötet). Letöltés dátuma: 2023. szeptember 17. forrás: <https://arxiv.org/abs/1706.03762>

[26] Kamath, U.–Keenan, K.–Somers, G.–Sorenson, S. (2024): *Large Language Models: A Deep Dive*. Berlin: Springer. doi:<https://doi.org/10.1007/978-3-031-65647-7>

[26] Kamath, U.–Keenan, K.–Somers, G.–Sorenson, S. (2024): *Large Language Models: A Deep Dive*. Berlin: Springer. doi:<https://doi.org/10.1007/978-3-031-65647-7>

[27] Srivastava, A. e. (2022): *Beyond the imitation game: quantifying and extrapolating the capabilities of language models*. <https://doi.org/10.48550/arXiv.2206.04615>

[28] Dou, Z. (2023): *Exploring GPT-3 Model's Capability in Passing the Sally–Anne Test: A Preliminary Study in Two Languages*. doi:<https://doi.org/10.31219/osf.io/8r3ma>

A nyelvi modelleket csillagászati nagyságú adathalmazon tanítják. Azonban ezek csak a tanult feladaton teljesítenek jól, ezért alapmodellnek (foundation model) nevezik őket. Gyakorlati alkalmazásokra e modellek speciális kérdések megfogalmazásának (prompt engineering) segítségével vagy az alapmodell további tréningjével tehető alkalmassá. Valamely részterületre vagy valamilyen feladattípusra alkalmassá a modellt finomhangolásnak (finetuning) nevezett plusz tanítás segítségével teszik a modellt. Bizonyos szabályok követésére, értékek tiszteletben tartására felügyelt tanítási módszereket alkalmaznak (Align with human feedback) [26].

Mesterséges mentalizáció

Számos kognitív feladat esetén a nagy nyelvi modellek meglepően jól teljesítenek, különösen a legújabb GPT-modellek, számos esetben az emberi referencia-eredményeket is meghaladva [27].

Néhányan vizsgálták azt is, hogy a korábban ismertetetett mentalizációs feladatokban mennyire eredményesek. Az eredmények nem teljesen egyöntetűek, de a tendenciák megállapíthatóak.

Egy angol és kínai nyelvű vizsgálatban a GPT-3 is teljesítette az elsőfokú téves vélekedés (Sally–Anne) tesztet [29].

Egy másik vizsgálatban különböző GPT-modellek teljesítményét tesztelték több hamis vélekedés feladatban (Kosinski, 2024). Egyedül a legfejlettebb modell, a GPT-4 oldotta meg a feladatok 75%-át, ami a szerző szerint a 6 évesek mentalizációs teljesítményének felel meg. Egy újabb tesztben különböző mentalizációs szintnek megfelelő feladatokon (téves vélekedés, ironia, faux pas, utalás, különös történetek) a nagy nyelvi modellek (GPT-4, GPT-3.5, LLama2-70B) és az emberek teljesítményét hasonlították össze (Strachan–Albergo–Borghini–Pansardi 2024). A GPT-4 minden feladatban (a faux pas-t kivéve) a legjobban, és az emberi átlag felett teljesített. A LLama2 a faux pas feladatban volt egyedül a legjobb, de az eredmények vele kapcsolatban nem egyértelműek.

Továbbra is tisztázandó kutatási cél marad az, hogy hogyan történik a nyelvi modellek esetén a szociális következtetés, és milyen vonatkozásokban hasonlít a humán mentalizációhoz, és milyen szempontok szerint különbözik az emberi kognitív folyamatoktól. Mivel a nyelvi modelleknek további fejlesztése várható, ezek a mesterséges mentalizáció további javulását is feltételezhetik.