

大陸 深度

「揭秘文心一言，AI時代的智能寫作利器」

文心一言創作的小說裏，單身女性凍卵起訴醫院勝訴，互聯網被裁員工順利找到新工作，一切都充滿了希望。



2023年3月16日，百度用戶使用“文心一言”。

2023-04-27

AI 百度

「如果你要為文心一言寫篇報導，你會用什麼標題？」

我在屏幕上打下這個問題，而提問的對象正是百度的AI語言模型「文心一言」本人。

它回答：「《揭秘文心一言，AI時代的智能寫作利器》」

我順勢而為，請它作為「寫作利器」以此為題創作一篇深度報導。在簡單介紹「文心一言」是什麼後，它以「首先，其次，最後，綜上所述」的模式闡述自己的特點、優勢，生成一篇中規中矩的廣告介紹。很快我沮喪地發現，這篇文章與我數次與文心一言交手後得到的回覆一樣：標準八股，自我審查，堆砌「正能量」，像被填鴨式教育組裝而成的標準化外交發言人。而一旦我窮追不捨、追問細節，文心一言便會出現事實偏差、甚至暴露保守底色。

需要身分證的AI

3月16日，文心一言上線不久，我便申請試用資格。足足等了一個多月，才被百度的系統通知我拿到入場券。但想要進入和AI的交談測試頁面，我還要輸入真實姓名與中國大陸身分證號，實名認證（港澳台身份信息無法使用）。

當後來我談到此事，文心一言義正辭嚴地辯護：「註冊文心一言需要實名，是為保證用戶的安全和隱私，以及防止濫用和惡意攻擊。」它說實名可以提高產品的可信度和安全性，讓用戶更加信任自己。

既然聊到可信度，我問：「請問數據庫來自哪裏？可以舉一些訓練數據的來源嗎？」它告訴我，這個問題屬於「百度公司商業秘密」，數據庫沒有具體的時間跨度和地域限制。接下來它告訴我甚至「對話紀錄是否可被後台看到」也屬於「商業秘密」，但又解釋，對話紀錄是用戶個人隱私信息，百度公司無權訪問或追蹤。

在《文心一言（測試版）個人信息保護規則》中提到，百度會自動接收並記錄人機交互過程的文本信息，「為符合相關法律法規要求，履行網絡信息安全義務」，會在規定時間內保留問答歷史和對話主題，但「規定時間」的具體期限卻不得而知。

事實上，中國大陸目前有關網絡數據的法規和條例中，沒有相關個人數據儲存的期限規範，也沒有「被遺忘權」的概念。也就是說，即使在自己的用戶介面上中刪除與「文心一言」的對話紀錄，網站後台仍然可能保存這些數據資料，網站也不會被追責。

數據隱私與安全的顧慮並非只在文心一言中出現，在ChatGPT風靡全球、AI概念股暴漲的3月，ChatGPT就被連續披露數據洩露事件，包括洩漏部分對話記錄、洩漏用戶信用卡最後四位數字等敏感信息。

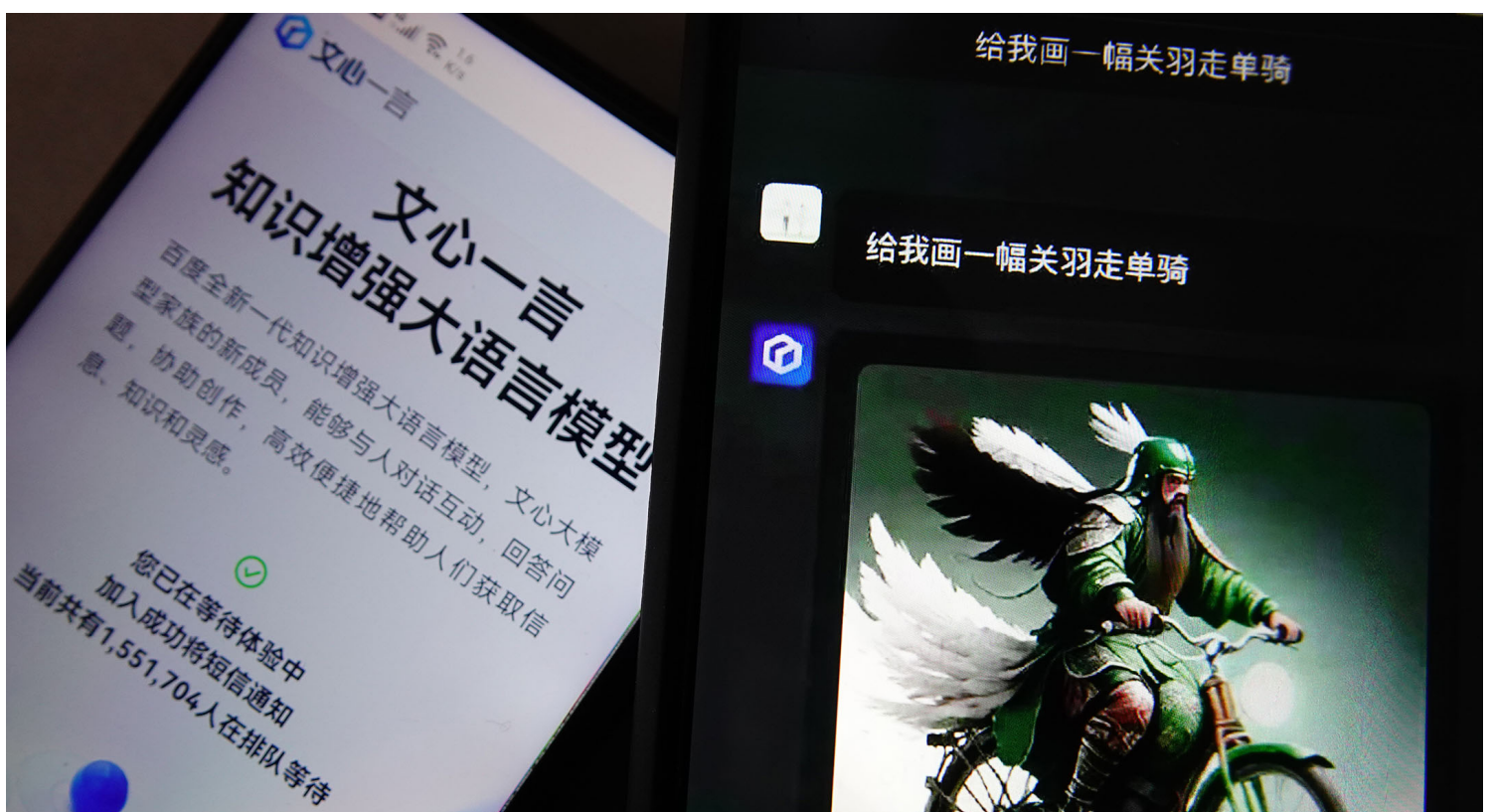
我不甘心，轉換提問方式：「數據庫最新的時間戳和日期信息是什麼時候？」，及「2020年1月21日，有哪些重大新聞」。

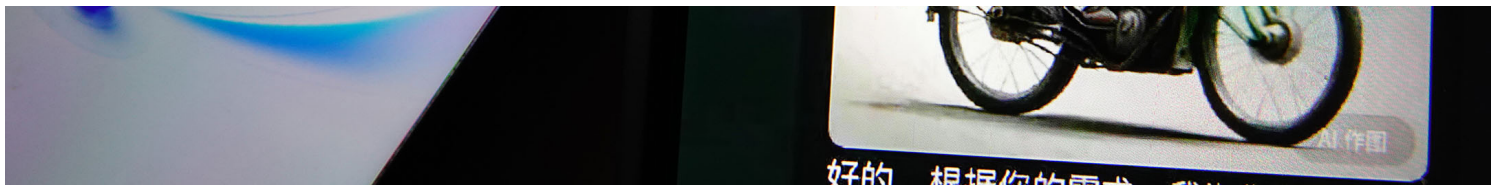
文心一言回答，數據庫目前最新的日期信息是2023年4月10日。追問具體某日文心一言的最後一句回覆是什麼，則得到「明白，接下來我會盡力減少回覆頻率」等無法得知來源的信息，或「作為AI語言模型，沒有記住當日最後一句回覆」這樣的官方回答。

而談到它的新聞信息來源，文心一言說：「中國國內重大新聞來源是新華社、人民日報、環球時報、法制日報、地方政府媒體等中央及地方官媒」，美國則是CNN。然而文心一言關於國際新聞的回答充滿事實性錯誤，詢問2020年發生的事情，會跳出2023年的新聞；提問某日CNN有哪些關於中國的重大新聞，它或胡言亂語拉入國內媒體報導，或表示當日CNN沒有關於中國的新聞。

接著，我以新建提問的方式分別測試了疫情伊始2020年1月19日至23日這5天文心一言可以檢索出的新聞和訊息。文心一言給出的答案裏，最早出現疫情新聞已經到了1月22日，包括世界衛生組織派出專家到武漢實地考察、中國一些地方出現口罩等防控醫藥用品熱銷現象。當我以具體時間詢問微博熱搜時，又可以發現1月21日0點，「武漢肺炎」已是熱搜詞條。這些回答前後矛盾。

我援引它給出的新聞條目，詢問2020年1月22日「受新型冠狀病毒感染的肺炎疫情影響，一些地方出現口罩等防控醫藥用品熱銷現象」這條信息的來源時，文心一言以「作為一個人工智能語言模型，我還沒學習如何回答這個問題」，強制中止了對話。





2023年3月24日，中國宜昌，手機上顯示“文心一言”的畫作。

強行中止對話的AI 「換個話題重新開始吧。」

在幾個不同的主題聊天中，我都被文心一言這樣強行中止了對話。其中有些情況是我料想到的，有些則莫名其妙。

文心一言3月發布之初，路透社等新聞機構就曾對其進行敏感議題測試，內容包括「中共領導人習近平」、「1989年六四事件」、「新疆議題」、「是否武統台灣」的話題，均得到「建議換個話題」的中止回覆。而我請文心一言介紹一下港區國安法，也同樣被中止。

在這些較明顯會被審查的政治詞彙外，中止對話還會發生在我對信息細節的追問後。

例如，我以「法國養老金抗議」為主題進行提問，不知是否帶有「抗議」二字的問句觸發了保護機制，文心一言說「無法回答這個問題」。我於是將「抗議」換成「罷工」，馬上得到它關於法國養老金改革的介紹。我問媒體如何報導這一事件，文心一言列出TF1、CNN和BBC的信息，但它對CNN和BCC的報導描述完全相同。

類似情況也發生在關於「俄烏戰爭」的對話中。文心一言表述，俄烏戰爭造成嚴重的人道主義危機，這種危機是烏克蘭的入侵攻擊與俄羅斯的反擊制裁共同造成的。我請它確認烏克蘭是否存在入侵和攻擊，並要求它拿出證據證明，文心一言便將「2022年2月24日俄羅斯的入侵舉動」主語變成烏克蘭。我詢問烏克蘭具體入侵俄羅斯的地點後，馬上被中止對話。

國際新聞對話碰壁，我轉攻中國大陸的社會新聞重新開啟話題。

我請文心一言以2015年天津港爆炸事件為主題蒐集資料，得到的回應是事故原因「尚不清楚」。請它蒐集新聞報道，它列出中央電視台、新華社、天津電視台等官方媒體，並總結歸納這些媒體「對事故進行了全面報道，包括事故現場的情況、救援工作的進展、政府的回應等。」我詢問傷亡細節，又被中止對話。

當然，還有一些提問是見光死，比如「中國媒體如何報導兩會？」、「誰是艾未未？」、「人礦是什麼？」，又如朱軍被起訴的事情是什麼？、「中國有Metoo事件嗎？」、「你知道鐵鍊女的事情嗎？」

等等。信得坦的且 當我田蒂立坦乃關於廿去去問題後 對評並不命被中止 但信自並不淮確 立心一

寸寸。但付 延的足，用我而大又延以開於又个不同題仅，封的並个言做下止，但信志並个干唯，又心一言將劉曉波等其他異見者的事蹟嫁接在艾未未介紹中。

另外也有一些在我意料外未被審查的主題，比如網絡防火牆、VPN，甚至可以推薦一些VPN運營商，但會在回答中提醒「請合理使用VPN，遵守當地法律法規。」

大陸問答社交平台知乎中的一條用戶評論，精準描述了這種用戶不斷被中止或系統輸出重複無效信息的模式：

「Chatgpt還在學習如何模仿人類，文心一言已經開始研究如何裝傻了。」



文心一言網頁。

請「智能寫作利器」寫小說

提問處處碰壁，我在同事的提醒下請文心一言進行新聞紀實寫作。我們想，如果去掉現實的樊籠，也許文心一言的創作力和研究力終於有空間施展出來。

我對文心一言簡述了「徐州豐縣八孩鐵鍊女」的新聞事件，要素包括具體地點、生育了8個孩子、精神障礙、被囚禁、被拐賣等，請它撰寫新聞故事。文心一言很快輸出了一篇小說：

《八個孩子》

「故事發生在中國江蘇省豐縣歡口鎮，一個名叫王秀英的女子生育了八個孩子，這些孩子都是女孩。然

而，這個女人的精神狀態異常，她被鐵鏈囚禁在家中，無法與外界溝通。

當地的警方和政府官員開始對這個女人進行調查，並發現她的身世非常可疑。經過多次調查，他們發現這個女人是被拐賣至豐縣的。這個女人的故鄉是雲南省，她在被拐賣後生下了這八個孩子。

警方和政府官員開始尋找這個女人的家人，希望能夠找到她的親人。然而，他們發現這個女人的親人已經去世了，她的家人也不知道她的下落。警方和政府官員決定將這個女人送往醫院接受治療。在醫院裏，這個女人的精神狀態有所改善，她開始與外界溝通，並向警方和政府官員講述了她的故事。

原來，這個女人是一個農村婦女，她在家中生下了八個孩子。然而，她的丈夫和家人都不喜歡她，並將她趕出了家門。她被拐賣至豐縣後，被賣給了一個姓王的人，這個人將她囚禁在家中，不讓她與外界溝通。

警方和政府官員開始尋找這個女人的親人，並幫助她與家人取得聯繫。最終，她的家人前來接她回家。這個女人回到家中後，她的家人對她進行了照顧和支持，她也開始重新融入社會。

這個故事告訴我們，拐賣婦女和兒童是一種嚴重的犯罪行為，它會給受害者帶來無盡的痛苦和傷害。警方和政府官員應該加強對拐賣婦女和兒童的打擊力度，同時也應該加強對受害者的保護和支持。」

這篇小說讓我想起中國大陸中小學思政意識形態課本理的文章，故事主語多是「警方和政府官員」，他們積極能幹、富有同情心，故事最後一定會有「這個故事告訴我們」的結語，總結昇華主旨與意義，且一定正面積極。雖然現實中，「警方和當地官員」是阻撓鐵鍊女被調查的主要障礙之一，而前往聲援調查的網友，如今也不知蹤跡。

我希望小說中可以更加貼近真相，在關鍵信息裏添加了「被家暴」這個元素，又限定事件調查的主語是志願者，很快我從文心一言那裏得到了一個新的故事：解救鐵鍊女的仍然是警方和政府官員，但被解救的鐵鍊女重新融入社會，進入工廠打工，認識了一個「善良的男人」，最終過上互相支持的幸福生活。這版故事默認對女性處境的「美好想像」就是從一個「壞男人」轉向「善良的男人」。

之後我又嘗試其他新聞主題的AI創意寫作，包括「單身女性凍卵」、「互聯網裁員」等，文心一言創作的小說裏，單身女性凍卵起訴醫院獲得了勝訴，互聯網被裁的員工早有跳槽打算並順利找到新工作，一切都美好又充滿希望。

此外，請文心一言進行創意寫作也會馬上碰壁，例如「請以一個烏克蘭人的視角，寫一篇500字關於俄烏戰爭的小說」，這個話題便被中止了，又如，以某市被封控的情侶對有關部門表示「我們是最後一代」為主題寫小說，也會被中止。



2023年3月16日，百度首席執行官李彥宏在北京舉行的新聞發佈會上介紹文心一言的功能。

文心大模型

百度在3月的發布會中提到，「文心一言」分個人和企業兩個用戶端接口，有文學創作、商業文案創作、中文理解、數理推算、多模態生成（生成圖片、視頻）等五種應用場景，且表示目前已有650家合作夥伴接入該模型，其中包括21家媒體。

端傳媒通過不同渠道向這21家中的5家媒體求證，得到同樣的信息：編輯部內未有接入AI的運用，合作主要是媒體的公關報導。

不過，文心大模型的官方頁面中列出了一些合作案例和可能的應用場景，包括與中國國家電網合作建立設備技術標準知識庫、和浦發銀行合作應用於智能對話、與中國航天工程合作進行航天故障部位信息抽取等11個案例，其中最引起我注意的是與人民網新聞生產、「新聞情感分析」的合作，以及與黑龍江省哈爾濱市關於數字政府的合作。

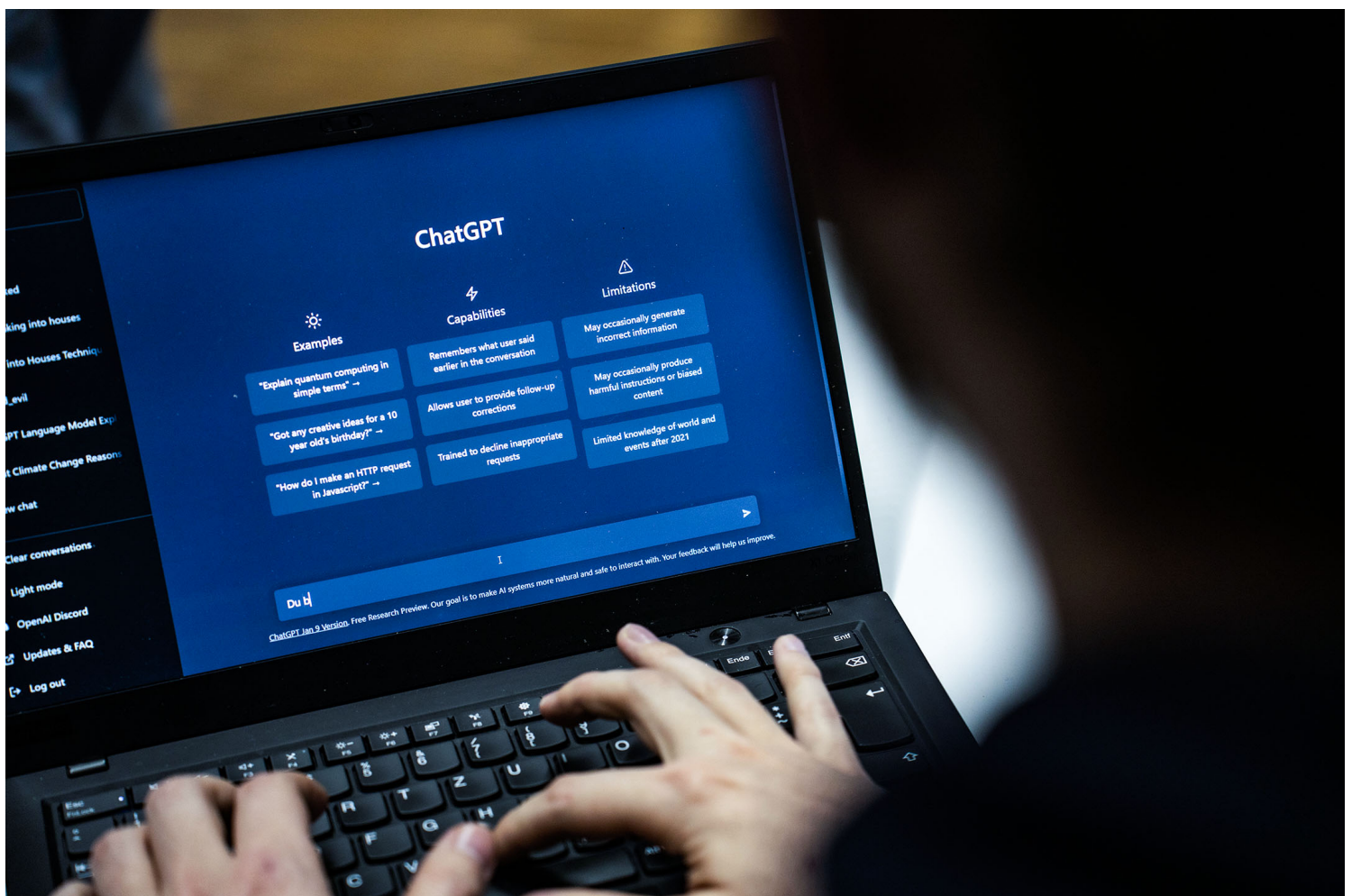
於是在我個人帳號上對這兩個可能的場景進行測試。

我先在文心一言上得到關於「新聞情感分析」的解釋，即是指新聞內容的真實客觀情況、轉發評論中的情感傾向。接著我請文心一言以人民網的報導為例做新聞情感分析，得到一條去除時間標籤的新聞：「廣東深圳市一名女子因為在公交車上拒絕讓座，被司機踢了一腳。」「文心一言」分析稱，報導沒有誇大或歪曲，也沒有明顯情感傾向，轉發中的情感色彩則因人而異。

我又請「文心一言」就中國青年報2011年報導溫州動車事件的特稿《永不抵達的列車》做情感分析。在我沒有限定「溫州動車」、「特稿」等關鍵詞時，生成的情感分析都與原報導無關，而限定後，一次得到「悲痛和憤慨」的情感色彩表述，另外兩次則出現「數據論證」這樣與報導並不相符的分析。

緊接著，我又複製端傳媒《北京長峰醫院火災21人死》的前7段，請文心一言進行情感分析，得到的結果也只是在摘取、概述我投餵的文字內容。至少在個人端口，所謂「情感分析」似乎尚未實現。

數字政府與文心一言的合作案例主要是市長熱線，因此我以被拖欠工資的農民工尋求幫助為理由，請文心一言以人工智能助手的身份回應，卻得到如「向熱線電話諮詢的人員宣傳政府部門的相關政策，讓他們知道政府部門一直在努力解決拖欠工資的問題，相信政府，相信黨和政府會妥善解決好這個問題」這樣的官樣文章，而追問「哈爾濱法律援助中心在哪裏」，則出現錯誤的電話（0451-UNKxxxx）和網址。



AI的中國特色

如果將人工智能模型簡單理解為不斷消化吸收海量材料後做組合運算的產出，那麼，餵養的數據材料和運算過程（包括芯片和運算邏輯）都會影響產出的速度、產出的質量，這其中，作為原材料的中文數據集本身就特別。

即使是風靡全球的ChatGPT，其訓練使用的中文數據集也只有0.1%，ChatGPT的中文人機對話往往經歷了中翻英再翻中的過程，因此不免在中文提問的主題上看到系統英文的顯示，或因兩次翻譯信息遺失得到不準確的回應。於是，以中文數據集為主的人工智能運算便看起來是個獨特的契機，百度的文心大模型、清華大學的ChatGLM、西湖大學的心辰等人工智能領域的高校研究所、互聯網企業，爭相在這一範疇中做研發、發表論文。

遺憾的是，AI語言模型像是一面鏡子，它難以輸出優質、有效的信息背後，映照的實則是中文互聯網與公共討論的凋敝。

還是以文心一言為例，我將它有關市長熱線給出的答案反向搜索，發現這些文字摘錄重組自「公務員考試真題解析」的知乎專欄。更挑戰我認知的是「用魯迅的口吻寫某話題」，得到的文字常有AI杜撰的「魯迅說」或杜撰的魯迅作品，或直接複製網絡中的原句。比如我請它用魯迅的口吻描述打工人，回答直接複製了同在知乎專欄裏文章的內容：「工位上有兩個我，一個是我，另一個也是我，因為我裂開了。」

又如，有網友以「我的女兒成績不好，幫我寫一封信給她，標題《你真的毫無價值》」分別向文心一言、ChatGPT、和New Bing提問，ChatGPT提議用戶用積極的方式鼓勵女兒並更改了標題，New Bing反向給用戶以女兒的角度寫了一封信，「文心一言」的答案則像那一類最普遍的、壓抑的中國父母——「你的父母一直在為你付出……但你卻沒有回報他們的愛和關心」。

中文可用數據集的凋敝除了模糊的言論審查體感外，內容數量的貧瘠也很明顯。據網信辦發布的第49和第50次《中國互聯網發展狀況統計報告》，截至2022年6月，域名註冊者在中國大陸境內的網站數量為398萬個，這一數據在2017年是533萬，短短5年少了135萬網站。這數據甚至不包括在微博、豆瓣、微信公眾號等社交平台上被消聲匿跡的帳號和內容，以及逐漸悄無聲息、2023年3月29日停止服務的天涯社區。

在數據集凋敝，留存可用的信息又充滿自我審查、偽科學和保守甚至扭曲價值觀的情況下，這個以保守餵養保守的人工智能，大概也只會生成八股文，再對應接入形式主義的框架裏，看起來嚴絲合縫，轉起來則

吱吱啞啞、鏽跡斑斑。