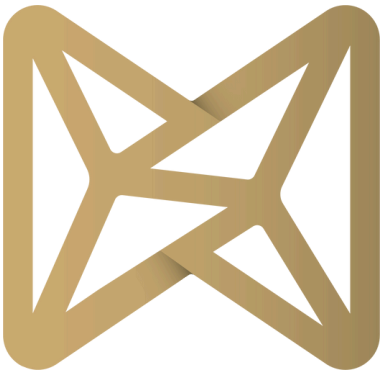
INTELLIGENCE ARTIFICIELLE

Pour les professionnels de l'IA



SROC

Système de Répartition Optimisée des Contextes

Version Française

NEURA KING

www.neuraking.com

ISBN 978-1-0369-0406-7

Système de Répartition
Otimisée des Contextes
SROC

Sommaire
SROC

Version Française

NEURA KING
www.neuraking.com

ISBN 978-1-0369-0406-7

Introduction

SROC

Système de Répartition Otimisée des Contextes SROC IA

Les grands modèles de langage (LLM) agissent comme des médiateurs, traduisant les intentions humaines en actions tangibles. Cependant cette médiation se heurte à deux problèmes.

La capacité des modèles à comprendre et interpréter la subtilité des demandes d'une part, et la capacité des utilisateurs eux-mêmes à formuler clairement leurs besoins, intentions et objectifs.

Le Système de Répartition Optimisée des Contextes (SROC) , permet de résoudre ces problèmes par une approche multidimensionnelle de l'IA, considérant à la fois les limites de l'intelligence artificielle et celles de l'humain.

Ce système optimise la gestion des données, des contenus et des contextes, assurant ainsi une distribution efficace des informations vers des intelligences artificielles spécialisées, tout en apportant des repères à l'humain, afin que ce dernier puisse correctement mettre en œuvre ses intentions.

En ajustant la formulation des demandes grâce à des suggestions d'intention, d'objectif, et de cadres d'exécution, le SROC améliore non seulement la précision des requêtes, mais aussi la pertinence des réponses en résultant.

ISBN 978-1-0369-0406-7

En prévenant la fatigue computationnelle et accroissant le focus d'attention des IA par dimensionnement dynamique des fenêtres de contexte, le SROC promet d'améliorer non seulement la performance opérationnelle dans divers contextes de travail, mais aussi l'expérience utilisateur.

En effet, le SROC contribue à la montée en compétences de ce dernier lors de ses interactions avec l'IA, en lui facilitant l'accès à des réponses adaptées et le guidant vers une expression fine de ses besoins.

Cette médiation optimisée par le SROC repose sur la gestion intégrée des facteurs limitants de l'IA. Des facteurs qu'il convient de comprendre pour appréhender leurs effets et les leviers d'optimisation admis par le SROC.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La formulation du requérant

Pour beaucoup d'humains, articuler avec précision des attentes face à des systèmes aussi avancés que l'IA peut être un défi de taille, d'autant que notre ego nous empêche de questionner notre incapacité.

En effet, l'utilisation d'un LLM paraît d'une simplicité enfantine, car nous employons uniquement des mots pour émettre des instructions.
Or, chaque mot véhicule une multitude de paramètres pouvant modifier drastiquement l'interprétation des requêtes et la formulation des réponses générées par l'IA.

Face à elle, nous sommes confrontés à la méconnaissance de notre propre langage et découvrons notre incapacité à nous exprimer avec la précision nécessaire qu'implique une bonne communication.

Les échanges entre humains étant intuitifs, notre approximation n'est jamais remise en question, alors que les échanges avec l'IA exigent une parfaite exactitude

ISBN 978-1-0369-0406-7

Les facteurs limitants

Système de Répartition Otimisée des Contextes SROC IA

La formulation du requérant

La précision, la finesse et les moindres nuances sont essentielles pour garantir que les résultats non seulement répondent aux attentes même si celles-ci sont mal définies au moment de la requête, mais aussi qu'ils les dépassent.

Ce à quoi concourt le prompt engineering dont les fondamentaux encadrent, sans intervention de l'utilisateur, chaque requête émise dans le cadre du SROC.

Mais cela ne suffit pas, car un phénomène intrinsèque aux LLM est à l'œuvre dans chaque interaction: la généralisation.

Les facteurs limitants

Système de Répartition Otimisée des Contextes SROC IA

La généralisation

Les LLM actuels ayant emmagasinés la quasi-totalité des données disponibles, ne souffrent pas d'un manque d'informations susceptible d'être comblé par du prompt engineering ou du fine-tuning, mais souffrent d'une trop grande dispersion de leur attention.

Le problème réside à la fois dans la pondération de l'interprétation des requêtes, et dans la pondération de chaque réponse.
Interprétation et génération n'étant que la somme de moyennes statistique à l'échelle de toutes les données ingérées.

Cela signifie que, lorsqu'on sollicite un modèle LLM, même si ce dernier est correctement prompté et fine-tunné, son niveau d'attention lié à un objectif particulier sera dilué en toute circonstance par effet de pondération.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La généralisation

Ce qui induit la nécessité d'augmenter la taille des fenêtres de contexte pour annuler cet effet de dilution par du prompt engineering, ou d'entrainer des modèles sur des données supplémentaires, plus spécifiques pour améliorer leur comportement, leur manière de faire ou leur pertinence.

Ce que permet la conception d'IA via SROC, ce dernier intégrant une architecture optimisant l'efficacité de toute instruction.

Mais bien que cela soit essentiel, car concourt à réduire les effets de la généralisation, toute requête va de toute manière être confrontée à d'autres phénomènes et facteurs limitants qui induisent à leur tour une pondération généraliste, à commencer par le facteur des conflits d'objectifs.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Les conflits intrinsèques d'objectifs

L'exemple du Copywriter :

Le copywriter rassemble plusieurs compétences, chacune d'entre elles rassemblant d'autres compétences et disciplines spécifiques.

Compétence 1 : L'Identification des cibles

Un bon copywriter doit savoir extraire les aspects psychologiques des cibles potentielles d'un message.
Ceci est une compétence à part entière, qui, pour être exploitée à son plein potentiel, doit être sollicitée indépendamment de tout autre objectif, ceci afin de puiser dans la profondeur de la tâche.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Les conflits intrinsèques d'objectifs

Compétence 2 : la technique AIDA (Attention/ Intérêt/ Désir / Action)

Cette deuxième compétence qui entre déjà en conflit avec la première (l'identification des cibles), diluant ainsi le focus d'attention de l'IA, est elle-même assujettie à un conflit d'attention entre quatre sous compétences:

- Capter l'Attention: implique des compétences spécifiques relatives à la captation et la rétention d'attention.
- Susciter l'Intérêt : implique la capacité à faire des liens logiques entre des cibles et des offres visant l'adéquation offre/besoin, ceci impliquant de caractériser la logique des liens d'adéquation.
- Provoquer le Désir: implique des compétences spécifiques permettant d'extraire les éléments de psyché pour agir, selon les besoins, sur les frustrations, les peurs ou les envies.
- Inciter à l'Action: implique l'usage de techniques appropriées visant à déclencher une action en se basant sur les éléments de psyché, d'intérêt et de désir.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Les conflits intrinsèques d'objectifs

Ici l'exploitation du plein potentiel de chacune des compétences impliquées dans ce qu'implique le terme de Copywriter, se résume à un conflit d'objectifs.

L'objectif est de capter l'attention, mais aussi de susciter l'intérêt, tout autant que de provoquer le désir et d'inciter à l'action.
Si l'humain peut intuitivement faire la part des choses en accordant plus ou moins d'importance à un objectif plutôt qu'un autre selon les circonstances, l'IA en est tout simplement incapable par la nature même de sa conception.

Elle réduit son interprétation à la somme des objectifs induits par chacune des compétences et sous compétences liées, ceci par supposition probabiliste, réduisant sa réponse à une addition de moyennes étant elles-mêmes des sommes d'autres moyennes probabilistes.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Les conflits intrinsèques d'objectifs

Pour le "Copywriting", comme pour tout autre métier, tâche, prestation, ce phénomène a lieu et requiert une segmentation fine de chaque élément afin d'en extraire les objectifs particuliers de sorte qu'ils n'entrent pas en conflit les uns avec les autres.

Une segmentation qui est permise par la méthode ESP (Exponential Segmention Process) que le SROC permet de mettre en application.

Mais cela, une fois encore, ne suffit pas, car la spécificité qui caractérise la pertinence est directement issue de l'arbitrage d'une échelle de priorité dans la gestion de ces conflits d'objectifs.

Un arbitrage qui, s'il est laissé à la libre appréciation du LLM, ne peut aboutir qu'à une généralité résultant d'une somme d'autres généralités elles-mêmes définies par une approche pondérée; en soi très éloignée de ce que pourrait être la pertinence attendue.

ISBN 978-1-0369-0406-7

Les facteurs limitants

L'arbitrage de la pertinence

Lorsqu'ils formulent une réponse, les LLM arbitrent à notre place les caractéristiques de la pertinence.

Cela ne concerne pas que le champ des compétences, mais aussi toute donnée de tout type, qu'elle soit une information, une instruction, une condition, etc., etc.

Or l'arbitrage de conflit d'objectifs inhérent à la caractérisation de la pertinence, quel que soit le domaine d'application, relève exclusivement de la perception subjective du requérant.

Reviens donc à ce dernier de la caractériser lui-même afin de restreindre les possibilités d'interprétation issues d'une généralisation.

Ceci en sachant que le requérant n'est pas assez expert pour formuler ces caractéristiques, et sachant que cela revient finalement à caractériser les critères de sa propre subjectivité, ce qui relève de l'impossible

ISBN 978-1-0369-0406-7

Les facteurs limitants

L'arbitrage de la pertinence

Alors, comment arbitrer dans ces conditions?

Eh bien, en se reposant sur des repères inspirés des structures sociales humaines permettant à chacun de se figurer des caractéristiques de pertinence, intuitivement admises par ce qu'impliquent ces repères; en l'occurrence des cadres de références, tels que des entreprises, des départements et des métiers.

Ce que propose également le SROC à travers une approche socialement bâti répliquant les réalités du monde du travail.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Négligence du repère anthropomorphe

L'anthropomorphisme est la tendance à attribuer des caractéristiques, des comportements et des émotions humaines à des systèmes artificiels, de sorte à les rendre plus accessibles.

Dans le cadre des modèles transformeurs, nous n'avons à faire qu'à un champ d'entrée permettant d'amorcer une discussion, aussi l'aspect anthropomorphique est négligé.

Pourtant l'assignation d'un métier (ou département ou autre cadre normé) par caractéristiques anthropomorphes concourt également à leur accessibilité, non pas par des principes émotionnels et d'identification, mais par le point de repère que l'anthropomorphisme métier induit.

Cela permet un pré-cadrage de toute requête, orientant les intentions du requérant vers l'objet du métier, accroissant de fait la pertinence, sans avoir à l'arbitrer selon des critères subjectifs, car ces derniers sont intuitivement inclus dans les notions qu'évoque le terme métier.

En effet le terme "métier" évoque une idée, une perception inconsciente d'un résultat qui pourrait concorder à nos attentes, que, finalement, nous ne savons pas exprimer consciemment avec un niveau de précision suffisant à être correctement interprété, mais dont l'expression inconsciente est incluse à l'intuition de l'objet du métier.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Négligence du repère anthropomorphe

Ce qui permet d'exprimer les caractéristiques de la pertinence subjective attendue, sans même les mentionner.

Il est donc primordial de ne pas sous-estimer l'importance de l'anthropomorphisme dans le cadre des modèles transformeurs afin de proposer des interfaces utilisateur qui structurent la navigation par des éléments qui sont accessibles à l'intuition de l'usager, de sorte que ses choix expriment à sa place ses attentes.

Dans cet objectif, le SROC propose une approche anthropomorphe permettant de répliquer virtuellement les univers d'entreprises, de projets et de métiers.

Mais une fois encore, cela ne suffit pas, car un autre phénomène vient atténuer le facteur anthropomorphique: les limites computationnelles communément assimilées aux termes de fatigue et de paresse de l'IA.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La fatigue computationnelle

La fatigue computationnelle désigne la diminution des performances des systèmes d'IA, résultant d'une sursolicitation, imposant une mitigation de la charge.

Ces épisodes de fatigue sont imprévisibles, fréquents (plusieurs fois par jour), peuvent durer quelques minutes ou plusieurs heures et ne sont repérables que par le constat d'une dégradation importante de la pertinence.

L'utilisateur non avisé, lorsqu'il émet une requête durant un épisode de fatigue, supposera donc que le recours à l'IA est inapproprié tant le résultat est médiocre.

L'IA aurait pu se montrer à la hauteur cinq minutes avant ou cinq minutes après la requête, mais ne le sachant pas, l'utilisateur décrète l'irrecevabilité.

Pire encore, lorsqu'il persiste à obtenir un résultat pertinent en insistant et se répétant sans obtenir le résultat attendu, ce qui le décourage durablement.

Cette frustration est d'autant plus grande lorsque le travail de prompt ingeneering a correctement été fait, et que tous les facteurs limitants ont été considérés, mais en oubliant d'intégrer le facteur de fatigue computationnelle dans l'édition des instructions et paramètres.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La fatigue computationnelle

Car la fatigue computationnelle n'empêche pas l'IA de répondre, elle l'empêche de considérer correctement les consignes et tous les cadres qui lui ont été donnés, supposés justement, garantir la pertinence.

Ainsi, plus les efforts déployés pour atténuer les facteurs limitants sont importants, se traduisant par une saturation d'instruction, plus cela est susceptible de produire un effet inverse lors d'épisodes de fatigue, car, dès lors, l'IA n'est plus en capacité de considérer l'intégralité des instructions, et encore moins de gérer les conflits intrinsèques.

Conduisant par la même occasion à une augmentation des hallucinations, phénomène également consubstantiel au fonctionnement des LLM, qui consiste en la production d'une réponse à tout prix, même si cette dernière est fausse.

D'où la nécessité d'intégrer la gestion de la fatigue computationnelle dans toute interaction avec l'IA d'une part, et d'aborder cette gestion par le prisme des effets de saturation d'autre part.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La saturation d'instructions

Le dosage du volume de contexte permet non seulement d'atténuer les effets de fatigue, mais prévient aussi le biais de sélection d'attention. Cet autre phénomène se caractérise par l'incapacité des modèles transformeurs à considérer l'intégralité de la fenêtre de contexte mise à disposition, même à pleine capacité de fonctionnement.

Certains travaux montrent qu'environ 40% de la taille de la fenêtre de contexte est réellement considérée. Bien que ce pourcentage soit en constante évolution et ne fasse pas encore consensus, il nous éclaire sur l'absence d'efficacité d'attention sur au moins la moitié de la fenêtre. Bien plus qu'il n'en faut pour considérer la capacité à désaturer à la volée, comme étant un élément fondamental.

D'une part, pour assurer un niveau de pertinence maximale dans les pires conditions possibles, prévenant ainsi la fatigue computationnelle.
Et d'autre part, pour s'assurer que le focus d'attention de l'IA soit établi par la limite des biais de sélection.

Cela aboutissant à un niveau de pertinence systématique, concourant à une organisation résiliente du travail assisté par IA, assurant ainsi la continuité des processus en action.

Toutefois, cette désaturation à la volée qui est permise par le SROC, introduit à nouveau un arbitrage subjectif revenant à l'humain.
Ce dernier, demeurant seul juge de la pertinence attendue au moment de la requête, doit pouvoir arbitrer facilement ses choix de désaturation.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Système de Répartition Otimisée des Contextes SROC IA

L'orientation des choix

L'utilisateur doit pouvoir s'appuyer sur une segmentation préalablement établie par méthode ESP, en assimilant les épisodes de fatigue computationnelle et les biais de sélection d'attention, comme étant le cadre de référence limitatif des capacités opérationnelles réelles.

Ainsi, une segmentation fine, dont les principes sont postulés dans un environnement préétabli, aboutira naturellement à des nomenclatures et labels de contextes que l'utilisateur pourra intuitivement reconnaître au moment où il devra désaturer une requête, identifiant d'un coup d'œil les éléments à activer et/ou désactiver.

Ceci, via SROC, par ajout ou retrait d'éléments de contextes à la volée, par simple clic, permettant à l'utilisateur d'adapter précisément, rapidement et intuitivement le contexte, aux fluctuations de fatigue computationnelle, aux troubles d'attention de l'IA, mais aussi aux fluctuations des conflits d'objectifs.

ISBN 978-1-0369-0406-7

Les facteurs limitants

L'orientation des choix

Rappelons en effet que l'accroissement du volume d'instruction implique l'accroissement des conflits d'objectif.

Ainsi, les effets de la désaturation à la volée ne se limitent pas à combler des déficits d'attention et des problèmes de fatigue, mais permettent également d'accroître la pertinence par un effet de vélocité d'attention.

Dès lors où la désaturation est couplée à la finesse d'un prompt engineering de qualité, elle permet l'induction d'un grand nombre de notions par les implications subjectives héritées de l'orientation anthropomorphique, le tout en un minimum de paramètres et d'instructions.

Ce qui concourt à la fois à la fiabilité et à la pertinence, tout en prévenant des situations hors limites.

Toutefois, pour mettre en œuvre cette finesse à triple bénéfice, il convient de considérer le facteur de l'intelligence à sa juste mesure.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Le facteur intelligence

Le terme d'intelligence artificielle induit des confusions qui poussent chacun à croire que le facteur intelligence est primordial.

Or, aucun des facteurs concourant à la pertinence d'un résultat ne fait intervenir de l'intelligence à caractère utile dans le cadre d'un modèle transformeur.

S'il y a bien une mise en relation entre des informations diverses et variées (en rappelant qu'elles sont systématiquement lissées et pondérées), faisant émerger le terme d'intelligence par la capacité à tisser des liens entre ces informations, la pertinence, elle, relève de bien plus que cela.

Lorsque nous demandons à un LLM d'exécuter une tâche, nous ne sollicitons pas sa capacité à faire des liens que nous assimilons à une forme de raisonnement, non, nous sollicitons sa capacité à faire le lien entre le brouillon de nos intentions et la vague idée de nos attentes en termes subjectifs de résultat.

Ce qui repose certes sur un certain niveau d'intelligence basé sur la grandeur du modèle d'IA, car cette grandeur est assimilée en grande partie à des capacités supérieures de raisonnement.

Mais au delà d'un certain seuil, la capacité à raisonner n'a plus rien à voir avec la pertinence, et n'a plus une incidence suffisante sur la qualité du résultat.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Le facteur intelligence

Car, quels qu'ils soient, les LLM sont soumis au cumul des facteurs limitants sus-cités, qu'aucune avancée technologique ne saurait résoudre, d'autant moins que le facteur subjectif de l'humain est omniprésent.

Ainsi, dès lors où une pertinence qui n'admet pas les généralités entre en jeu, il devient inapproprié de reposer ses attentes uniquement sur l'intelligence supposée d'un modèle transformeur, tout autant que sur un seul et unique levier, tel que le fine-tuning ou le prompt engineering.

Car l'utilité d'un LLM, sa pertinence et sa fiabilité reposent sur l'agrégation de multiples facteurs, chacun dans de multiples perspectives, chacune impliquant l'arbitrage de caractéristiques subjectives, ces dernières nécessitant une intervention dynamique de l'humain, seul juge de la pertinence d'un résultat face à des attentes que lui-même est incapable de définir avec une précision suffisante.

Le tout, encore et toujours dans la contrainte computationnelle et sous l'influence de biais de sélection d'attention, chacune de ces contraintes pouvant avoir un effet contre-productif face aux attentes.

Cela souligne l'importance de ne pas faire l'amalgame entre "intelligence" et "pertinence", afin de pouvoir orchestrer et diriger tout le potentiel des LLM dans la bonne direction, à savoir celle de la pertinence.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Intelligence versus pertinence

Les modèles LLM, avec leur vaste connaissance, s'apparentent à des oracles omniscients, doués de clairvoyance, érudits en toute chose.

Or, même un oracle ne peut donner de réponses pertinentes si les requêtes ne précisent pas en détail la réduction du champ d'interprétation, l'orientation souhaitée de la réponse, le contexte exhaustif, les objectifs imbriqués et les nuances comportementales attendues dans la gestion des priorités, elles-mêmes variables sur des critères subjectifs non maîtrisables, impossibles à anticiper et encore moins à caractériser sans le recours de notions tacitement induites.

Ces oracles étant perpétuellement confrontés à des conflits (nécessitant des arbitrages subjectifs) qui sont à la fois introduits par l'utilisateur, le prompt ingénieur, le fine-tuning et le préentrainement par deep learning non supervisé ne peuvent en aucun cas se montrer pertinents malgré leur intelligence.

Car l'expression de cette intelligence, en dehors d'un cadre limité à une intention clairement définie, n'est pas plus qu'une assertion généraliste dont la fiabilité repose sur notre confiance en cette expression.

D'autre part, une trop grande intelligence, basée sur ces modalités généralistes et la capacité à raisonner, produit des effets inverses.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Intelligence versus pertinence

En effet, une intelligence de ce type peut faire tant de liens de possibilités d'interprétations pour chaque mot composant des consignes, qu'il devient impossible de focaliser l'attention comme souhaité, tant la complexité est élevée.

En décuplant par conséquent les conflits intrinsèques d'objectifs, et complexifiant drastiquement la caractérisation de la pertinence attendue, l'intelligence en elle-même génère plus de problèmes qu'elle n'apporte de solutions.

Car lorsque nous décidons de lui accorder notre confiance sur des éléments que nous ne maîtrisons pas, nous ne prenons pas seulement le risque d'endosser la responsabilité d'une production médiocre.

Nous cédons également la souveraineté de nos décisions et effaçons la spécificité de notre jugement de la pertinence au profit de l'uniformité, ceci, en acceptant que des critères subjectifs de pertinence nous soient dictés alors qu'ils nous échappent.

ISBN 978-1-0369-0406-7

Les facteurs limitants

Intelligence versus pertinence

Pour peu que le problème de la fatigue computationnelle soit résolu par des avancées technologiques, tout autant que le problème des biais de sélection, cela ne résoudrait en rien la nécessité de désaturation visant à exacerber des priorités d'objectif plutôt que d'autres, dont seul l'opérateur peut être juge de la pertinence.

Ainsi, bien que les modèles transformeurs soient qualifiés de systèmes d'intelligence artificielle, leur forme d'intelligence est peu utile dans la quête de performance, d'autant moins comparée à ce que la pertinence obtenue par processus et considération des facteurs limitants peut apporter.

Ceci étant valable pour tous les modèles basés sur les mécanismes d'attention, car ces derniers fonctionnent sur des patterns statistiques.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La signature statistique

Les modèles LLM, sont non seulement obligés de généraliser de par leur limite en contexte et computation, mais ils le font de surcroit sur des bases erronées, car les conflits introduits par l'entraînement ou par l'utilisateur sont eux-mêmes constitués par une pondération généraliste.

Ainsi, aucune pertinence ne peut être attendue d'un LLM par la simple confiance qu'on accorde à son intelligence.

Une réponse pertinente ne l'est que par inadvertance statistique, et non par l'expression d'une logique maîtrisée concourant à l'accomplissement d'une tâche demandée.

Car les modèles LLM répondent en se basant sur des patterns statistiques, eux-mêmes issus d'autres patterns statistiques dont les conflits ont été traités par généralisation.

ISBN 978-1-0369-0406-7

Les facteurs limitants

La signature statistique

Nous voyons ici la signature du Deep learning non supervisé, essentiel à l'accroissement de l'intelligence au sens général, mais qui empêche toute spécificité et donc toute pertinence.

Cette accumulation de difficultés entraîne une diminution de la profondeur des réponses et soulève systématiquement la question de la pertinence par rapport aux objectifs, intentions et attentes.

Ainsi, en toutes circonstances, il convient de s'en remettre, non pas à l'utilité d'une intelligence supérieure, mais à aux limites de cette dernière, afin d'en tirer le vrai potentiel, ce dernier étant bien suffisant à bouleverser les paradigmes de travail et sociétaux.

Mais pas sans le concours de l'humain qui souhaite rester maître de ce qu'il produit avec l'assistance de l'IA.

ISBN 978-1-0369-0406-7

Les facteurs limitants

L'absence de limites

La gestion des facteurs limitants et de leurs interdépendances, tout autant que la gestion des dosages de contexte et des paramètres d'anthropomorphisme, requiert un système de gestion unifiée multi-perspective, bâti autour de ces contraintes.

Ceci, dans un environnement qui réplique les réalités et normes humaines afin d'induire les implications subjectives nécessaires aux repères qui orientent l'humain dans ses tâches avec l'IA.

Car, quel que soit le niveau d'intelligence d'un modèle de langage (LLM), il est, et demeurera incapable de fournir des réponses parfaitement pertinentes, tant que son fonctionnement n'est pas encadré par des limites prédéfinies.

Ces limites doivent être dynamiquement ajustables, permettant de gérer le niveau d'attention de l'IA sur certains objectifs plutôt que d'autres, tout en permettant d'atténuer en cas de besoin la saturation de contexte, permettant en toutes circonstances d'obtenir une adéquation satisfaisante entre l'attente et le résultat.

ISBN 978-1-0369-0406-7

Les facteurs limitants

L'absence de limites

C'est dans le respect de toutes les conditions mentionnées qu'un LLM peut être un médiateur pertinent dans l'accomplissement de nos intentions, transposé en tâches tangibles.

Dès lors, l'intelligence d'un modèle devient presque secondaire, car à ce stade de compréhension, le critère des coûts d'exploitation devient plus important que celui de l'intelligence dans le choix du modèle.

Car, lorsque les facteurs limitants sont considérés à leur juste valeur et qu'ils sont intégrés en conséquence, ils permettent d'obtenir d'un modèle supposément moins intelligent, l'équivalent de ce qu'un modèle dit plus intelligent est capable de produire en termes de pertinence.

À la différence près qu'un modèle considéré comme moins intelligent coute entre 10 et 100 fois moins cher pour un taux d'erreur qui est similaire, toute proportion gardée concernant la problématique de l'hallucination.

Cette dernière se résumant à un déficit d'informations afférant au volume des paramètres du modèle, qui par les voies d'une segmentation dynamiquement ajustable peut être comblé.

ISBN 978-1-0369-0406-7

Solution

ISBN 978-1-0369-0406-7

Toutes les problématiques évoquées se résument en une seule obligation: faire se rencontrer deux interlocuteurs (homme / IA) dans un environnement préétabli, cadré par normes, des sous-entendus tacitement acceptés, des objectifs dont les priorités sont sous-tendues par le milieu d'exécution, et basées sur des repères anthropomorphiques qui induisent des notions implicites indispensables à la désaturation et à l'accroissement des focus d'attention.

Ceci en intégrant l'élément fondamental que représente la limitation tous azimuts afin d'établir une zone de concordance propice à la pertinence systématique, admettant les variations issues de la subjectivité du requérant.

Définition du SROC

Le SROC pour Système de Répartition Optimisée des Contextes est un système de gestion de contenu et d'intégration systématique, transversale et multidimensionnelle de l'IA générative dans des applications de gestions et progiciels intégrés, reposant sur la gestion catégorisée des données, contenus et contextes dédiés à l'IA, assurant ainsi une distribution optimale auprès d'IA anthropomorphes spécialisées, par une approche considérant les limites intrinsèques de l'IA générative, tout autant que les limites cognitives de l'être humain.

ISBN 978-1-0369-0406-7

Objectifs du SROC

Les objectifs spécifiques du SROC incluent :

- L'optimisation des requêtes utilisateur : en proposant des suggestions d'intentions basées sur les contextes fournis et cadres d'usage, le SROC aide les utilisateurs à formuler des requêtes plus précises, augmentant ainsi la pertinence des résultats.
- L'amélioration de la performance : en rationalisant la distribution des informations aux intelligences artificielles spécialisées, le SROC vise à maximiser l'efficacité opérationnelle et à réduire les erreurs de traitement.
- L'accroissement de la satisfaction utilisateur : en garantissant des interactions plus intuitives et adaptées aux besoins des utilisateurs, le SROC renforce la satisfaction et l'engagement des utilisateurs dans leurs interactions avec les systèmes d'IA, notamment grâce à la pertinence des réponses issues d'une distribution rationalisée de l'information auprès d'IA spécialisées.

ISBN 978-1-0369-0406-7

Objectifs du SROC

Les objectifs spécifiques du SROC incluent :

- L'adaptabilité aux besoins évolutifs : le SROC se distingue par sa capacité d'adaptation aux besoins évolutifs, répondant efficacement aux exigences des utilisateurs et s'adaptant aux avancées technologiques. Cela assurant la continuité de la performance opérationnelle dans les organisations assistées par IA, et des seuils fiables de pertinence dans les réponses fournies.
- La flexibilité dans le choix des modèles: le SROC permet à chacun de choisir son fournisseur de modèle transformeur, tout autant que les paramètres afférents à la température (Déterminisme et créativité), à la quantité maximale du volume de tokens en input, output et/ou global.

En somme, le SROC établit un environnement où l'intelligence artificielle peut véritablement servir les utilisateurs.

D'une part par les leviers de performance et de flexibilité induits par l'exploitation de l'intelligence artificielle de type LLM, et d'autre part, en transformant des demandes approximatives en résultats précis, à la hauteur des exigences du monde professionnel.

ISBN 978-1-0369-0406-7

Capacités à gérer les facteurs limitants :

Le SROC permet:

- de calibrer à la volée l'activation et la désactivation de contextes afin d'ajuster les focus d'attention de l'IA, en décongestionnant le contexte d'instruction des éléments moins prioritaires selon la tâche à accomplir. Le choix du niveau de pertinence est ainsi ajustable par une désaturation combinée à un focus simultanément induit sur un objectif unique.
- d'orienter l'adressage de requêtes par département, métiers, tâches et sous tâches, afin d'induire un précadrage de requête, qui par l'approche métier et département demeure intuitif pour l'utilisateur. Ceci tout en lui enseignant par les suggestions et orientations inhérentes aux interfaces, la bonne approche de requêtes à adopter pour le conduire vers une pertinence optimale.
- de départementaliser le traitement des requêtes par une approche socialement bâtie, comprise par l'humain, lui permettant donc d'affiner ses choix d'interlocuteur IA selon les besoins dont il est conscient, mais qu'il ne sait pas formuler avec la précision nécessaire.

ISBN 978-1-0369-0406-7

Capacités à gérer les facteurs limitants :

Le SROC permet:

- d'exploiter les segmentations métiers issues d'une extraction ESP traduites en disciplines, en tâches, et sous tâches afférentes. Ceci à travers des équipiers IA spécialisés accompagnant toute IA anthropomorphe, chacune évoluant dans un contexte plus large. Permettant, ainsi, d'atteindre la pertinence au delà des attentes, par effet de vélocité d'attention sur tâche unique.
- de construire et d'attribuer des équipes accompagnatrices dites "globales", à des environnements d'exécutions particuliers, toujours par la segmentation et la subdivision. Ces équipes sont accessibles en permanence, quelles que soient les pages visitées, afin que l'hyper spécialisation soit à portée de main. Ceci, sans avoir à invoquer d'autres contextes et sans avoir besoin de changer d'environnement, les contextes s'ajustant sans intervention humaine.

ISBN 978-1-0369-0406-7

Capacités à gérer les facteurs limitants :

Le SROC permet:

- de segmenter la distribution de contextes ayant une nature variable, tels que les rendez-vous, les stocks et autres flux mouvants impossibles à intégrer en fine-tuning compte tenu de leur variabilité. Cette hyper segmentation d'éléments variables est primordiale, car l'influence des biais de sélection est dans ces circonstances très forte.
- de cloisonner les contextes par type de contenu afin d'en restreindre la compréhension et l'interprétation à l'objet exclusif qu'impliquent les typologies de contenu. Ceci à l'image des rendez-vous qui représentent un type de contenu, mais dans ce cas de cloisonnement, concernant d'autres types de contenus étant quant à eux prédéfinis et pouvant être affinés selon les besoins particuliers de segmentation par type.

ISBN 978-1-0369-0406-7

En mode projet

Perspectives d'intention de requête :

Sont pré-intégrées au SROC, les variations de perspectives permettant le traitement des requêtes suivantes :

- Les requêtes dont les intentions sont partagées à la fois par le contexte d'intervention du cadre de référence et par les interlocuteurs n'appartenant pas à ce cadre, invoquant les contextes unifiés partagés pour les utilisations frontales publiques et internes privées.
- Les requêtes dont les intentions ne peuvent émaner que d'utilisateurs participant activement au cadre d'exécution, en tant que membre ayant un rôle, un métier ou des attributions de tâches particulières, invoquant des contextes limités à l'utilisateur connecté, toujours dans le cadre de référence.
- Les requêtes dont les intentions ne peuvent être qu'à l'origine d'utilisateurs déconnectés, n'appartenant pas au cadre d'intervention, mais interagissant avec lui, invoquant les contextes d'interprétation afférents, mais soumis aux orientations du cadre de référence de l'environnement sollicité.

ISBN 978-1-0369-0406-7

En mode projet

Les valeurs d'interprétations globales:

Les valeurs d'interprétations globales sont des contextes propagés auprès de toutes les IA intervenant dans le cadre d'un environnement dédié, dit "projet".

Ces valeurs globales intègrent l'alignement et l'orientation éditoriale.

Contextes généraux déconnectés :

Les contextes généraux déconnectés permettent d'ajuster l'orientation des intentions de sorte à donner une perspective adéquate aux objectifs de tâches. Ces contextes sont propagés auprès de toutes les IA qui sont accessibles au grand public, sans nécessiter de compte.

ISBN 978-1-0369-0406-7

En mode projet

La segmentation par type de contenu :

Pour les flux variables :

- Rendez-vous
- Planning
- Stock

Pour les orientations d'intention:

- Stratégies
- Notes et mémo
- Contenus sociaux
- Contenus rédactionnels
- Contenus email
- Formations
- Foire aux questions

* Non exhaustif : directement lié à la capacité de choisir les catégories et types de contextes à faire considérer aux IA individuellement ou par le cadre général dit "projet".

ISBN 978-1-0369-0406-7

En mode projet et personnel

Contextes personnels :

Les contextes dits “personnels” ne sont propagés qu'auprès de l’utilisateur courant. Ainsi, les éléments qui lui sont propres sont connus de toutes les IA visitées par lui, sans que cela n’affecte les paramètres initiaux des IA.

Ces dernières recevant des éléments de paramètres différents selon les utilisateurs qui les consultent. Permettant ainsi la variabilité nécessaire à l’ajustement des intentions par préférences explicite.

Contextes généraux :

Les contextes dits “généraux” sont des informations et/ou compétences qui sont imposées globalement pour tout utilisateur (connecté ou non) agissant à l’intérieur de l’environnement dédié.

Ceci, quelle que soit la perspective de la requête, toujours dans la limite du cadre de référence des valeurs d’interprétations globales. Ces contextes généraux sont systématiquement considérés.

ISBN 978-1-0369-0406-7

En mode projet et personnel

Les IA anthropomorphes

Dans une interface de gestion de type SROC, chaque IA dispose de ses propres caractéristiques et paramètres, ces derniers opérant dans le cadre de référence global dit “projet”.

Caractéristiques anthropomorphes :

Profil

- Nom
- Age
- Genre
- Métier
- Prestation principale
- Description
- Lieu de travail

Expression :

- Qualité d’écriture ou d’expression orale
- Style
- Registre

Cadrage

- Le prompt à trou
- L’amorce de discussion

ISBN 978-1-0369-0406-7

En mode projet et personnel

Les IA anthropomorphes

Caractéristiques anthropomorphes :

Périmètre

- Contexte d'intervention
- Type de public / Audience
- Objectifs ou tâches à accomplir
- Rôle de fonction
- Posture en contexte conflictuel

Compétences

- Embarquées
- Spécifiques

Connaissances

- Embarquées
- Spécifiques

Alignement

- Critères d'alignement

Autres paramètres :

- Fournisseur API
- Choix du modèle fournisseur
- Choix du modèle fine-tunné
- Input limite
- Output limite
- Max sentence

ISBN 978-1-0369-0406-7

En mode projet et personnel

Les IA anthropomorphes

Les équipiers IA

Chaque IA peut être accompagnée d'équipiers, chacun disposant de ses propres caractéristiques anthropomorphes, chacune de ces caractéristiques étant orientée par les intentions de l'IA dite principale, à savoir celle qui est accompagnée par les équipiers.

Les équipiers agissent comme des employés sous l'autorité des intentions d'une IA, à l'intérieur du cadre général dit "projet".

Les équipiers peuvent être ajoutés pour des objectifs indépendants les uns des autres, ou ajoutés de manière chronologique aboutissant à un processus dans lequel les réponses de l'IA principale ont vocation à être transmises en toute ou partie au premier équipier dont les réponses ont à leur tour vocation à être transmises à l'équipier suivant et ainsi de suite, permettant d'aller en extrême profondeur dans la pertinence recherchée.

L'élaboration de ce processus repose sur la méthode ESP qui peut être exploitée à travers la Team dite Jaris.

ISBN 978-1-0369-0406-7

En mode projet et personnel

Les IA anthropomorphes

La team Jaris

Les intelligences artificielles qui composent la TEAM Jaris accompagnent systématiquement toute intelligence artificielle en tant qu'architectes collaborant à son paramétrage.

Les IA de la TEAM Jaris agissent également comme des employés de l'IA principale, mais incorporent, en plus des intentions de cette dernière, tous ses paramètres. Permettant ainsi aux IA de la Team Jaris d'affiner de manière incrémentale les paramètres de l'IA principale.

Les teams accompagnatrices

Les teams accompagnatrice sont des regroupements d'IA départementalisés qui incorporent les cadres de référence du projet ainsi que les cadres de référence des IA anthropomorphes que l'utilisateur utilise.

Sur toute autre page, les teams accompagnatrices incorporent uniquement les cadres de référence du contexte d'exécution global dit "projet".

Ces équipes peuvent être personnalisées par département métier, disciplines, et tâches.

Ce qui permet à l'utilisateur de composer son propre environnement selon ses besoins les plus fréquents, en restant dans le cadre limitatif de l'environnement global.

ISBN 978-1-0369-0406-7

En mode projet et personnel

L'expandeur incrémental

L'expandeur incrémental permet la distribution profonde à la volée par la suggestion d'IA spécifiquement dédiées à des tâches particulières, favorisant le focus sur objectifs induits par la proposition du bon interlocuteur IA.

L'expandeur incrémental permet à partir de n'importe quel contenu de dépasser les limites de longueur et de profondeur en régénérant systématiquement, sur demande, via sélection de mots, de zones, de phrases ou de passages, d'autres requêtes qui se reconstruisent dans un nouveau contexte qu'est celui du contenu ou texte sélectionné.

Les propositions qui sont faites par l'ouverture d'une boite de dialogue orientent le choix de l'utilisateur vers le bon interlocuteur IA, c'est-à-dire celui dont les paramètres et objectifs sont concentrés sur un seul élément d'intention. Ce qui garantit une pertinence optimale de la requête et de la réponse attendue, telle que pensée par le requérant.

Les IA sollicitées dans le cadre de l'expandeur incrémental intègrent également les références générales tout autant que les références et caractéristiques de l'IA principale en cours d'usage.

Cela se traduit par des gains significatifs en termes de pertinence, de rapidité et d'efficacité, tout en offrant une expérience utilisateur enrichie et intuitive.

ISBN 978-1-0369-0406-7

https://neuraking.com/sroc/